We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900
Open access books available

185,000
International authors and editors

200M
Downloads

154
Countries delivered to

Our authors are among the

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS

BOOK CITATION INDEX

INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Bayesian vs Frequentist Power Functions to Determine the Optimal Sample Size: Testing One Sample Binomial Proportion Using Exact Methods

Valeria Sambucini

Abstract

In order to avoid the drawbacks of sample size determination procedures based on classical power analysis, it is possible to define analogous criteria based on 'hybrid classical-Bayesian' or 'fully Bayesian' approaches. We review these conditional and predictive procedures and provide an application, when the focus is on a binomial model and the analysis is performed through exact methods. The distinction between analysis and design prior distributions is essential for the practical implementation of the criteria: some guidelines for choosing these priors are discussed, and their impact on the required sample size is examined.

Keywords: analysis and design prior distributions, binomial proportion, Bayesian power functions, conditional and predictive approach, sample size determination, saw-toothed behaviour of power

1. Introduction

The calculation of an adequate sample size is a crucial aspect in the design of experiments. Researchers need to select the appropriate number of participants required to ensure ethically and scientifically valid results. If samples are too large, time and resources are wasted, often for minimal gain. On the other hand, too small samples may lead to inaccurate results. Therefore, sample size determination (SSD) plays a very important role in the design aspect of studies in many fields, especially in the context of clinical trials where, in addition to economical problems, investigators have to deal with important ethical implications.

Sample size determination (SSD) methods, when the focus is on hypothesis testing, are typically related to the concept of *power function*. Let us denote the parameter of interest by θ and

let us assume that we are interested in testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$, where Θ_0 and Θ_1 form a partition of the parameter space Θ. The most widely used frequentist SSD criterion consists in choosing the minimal sample size that guarantees a given power, for a fixed type I error rate, under the assumption that θ is equal to a suitable *design value*, $\theta^D \in \Theta_1$. In practice, the idea is to ensure a sufficiently large probability of obtaining a statistically significant result (i.e. of rejecting the null hypothesis), when the true value of θ belongs to the alternative hypothesis and is equal to θ^D. In many textbooks (see [1–3], among others) sample size formulas, derived using this procedure, are provided in many occurring situations, under different hypothesis testing and based on both categorical and quatitative data.

In the frequentist criterion described above, a crucial role is played by the design value that the trial is designed to detect with high probability, whose uncertainty is not accounted for. In fact, the local optimality is one of the most criticized aspects of the method. Moreover, this frequentist procedure does not allow to take into account pre-experimental information about θ, for instance available from previous studies. By adopting a 'hybrid classical-Bayesian approach' or a 'fully Bayesian approach', it is possible to define analogous criteria for sample size selection that allow the researcher to avoid the problem of the local optimality or/and to introduce possible prior information in the SSD process.

In this chapter, we illustrate how to construct frequentist and Bayesian power functions, based on both conditional and predictive approaches, and how to use them to determine the optimal sample size. An essential element of the method is the use of two different prior distributions for the parameter of interest, which play two distinct roles in the criteria. The importance of this distinction in sample size determination problems has been stressed by several authors (see, for instance, [4–9] among others). The rest of the chapter is organized as follows: in Section 2, we review both the frequentist conditional and predictive procedures based on power analysis to determine the optimal sample size. Section 3 provides a description of analogous methods based on Bayesian power functions. Then, in Section 4, we formalize different SSD criteria that depend on the shape of the power curves as a function of the sample size and, as a consequence, on the nature of the data distributions. Furthermore, in Section 5, we illustrate an application of the frequentist and Bayesian SSD procedures, when the parameter of interest is a single binomial proportion. Finally, Section 6 contains a brief final discussion.

2. Frequentist power functions and SSD methods

Let us consider a parameter of interest θ and assume that we are interested in testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$, where Θ_0 and Θ_1 form a partition of the parameter space Θ. Moreover, let Y_n be the random result of the experiment that is typically a suitable statistic used to summarize the data relevant to the parameter θ. In the notation, we have highlighted that Y_n depends on the sample size n. Finally, we denote by $f_n(y_n | \theta)$ the sampling distribution of Y_n.

The power function is defined as the probability of obtaining a statistically significant result that leads to reject the null hypothesis H_0, when the actual value of the parameter is θ. In a frequentist approach, the investigator is firstly required to specify a fixed level α for the type

I error probability that one is willing to tolerate. This significance level is typically set equal to 0.05 and is used to obtain the rejection region of H_0, denoted by R_{H_0}, that represents an appropriate subset of outcomes that—if observed—lead to the rejection of H_0. Therefore, given a frequentist test of size α, Y_n is considered a statistically significant result if it belongs to R_{H_0}. Consequently, in general terms, the power function is defined as

$$\eta(n, \theta) = \mathbb{P}_\theta(Y_n \in R_{H_0}), \qquad (1)$$

where $\mathbb{P}_\theta$ is the probability measure associated with a suitable distribution of Y_n.

In order to exploit the frequentist power function in Eq. (1) for sample size determination purposes, investigators can adopt two different approaches: the conditional and the predictive one. The conditional approach is certainly the most widely known and used, when performing sample size calculations based on pre-study power analysis. It requires the specification of a suitable *design value* for θ, denoted by θ^D, that belongs to the alternative hypothesis and is considered a relevant value important to detect. By assuming that the true value of the parameter is equal to θ^D, we obtain the *frequentist conditional power* given by

$$\eta_F^C(n, \theta^D) = \mathbb{P}_{f_n(\cdot|\theta^D)}(Y_n \in R_{H_0}), \qquad (2)$$

where $\mathbb{P}_{f_n(\cdot|\theta^D)}$ is the probability measure associated with the sampling distribution of Y_n when $\theta = \theta^D$. Since θ^D has to be selected within the subspace Θ_1, the conditional frequentist power can be interpreted as the probability of correctly rejecting H_0, when the true value of the parameter belongs to the alternative hypothesis and is exactly equal to θ^D. Then, the sample size determination criterion consists in choosing the minimal sample size that guarantees a desired level for $\eta_F^C(n, \theta^D)$. In practice, the idea is to ensure a sufficiently large probability of rejecting H_0, when the true θ belongs to the alternative hypothesis and, more specifically, it is equal to $\theta^D \in \Theta_1$.

The SSD procedure based on the power function in Eq. (2) is strongly affected by the choice of θ^D. In order to account for uncertainty in the specification of the design value and to avoid local optimality, it is natural to incorporate Bayesian concepts into the sample size determination process. By adopting a 'hybrid classical-Bayesian approach', it is possible to model uncertainty on the appropriate design value for θ through the elicitation of a prior distribution, denoted by $\pi^D(\theta)$ and called *design prior*. This prior is used to compute the marginal or prior predictive distribution of the data by averaging the sampling distribution as follows:

$$m_n^D(y_n) = \int_\Theta f_n(y_n|\theta)\pi^D(\theta)d\theta. \qquad (3)$$

Therefore, the design prior cannot be a non-informative improper distribution in order to have $m_n^D(y_n)$ well defined. In any case, the elicitation of a non-informative $\pi^D(\theta)$ would not be reasonable choice. In fact, the design prior is used to introduce uncertainty on the suitable design value for θ that we need to specify when using the SSD procedure previously described and the possible guessed values have to belong to the subspace Θ_1. Thus, $\pi^D(\theta)$ serves to describe a design scenario of interest that supports values of θ under the alternative hypothesis:

it has to be an informative distribution that assigns a negligible probability to values of θ under the null hypothesis.

Once the design prior has been elicited, the idea is to average the conditional frequentist power with respect to it by computing

$$\int_\Theta \eta_F^C(n,\theta)\pi^D(\theta)d\theta = \int_\Theta \left[\int_{R_{H_0}} f_n(y_n|\theta)dy_n\right]\pi^D(\theta)d\theta$$
$$= \int_{R_{H_0}} m_n^D(y_n)dy_n. \tag{4}$$

This leads to the *frequentist predictive power* that is given by

$$\eta_F^P(n,\pi^D) = \mathbb{P}_{m_n^D(\cdot)}(Y_n \in R_{H_0}), \tag{5}$$

where $\mathbb{P}_{m_n^D(\cdot)}$ is the probability measure associated with the marginal distribution of Y_n obtained using $\pi^D(\theta)$. The power function in Eq. (5) expresses the probability of making a correct decision by rejecting H_0, when θ actually belongs to the subspace defined under the alternative hypothesis, where we can assume that it is distributed according to the design prior. Therefore, the corresponding SSD criterion requires to select the minimum n to achieve a desired level for $\eta_F^P(n,\pi^D)$.

Note that if $\pi^D(\theta)$ is chosen as a point mass distribution centred on θ^D, no uncertainty on the relevant design values is taken into account and the marginal distribution coincides with the sampling one. In this case, there is no difference between the frequentist power functions obtained under the conditional and the predictive approach.

3. Bayesian power functions and SSD methods

In the previous section, we have described how to select the sample size through power functions by assuming that a frequentist analysis will be performed at the end of the study. In both the frequentist conditional and predictive powers, the decision about the two hypotheses is based on the construction of the rejection region of H_0 of a classical test of fixed size α. A major limitation to the fully classical and the hybrid classical-Bayesian approaches previously introduced is the inability to incorporate past experience and information about the unknown parameter, as well as expert prior opinions. The use of a 'fully Bayesian approach' allows to take into account important knowledge and belief about θ when planning the study.

It is well known that the information available before starting the study can be expressed by introducing a prior distribution for θ, $\pi^A(\theta)$, which in this context is typically called *analysis prior* to distinguish it from the design prior. It is worth pointing out that $\pi^A(\theta)$ is the usual prior distribution employed in a Bayesian analysis: it formalizes pre-experimental knowledge, often represented by historical data, and subjective opinions of experts and is used to compute the posterior distribution of the parameter, $\pi_n^A(\theta|y_n) \propto f_n(y_n|\theta)\pi^A(\theta)$. Moreover, it is often chosen

as a non-informative distribution to avoid the inclusion of external evidence in the posterior inference.

Let us recall that, in general terms, a power function is defined as the probability of obtaining a significant result, i.e. a result that leads to the rejection of the null hypothesis. Then, to exploit this function as a useful tool to determine the optimal sample size, we need to compute it under the assumption that the alternative hypothesis is true. In practice, we have to consider a design scenario where the true θ belongs to Θ_1, so that the power function represents the probability of making a correct decision. Therefore, to define power functions from a Bayesian point of view, first of all we need to decide when we reject the null hypothesis in a Bayesian setting, that is we have to establish the condition for the 'Bayesian significance'. Following Spiegelhalter et al. [10], we define the result Y_n as 'significant from a Bayesian perspective' if the corresponding posterior probability that θ belongs to the alternative hypothesis is sufficiently large, that is if

$$\mathbb{P}_{\pi_n^A(\cdot|Y_n)}(\theta \in \Theta_1) > \lambda, \tag{6}$$

where $\mathbb{P}_{\pi_n^A(\cdot|Y_n)}$ denotes the probability measure associated with the posterior distribution of θ computed using the analysis prior and $\lambda \in (0, 1)$ represents a suitably specified threshold. Let us stress that, since we are dealing with a pre-experimental problem, the posterior probability in Eq. (6) is a random variable, depending on a random result that has not yet been observed. In order to construct Bayesian power functions, we need to compute the probability of obtaining a Bayesian significant result. Similar to what we have seen in the frequentist case, we can use two alternative distributions of the data, according to the approach we decide to adopt.

The *conditional approach* realizes the pre-experimental assumption that the alternative hypothesis is true, by fixing a design value $\theta^D \in \Theta_1$, which is considered relevant and important to detect. Then the sampling distribution of Y_n conditional on θ^D, $f_n(\cdot | \theta^D)$, is used to compute the probability of getting Bayesian significance. In this way, we obtain the *Bayesian conditional power*

$$\eta_B^C(n, \theta^D) = \mathbb{P}_{f_n(\cdot|\theta^D)}\left(\mathbb{P}_{\pi_n^A(\cdot|Y_n)}(\theta \in \Theta_1) > \lambda\right). \tag{7}$$

The *predictive approach*, instead, aims at avoiding the problem of local optimality in the SSD procedure by introducing a design prior for θ, $\pi^D(\theta)$, that accounts for additional uncertainty involved in the choice of the design values θ^D. Then, the prior predictive distribution of Y_n, $m_n^D(\cdot)$, is computed and used in place of the sampling distribution conditional on θ^D. This leads to the *Bayesian predictive power*

$$\eta_B^P(n, \pi^D) = \mathbb{P}_{m_n^D(\cdot)}\left(\mathbb{P}_{\pi_n^A(\cdot|Y_n)}(\theta \in \Theta_1) > \lambda\right). \tag{8}$$

Both the power functions in Eqs. (7) and (8) express the probability of rejecting H_0 under a Bayesian framework, assuming that the true θ actually belongs to H_1. In fact, we assume that θ is equal to a specific value under the alternative hypothesis (conditional approach) or that θ is in the specific subspace defined under the alternative hypothesis, where we can assume that it is distributed according to the design prior (predictive approach). The sample size

determination criteria, therefore, require to select the minimal sample size to ensure a sufficiently large level for $\eta_B^C(n, \theta^D)$ or $\eta_B^P(n, \pi^D)$. Moreover, note that, when the specified design prior distribution assigns the whole mass probability to θ^D, the two Bayesian power functions coincide, leading to the same optimal sample size.

4. SSD criteria according to the nature of the distribution of Y_n

In this section, we explicitly formalize the SSD criteria based on frequentist and Bayesian power functions, according to the nature of the random result Y_n. When Y_n has a continuous distribution, each of the power functions previously introduced shows a monotonically increasing behaviour as a function of n. In this case, the SSD criteria sensibly select the minimum sample size to guarantee the desired level of power, that is

$$n_F^C = \min\{n \in \mathbb{N} : \eta_F^C(n, \theta^D) > \gamma\}, \tag{9}$$

$$n_F^P = \min\{n \in \mathbb{N} : \eta_F^P(n, \pi^D) > \gamma\}, \tag{10}$$

$$n_B^C = \min\{n \in \mathbb{N} : \eta_B^C(n, \theta^D) > \gamma\}, \tag{11}$$

$$n_B^P = \min\{n \in \mathbb{N} : \eta_B^P(n, \pi^D) > \gamma\}, \tag{12}$$

for a conveniently chosen threshold $\gamma \in (0, 1]$. Let us remark that in the notation for the optimal sample sizes, as well as in the notations for the power functions, the subscripts are used to specify the approach (frequentist or Bayesian) adopted at the analysis stage. The superscripts, instead, indicate the appoach (conditional or predictive) used to represent the design expectations. An application of the criteria formalized above is provided by Gubbiotti and De Santis [11], where it is assumed that the statistic Y_n follows a normal distribution with mean equal to θ and known variance.

However, it may happen that $\eta_F^C(n, \theta^D)$, $\eta_F^P(n, \pi^D)$, $\eta_B^C(n, \theta^D)$ and $\eta_B^P(n, \pi^D)$ are not monotonically increasing functions of the sample size: this occurs when dealing with discrete distributions of Y_n. In these cases, the power functions show a basically increasing behaviour as a function of n, but with some small fluctuations. A suitable SSD criterion has to take into account this kind of behaviour. For instance, instead of selecting the smallest sample size that attains the condition of interest, it can be considered more appropriate to select the smallest sample size in such a way that the condition is fulfilled also for all the sample size values greater than it. Given a threshold $\gamma \in (0, 1)$, the corresponding SSD criteria are

$$n_F^C = \min\{n^* \in \mathbb{N} : \eta_F^C(n, \theta^D) > \gamma, \ \forall n \geq n^*\}, \tag{13}$$

$$n_F^P = \min\{n^* \in \mathbb{N} : \eta_F^P(n, \pi^D) > \gamma, \ \forall n \geq n^*\}, \tag{14}$$

$$n_B^C = \min\{n^* \in \mathbb{N} : \eta_B^C(n, \theta^D) > \gamma, \ \forall n \geq n^*\}, \tag{15}$$

$$n_B^P = \min\{n^* \in \mathbb{N} : \eta_B^P(n, \pi^D) > \gamma, \ \forall n \geq n^*\}. \tag{16}$$

In this way, it is possible to avoid the paradox of having the condition of interest fulfilled for the selected sample size, but not satisfied for some larger values of n any longer.

5. Single binomial proportion using exact methods

In this section, we focus on exact procedures for one-sample testing problem with binary response. For instance, in a clinical context, we could be interested in evaluating the efficacy of a new experimental treatment or drug that is received at the same dose by all the n patients enrolled in the trial. No comparisons with other therapies are involved. A binary response variable, which assumes value 1 if clinicians classify the patient as a responder to the therapy and 0 otherwise, is considered and, therefore, the parameter of interest θ is the true response rate (i.e. an unknown proportion). In these one-arm studies, θ is compared with a fixed target value, say θ_0, that should ideally represent the response rate for the current 'gold standard' therapy and that is typically obtained through historical data. Values of θ greater than θ_0 suggest that the experimental drug can be considered sufficiently effective and, therefore, the following hypotheses are considered

$$H_0 : \theta = \theta_0 \quad \text{and} \quad H_1 : \theta > \theta_0. \tag{17}$$

This kind of single-arm studies is typically conducted in phase II of clinical trials, whose primary goal is not to definitively assess the efficacy of new drugs, but to screen out those that are ineffective. In practice, in the clinical development process of a new drug, phase II aims at avoiding that not sufficiently promising treatments reach phase III, where randomized controlled trials, based on large patients groups, are generally conducted.

It is important to point out that the power functions based on exact procedures usually do not have explicit forms. Hence, exact formulas for sample size calculations cannot be obtained. However, it is possible to proceed numerically by evaluating the conditions of interest for different increasing or decreasing values of the sample size, until reaching the optimal one. In the following sections, we provide the expressions of the frequentist and Bayesian power functions for non-comparative studies with binary responses. The saw-toothed shape of the power curves as a function of n is shown and, hence, the conservative criteria illustrated in the previous section are adopted. All the graphical and numerical results have been obtained by using the R programming language [12].

5.1. Frequentist conditional power

In the statistical context described above, the number of responders out of the n patients treated with the new drug (i.e. the number of successes in n trials) is the natural statistic Y_n we have to consider and its sampling distribution is

$$f_n(y_n|\theta) = \text{bin}(y_n; n, \theta), \quad \text{for } y_n = 0, ..., n, \tag{18}$$

where $\text{bin}(\cdot; n, \theta)$ denotes the probability mass function of a binomial distribution of parameters n and θ.

Let us consider the two hypotheses in Eq. (17). For a fixed significance level α and assuming that H_0 is true, there exists a non-negative integer r between 0 and n such that

$$\sum_{i=r}^{n} \text{bin}(i; n, \theta_0) \leq \alpha \quad \text{and} \quad \sum_{i=r-1}^{n} \text{bin}(i; n, \theta_0) > \alpha. \tag{19}$$

Then, the rejection region at α level is $R_{H_0} = \{y_n \in \{0, 1, ..., n\} : y_n \geq r\}$, where the critical value r can be expressed in symbols by

$$r = \min\left\{ k \in \{0, 1, ..., n\} : \sum_{i=k}^{n} \text{bin}(i; n, \theta_0) \leq \alpha \right\}. \tag{20}$$

For a given design value θ^D, that has to be specified under the alternative hypothesis, the frequentist conditional power is provided by

$$\begin{aligned} \eta_F^C(n, \theta^D) &= \mathbb{P}_{f_n(\cdot | \theta^D)}(Y_n \in R_{H_0}) \\ &= \sum_{y_n=r}^{n} \text{bin}(y_n; n, \theta^D). \end{aligned} \tag{21}$$

In practice, $\eta_F^C(n, \theta^D)$ is obtained by the sum of the probabilities of the all the outcomes that belong to R_{H_0}, when we assume that the true θ is equal to the design value.

Figure 1 shows the behaviour of the frequentist conditional power as a function of n, when $\theta_0 = 0.2$, $\theta^D = 0.4$ and $\alpha = 0.05$. It is evident that $\eta_F^C(n, \theta^D)$ is not a monotonically increasing function of the sample size, because of the discrete nature of the sampling distribution of Y_n.
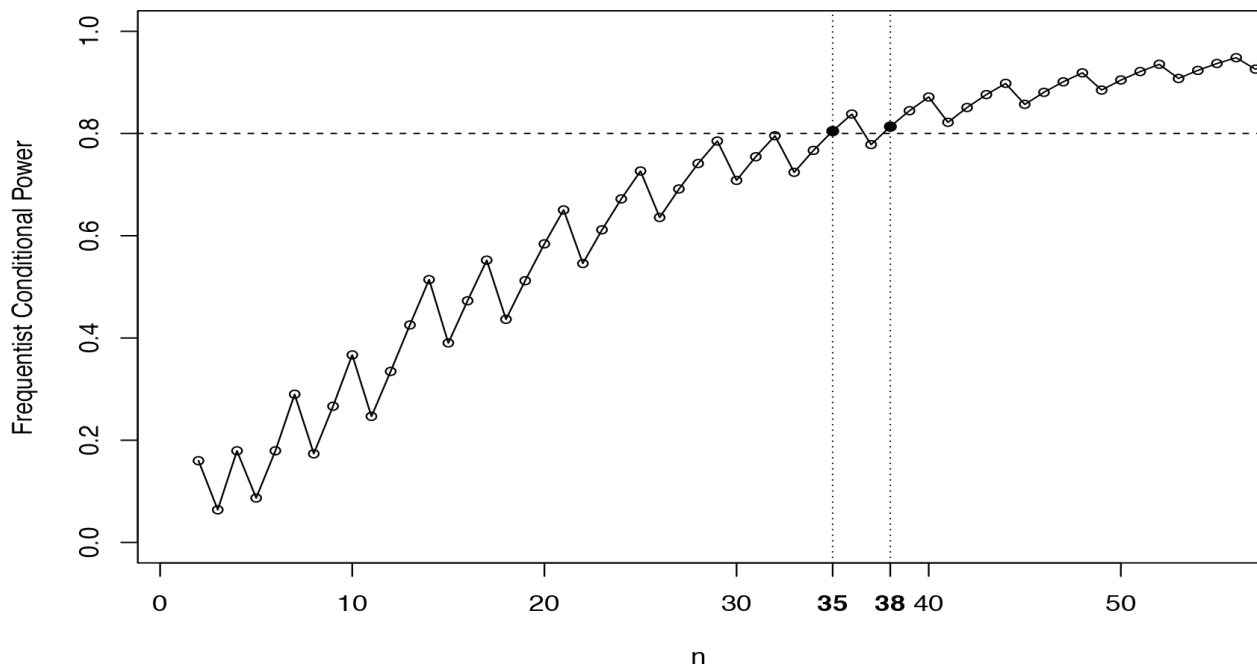

Figure 1. Behaviour of $\eta_F^C(n, \theta^D)$ as a function of n, when $\theta_0 = 0.20$, $\theta^D = 0.4$ and $\alpha = 0.05$.

The reasons for this saw-toothed behaviour can be clarified by the numerical results presented in **Table 1**. Here, for all the possible values of the sample size between 3 and 50, we provide not only the level of the frequentist conditional power used to obtain **Figure 1**, but also the corresponding critical value r and the actual value for the type I error probability. Obviously, this latter value is always below the fixed threshold 0.05. Note that whenever the sample size is increased by one unit, the corresponding critical value r may also increase or it may remain constant. In the second case, both the actual type I error rate and the conditional frequentist power grow up; otherwise, if also the critical value changes by one unit, they both get smaller. To help in reading the table, the colours white and grey are used alternately to highlight blocks

n	r	$\eta_F^C(n,\theta^D)$	Actual type I error rate	n	r	$\eta_F^C(n,\theta^D)$	Actual type I error rate
3	3	0.0640	0.0080	27	10	0.6913	0.0304
4	3	0.1792	0.0272	28	10	0.7412	0.0391
5	4	0.0870	0.0067	29	10	0.7853	0.0493
6	4	0.1792	0.0170	30	11	0.7085	0.0256
7	4	0.2898	0.0333	31	11	0.7546	0.0327
8	5	0.1737	0.0104	32	11	0.7954	0.0411
9	5	0.2666	0.0196	33	12	0.7242	0.0216
10	5	0.3669	0.0328	34	12	0.7669	0.0274
11	6	0.2465	0.0117	35	12	0.8048	0.0344
12	6	0.3348	0.0194	36	12	0.8380	0.0424
13	6	0.4256	0.0300	37	13	0.7783	0.0231
14	6	0.5141	0.0439	38	13	0.8136	0.0288
15	7	0.3902	0.0181	39	13	0.8446	0.0355
16	7	0.4728	0.0267	40	13	0.8715	0.0432
17	7	0.5522	0.0377	41	14	0.8219	0.0242
18	8	0.4366	0.0163	42	14	0.8509	0.0298
19	8	0.5122	0.0233	43	14	0.8762	0.0362
20	8	0.5841	0.0321	44	14	0.8979	0.0436
21	8	0.6505	0.0431	45	15	0.8570	0.0250
22	9	0.5460	0.0201	46	15	0.8807	0.0304
23	9	0.6116	0.0273	47	15	0.9012	0.0366
24	9	0.6721	0.0362	48	15	0.9187	0.0437
25	9	0.7265	0.0468	49	16	0.8851	0.0256
26	10	0.6358	0.0232	50	16	0.9045	0.0308

Table 1. Numerical calculations related to **Figure 1**: sample sizes, corresponding critical values, frequentist conditional power and actual values for the type I error rate, when θ_0 = 0.20, θ^D = 0.4 and α = 0.05.

of sample sizes with the same critical value: within each block both the power and the actual type I rate monotonically raise as n increases. But, in correspondence with the first sample size of the subsequent block, they both decrease. This determines the basically increasing behaviour of the power as a function of n, with some small fluctuations, which is represented in **Figure 1**. For additional discussion about the saw-toothed shape of the frequentist power function, the reader is referred to Chernick and Liu [13].

Now, the problem of which sample size we should select arises because of the non-monotonic behaviour of $\eta_F^C(n, \theta^D)$. If we set the desired threshold γ for the power equal to 0.8, we have that the smallest sample size that meets the power requirement is $n = 35$. At that sample size, the critical value is 12 and the power level is 0.8048. Then for $n = 36$, the critical value is still 12 and the power increases to 0.8380. However, the power drops below 0.8 to 0.7783, when $n = 37$, at which $r = 13$, and rises again over 0.8 when $n = 38$. Then $\eta_F^C(n, \theta^D)$ never decreases below 0.8 for sample sizes greater than 38. Therefore, instead of selecting the smallest n that attains the power condition, it can be more appropriate to consider the more conservative sample size criterion formalized in Section 4, according to which the optimal sample size is selected as

$$n_F^C = \min\{n^* \in \mathbb{N} : \eta_F^C(n, \theta^D) > \gamma, \ \forall n \geq n^*\}. \tag{22}$$

The criterion ensures that the power will not decrease below the desired threshold for any larger sample size: in our specific case, it consists in selecting $n = 38$, instead of $n = 35$.

5.2. Frequentist predictive power

In order to model uncertainty in the specification of the design value, we need to adopt the hybrid classical-Bayesian approach described previously. We introduce a beta design prior density for θ, $\pi^D(\theta) = \text{beta}(\theta; \alpha^D, \beta^D)$, that is used to obtain the prior predictive distribution of the data. It is well known that by averaging the binomial sampling $f_n(y_n | \theta)$ with respect to the beta design prior, we obtain the following marginal distribution

$$m_n^D(y_n) = \text{beta-bin}(y_n; \alpha^D, \beta^D, n), \ \text{for} \ y_n = 0, ..., n, \tag{23}$$

where beta-bin$(\cdot; \alpha^D, \beta^D, n)$ denotes the probability mass function of a beta-binomial distribution with parameters (α^D, β^D, n).

The design prior $\pi^D(\theta)$ can be elicited in many different ways. One useful possibility consists in (i) setting the prior mode equal to the fixed design value θ^D, which investigators would choose within the subset under H_1 when using the conditional approach, and (ii) regulating the concentration of the distribution around its mode according to the degree of uncertainty one wishes to express. This can be done by using for the hyperparameters of $\pi^D(\theta)$ the following expressions:

$$\alpha^D = n^D \theta^D + 1 \quad \text{and} \quad \beta^D = n^D(1 - \theta^D) + 1, \tag{24}$$

where θ^D is the prior mode and n^D is a design parameter that can be interpreted as *prior sample size*. The larger the n^D, the smaller the variance of the beta design prior. Therefore, we need to

increase n^D if we want to reduce uncertainty on the guessed values of θ. More specifically, if we set $n^D = \infty$, the design prior of θ assigns all the probability mass to θ^D: in this case, no uncertainty is involved and the marginal distribution of the data coincides with the sampling distribution conditional on θ^D. We thus must set $n^D < \infty$ to distinguish between conditional and predictive approaches. In particular, once a prior mode θ^D has been selected, the researcher can choose n^D by assuring a large level (say very close to 1) for $\mathbb{P}_{\pi^D(\cdot)}(\theta > \theta_0)$, that is the probability assigned by $\pi^D(\theta)$ to the event $\theta > \theta_0$. Let us assume, for instance, that $\theta_0 = 0.2$ and consider three possible choices for θ^D (i.e. 0.3, 0.4 and 0.5). For each of them, we compute the smallest n^D such that $\mathbb{P}_{\pi^D(\cdot)}(\theta > \theta_0)$ is about equal to 0.999, and the behaviour of the corresponding design priors is shown in **Figure 2(a)**. Clearly, if the prior mode approaches θ_0, we need to increase n^D to guarantee that $\mathbb{P}_{\pi^D(\cdot)}(\theta > \theta_0) \simeq 0.999$. Moreover, for a fixed prior mode θ^D, if we decided to decrease the value of n^D with respect to the one used in the graph, $\mathbb{P}_{\pi^D(\cdot)}(\theta > \theta_0)$ would decrease. In fact, n^D has been specified in order to express the minimum degree of prior enthusiasm about the efficacy of the treatment necessary to have the prior probability that θ exceeds the target θ_0 at least equal to the chosen level 0.999. An alternative way of proceeding consists in choosing n^D by ensuring a fixed level for the prior probability assigned to a symmetrical interval around the prior mode. For instance, if we set $\theta^D = 0.4$, we can find that 255, 111 and 60 are the values of n^D such that it is about equal to 0.999 the probability that $\pi^D(\theta)$ assigns to the intervals (0.3, 0.5), (0.25, 0.55) and (0.2, 0.6), respectively. The corresponding design prior distributions are shown in **Figure 2(b)**. It is important to point out that all the design densities, represented in both the graphs of **Figure 2**, express uncertainty in the suitable design value that it is worthwhile to consider when applying the SSD criteria based on power analysis. Thus, all the distributions assign a negligible probability to values of θ smaller than θ_0, which are those values specified under H_0.
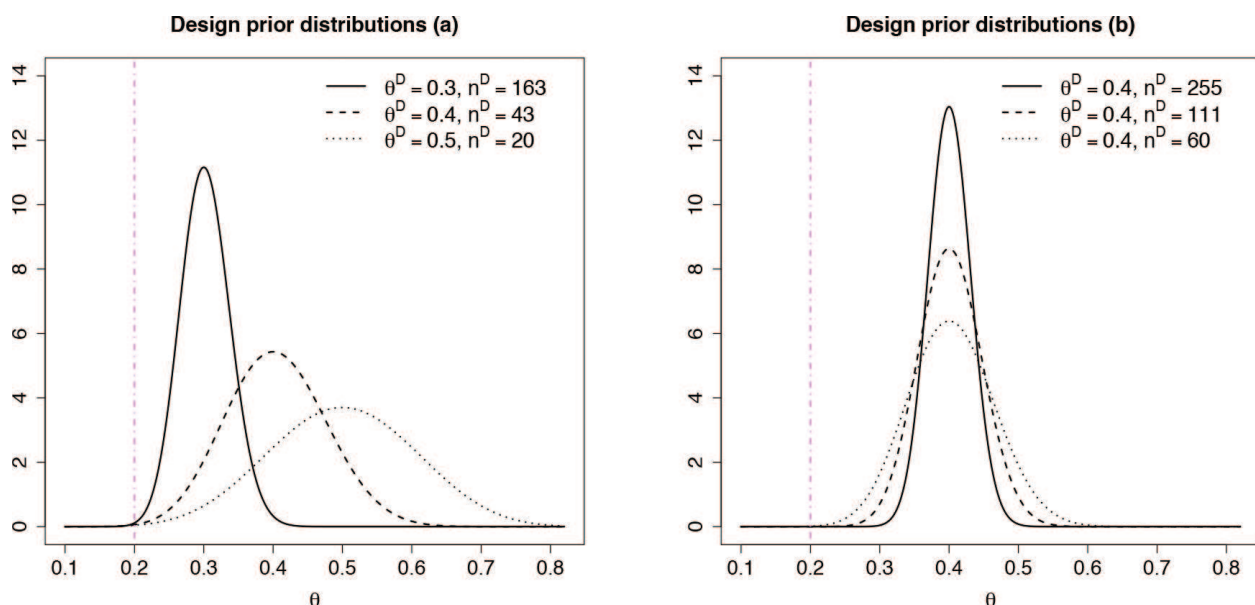

Figure 2. Possible choices of the design prior distribution, when $\theta_0 = 0.2$.

Once $\pi^D(\theta)$ has been specified, the frequentist predictive power can be obtained by computing the probability of rejecting the null hypothesis at α level with respect to $m_n^D(y_n)$. Hence, we have

$$\eta_F^P(n, \pi^D) = \mathbb{P}_{m_n^D(\cdot)}(Y_n \in R_{H_0})$$
$$= \sum_{y_n=r}^n \text{beta-bin}(y_n; \alpha^D, \beta^D, n), \tag{25}$$

where r is the critical value provided in Eq. (20). In practice $\eta_F^P(n, \pi^D)$ is given by the sum of the probabilities of the all the outcomes inside R_{H_0}, computed under a design scenario according to which the true θ belongs to the interval $(\theta_0, 1)$, where it is distributed according to the design prior density. Let us remark again that if the design prior is a point mass distribution on θ^D (i.e. $n^D = \infty$), we have that the frequentist power functions, conditional and predictive coincide.

Similarly to the frequentist conditional power, also the predictive one presents a saw-toothed shape as a function of n, since $m_n^D(y_n)$ is a discrete distribution. Therefore, we suggest to adopt the conservative approach previously described and to select

$$n_F^P = \min\{n^* \in \mathbb{N} : \eta_F^P(n, \pi^D) > \gamma, \ \forall n \geq n^*\}, \tag{26}$$

for a fixed desired threshold γ. **Figure 3** shows the behaviour of the frequentist predictive power as a function of n for different choices of the design prior, when $\theta_0 = 0.2$ and $\alpha = 0.05$. More specifically, we consider the three $\pi^D(\theta)$ plotted in **Figure 2(b)** that are all centred on $\theta^D = 0.4$, but with different degrees of concentrations regulated by the n^D value. In each graph, we highlight which is the optimal sample size obtained according to the criterion in Eq. (26) when $\gamma = 0.8$. Note that the larger the n^D, the smaller the degree of uncertainty we introduce through the design prior and, as a consequence, the smaller the optimal sample size. In fact, we obtain the optimal values 46, 42 and 39, for n^D equal to 60, 111 and 255, respectively. If we set $n^D = \infty$, we would retrieve the conditional criterion in Eq. (22), where no uncertainty is considered in specifying the design value, and the optimal n would be equal to 38 (see
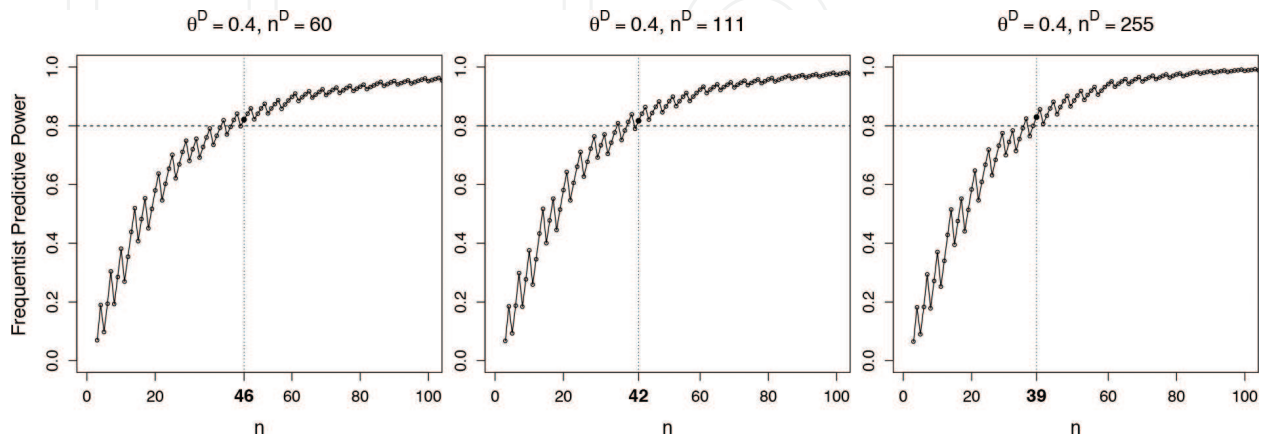

Figure 3. Behaviour of $\eta_F^P(n, \pi^D)$ as a function of n for different choices of the design prior distribution, when $\theta_0 = 0.2$ and $\alpha = 0.05$.

Figure 1). Moreover, let us fix again $\theta_0 = 0.2$, $\alpha = 0.05$ and $\gamma = 0.8$ and consider the three design prior distributions in **Figure 2(a)**, which are characterized by different prior modes. The evident difference between the prior scenarios represented by these design priors clearly affects the optimal sample size: we obtain the optimal values 157, 46 and 23, for $(\theta^D, n^D) = (0.3, 163)$, $(\theta^D, n^D) = (0.4, 43)$ and $(\theta^D, n^D) = (0.5, 20)$, respectively.

5.3. Bayesian conditional power

When we decide to adopt a Bayesian approach to establish the statistical significance of the result, we need to introduce an analysis prior distribution for θ. In our specific case, it is computationally convenient to specify a beta analysis prior, $\pi^A(\theta) = \text{beta}(\theta; \alpha^A, \beta^A)$: in this way, from conjugate analysis we obtain that the corresponding posterior distribution is still a beta density with updated parameters,

$$\pi_n^A(\theta|y_n) = \text{beta}(\theta; \alpha^A + y_n, \beta^A + n - y_n). \tag{27}$$

Through $\pi^A(\theta)$, the researcher can incorporate in the SSD procedure pre-experimental knowledge, as well as sceptical or enthusiastic expert prior opinions about the efficacy of the experimental treatment. However, one of the most common ways of proceeding is to choose a non-informative—or based on very weak information–density, to let the posterior distribution be based almost entirely on the evidence in the data. We could, therefore, specify $\pi^A(\theta) = \text{beta}(\theta; 1, 1)$ or consider the non-informative Jeffreys prior. Alternatively, if we want to use informative analysis prior distributions, we can express the hyperparameters in terms of the prior mode θ^A and the prior sample size n^A, that is

$$\alpha^A = n^A\theta^A + 1 \quad \text{and} \quad \beta^A = n^A(1 - \theta^A) + 1. \tag{28}$$

In this way, for instance, it is possible to express scepticism or optimism about large treatment effects by setting θ^A less or higher than the target θ_0, respectively. Obviously, when $\theta^A < \theta_0$, the larger the n^A, the larger the degree of scepticism we wish to express; while, when $\theta^A > \theta_0$ larger values of n^A are used to increase the degree of enthusiasm we desire to take into account. However, the value $n^A = 1$ is often used to have a weakly informative prior distribution. The upper panel of **Figure 4** shows three possible choices for the analysis prior when $\theta_0 = 0.2$. These distributions are obtained by fixing the prior mode θ^A and, then, selecting n^A so that $\mathbb{P}_{\pi^A(\cdot)}(\theta > \theta_0)$ (i.e. the probability assigned by $\pi^A(\theta)$ to the event $\theta > \theta_0$) is about equal to a desired level. More specifically, we have considered (i) a sceptical prior mode $\theta^A = 0.1$ and $\mathbb{P}_{\pi^A(\cdot)}(\theta > \theta_0) \simeq 0.4$, (ii) a neutral prior mode $\theta^A = 0.2$ and $\mathbb{P}_{\pi^A(\cdot)}(\theta > \theta_0) \simeq 0.6$ and finally (iii) an enthusiastic prior mode $\theta^A = 0.3$ and $\mathbb{P}_{\pi^A(\cdot)}(\theta > \theta_0) \simeq 0.8$. The corresponding values of n^A are 7, 14 and 4, respectively. These densities will be used to illustrate how the optimal sample sizes based on Bayesian powers are affected by the information formalized through the analysis priors.

The random result Y_n is defined as 'significant' from a Bayesian perspective, if the corresponding posterior probability that $\theta > \theta_0$ is sufficiently large. In symbols, we decide to reject the null hypothesis, on the basis of the result Y_n, if the following condition is satisfied.
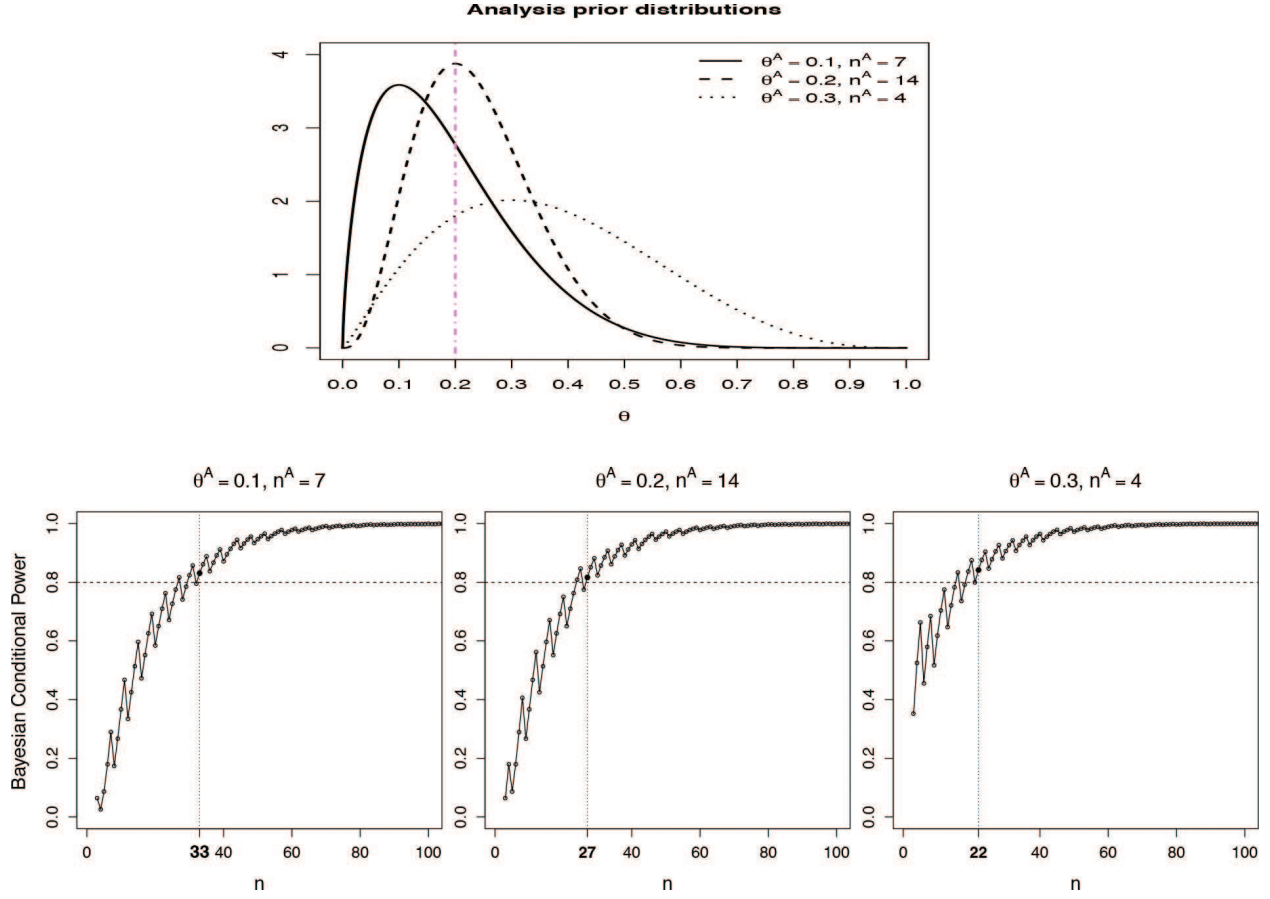
Figure 4. Upper panel: possible choices of the analysis prior distribution, when $\theta_0 = 0.2$. Lower panel: behaviour of $\eta_C^B(n, \theta^D)$ as a function of n for each of the analysis prior distributions represented in the upper panel, when $\theta_0 = 0.2$, $\theta^D = 0.4$ and $\lambda = 0.9$.

$$\mathbb{P}_{\pi_n^A(\cdot|Y_n)}(\theta > \theta_0) > \lambda, \tag{29}$$

where $\mathbb{P}_{\pi^A(\cdot|Y_n)}$ is the probability measure associated with the posterior distribution in Eq. (27) and $\lambda \in (0, 1)$ is a pre-specified threshold. It is worth noting that, for a given value of n, the posterior quantity $\mathbb{P}_{\pi_n^A(\cdot|Y_n)}(\theta > \theta_0)$ is an increasing function of Y_n. As a consequence, we can find a non-negative integer $\tilde{r}$ between 0 and n, such that

$$\mathbb{P}_{\pi_n^A(\cdot|\tilde{r})}(\theta > \theta_0) > \lambda \quad \text{and} \quad \mathbb{P}_{\pi_n^A(\cdot|\tilde{r}-1)}(\theta > \theta_0) \leq \lambda, \tag{30}$$

and we can claim that H_0 is rejected if the observed number of responders y_n is equal to or greater than $\tilde{r}$. In practice, $\tilde{r}$ represents the smallest number of successes such that the condition for the Bayesian significance is satisfied, and in symbols it can be expressed by

$$\tilde{r} = \min\left\{ k \in \{0, 1, ..., n\} : \mathbb{P}_{\pi_n^A(\cdot|k)}(\theta > \theta_0) > \lambda \right\}. \tag{31}$$

By considering a fixed design value θ^D greater than θ_0, the Bayesian conditional power is therefore obtained as

$$
\begin{aligned}
\eta_B^C(n, \theta^D) &= \mathbb{P}_{f_n(\cdot|\theta^D)}\left(\mathbb{P}_{\pi_n^A(\cdot|Y_n)}(\theta > \theta_0) > \lambda\right) \\
&= \sum_{y_n=\tilde{r}}^{n} \text{bin}(y_n; n, \theta^D).
\end{aligned}
\tag{32}
$$

Essentially, it is given by the sum of the probabilities of all the Bayesian significant results, computed assuming that the true θ is equal to θ^D.

Since we are dealing with discrete data, also this power function is not monotonically increasing as a function of n. Let us assume that $\theta_0 = 0.20$, $\theta^D = 0.4$ and $\lambda = 0.9$. The detailed calculations shown in **Table 2** can help to understand why $\eta_B^C(n, \theta^D)$ has the typical saw-toothed behaviour. For each sample size between 3 and 50, the table provides the corresponding value of $\tilde{r}$, the level of the Bayesian conditional power and the posterior probability that θ exceeds θ_0 conditional on the result $\tilde{r}$. Clearly, these latter values are always larger than the threshold λ that is 0.9. The white and grey colours are used alternately to highlight blocks of sample sizes with the same value of $\tilde{r}$ associated. When the sample size grows, but $\tilde{r}$ remains constant, $\mathbb{P}_{\pi_n^A(\cdot|\tilde{r})}(\theta > \theta_0)$ decreases, while $\eta_B^C(n, \theta^D)$ increases. However, when both n and $\tilde{r}$ are simultaneously increased by one unit, $\mathbb{P}_{\pi_n^A(\cdot|\tilde{r})}(\theta > \theta_0)$ jumps up, while the Bayesian power drops.

Because of the saw-toothed nature of the power curve, for a fixed threshold γ, the optimal sample size is selected using the conservative criterion, that is

$$
n_B^C = \min\left\{n^* \in \mathbb{N} : \eta_B^C(n, \theta^D) > \gamma, \ \forall n \geq n^*\right\}.
\tag{33}
$$

The lower panel of **Figure 4** shows the behaviour of the Bayesian conditional power as a function of n for each of the three analysis prior density plotted in the upper panel, when $\theta_0 = 0.2$, $\theta^D = 0.4$ and $\lambda = 0.9$. In each graph, it is indicated the optimal sample size according to the criterion in Eq. (33) for $\gamma = 0.8$. As expected, as we move from sceptical prior opinions towards more enthusiastic beliefs about the efficacy of the experimental treatment, the required sample size decreases.

5.4. Bayesian predictive power

Besides introducing pre-experimental information, if we also wish to model uncertainty on the design value, we have to consider the Bayesian predictive power. Therefore, as described in Section 5.3, we elicit an analysis prior distribution to obtain the beta posterior density $\pi_n^A(\theta|y_n)$. Moreover, following the indications provided in Section 5.2, we introduce a design prior distribution to construct the marginal distribution $m_n^D(y_n)$.

The Bayesian predictive power is computed by adding the probabilities of all the Bayesian significant results, computed under the design scenario expressed through the design prior. Thus, we have

$$
\begin{aligned}
\eta_B^P(n, \pi^D) &= \mathbb{P}_{m_n^D(\cdot)}\left(\mathbb{P}_{\pi_n^A(\cdot|Y_n)}(\theta > \theta_0) > \lambda\right) \\
&= \sum_{y_n=\tilde{r}}^{n} \text{beta-bin}(y_n; \alpha^D, \beta^D, n),
\end{aligned}
\tag{34}
$$

n	$\tilde{r}$	$\eta_B^C(n, \theta^D)$	$\mathbb{P}_{\pi_n^A(\cdot\mid\tilde{r})}(\theta > \theta_0)$	n	$\tilde{r}$	$\eta_B^C(n, \theta^D)$	$\mathbb{P}_{\pi_n^A(\cdot\mid\tilde{r})}(\theta > \theta_0)$
3	3	0.0640	0.9263	27	9	0.8161	0.9077
4	4	0.0256	0.9703	28	10	0.7412	0.9464
5	4	0.0870	0.9558	29	10	0.7853	0.9354
6	4	0.1792	0.9377	30	10	0.8237	0.9230
7	4	0.2898	0.9159	31	10	0.8566	0.9092
8	5	0.1737	0.9618	32	11	0.7954	0.9460
9	5	0.2666	0.9476	33	11	0.8310	0.9356
10	5	0.3669	0.9304	34	11	0.8617	0.9239
11	5	0.4672	0.9102	35	11	0.8877	0.9110
12	6	0.3348	0.9559	36	12	0.8380	0.9460
13	6	0.4256	0.9422	37	12	0.8667	0.9362
14	6	0.5141	0.9260	38	12	0.8911	0.9252
15	6	0.5968	0.9075	39	12	0.9118	0.9131
16	7	0.4728	0.9518	40	13	0.8715	0.9464
17	7	0.5522	0.9388	41	13	0.8945	0.9371
18	7	0.6257	0.9237	42	13	0.9140	0.9267
19	7	0.6919	0.9065	43	13	0.9305	0.9153
20	8	0.5841	0.9491	44	13	0.9441	0.9028
21	8	0.6505	0.9367	45	14	0.9164	0.9381
22	8	0.7102	0.9226	46	14	0.9320	0.9284
23	8	0.7627	0.9067	47	14	0.9450	0.9176
24	9	0.6721	0.9474	48	14	0.9558	0.9059
25	9	0.7265	0.9357	49	15	0.9336	0.9394
26	9	0.7745	0.9225	50	15	0.9460	0.9301

Table 2. Numerical calculations to explain the saw-toothed behaviour of $\eta_B^C(n, \theta^D)$ as a function of n: sample sizes, the corresponding value of $\tilde{r}$, the Bayesian conditional power and the posterior probability that $\theta > \theta_0$ when the observed result is equal to $\tilde{r}$ successes, for $\theta_0 = 0.20$, $\theta^D = 0.4$ and $\lambda = 0.9$.

where $\tilde{r}$ is given in Eq. (31). Obviously, also $\eta_B^P(n, \pi^D)$ shows the typical saw-toothed behaviour as a function of n, because of the discrete nature of the beta-binomial marginal distribution of y_n. Therefore, given a desired threshold γ and according to the suitable conservative approach previously used, we select the optimal sample size as

$$n_B^P = \min\{n^* \in \mathbb{N}: \eta_B^P(n, \pi^D) > \gamma, \ \forall n \geq n^*\}. \tag{35}$$

θ^D	n^D	$\theta^A = 0.1$	$\theta^A = 0.2$	$\theta^A = 0.3$
		$n^A = 7$	$n^A = 14$	$n^A = 4$
(a) Design prior distributions in **Figure 2(a)**				
0.3	163	120	109	94
0.4	43	37	31	22
0.5	20	21	18	11
(b) Design prior distributions in **Figure 2(b)**				
0.4	60	37	31	22
0.4	111	33	31	22
0.4	255	33	27	22

Table 3. n_B^P for different choices of the analysis and the design priors, when $\theta_0 = 0.2$ and $\lambda = 0.9$.

In **Table 3** we provide the values of n_B^P, for different choices of the analysis and the design prior densities. More specifically, we consider the three analysis priors plotted in the upper panel of **Figure 4** and the design prior distributions represented in both the panels of **Figure 2**, when $\theta_0 = 0.2$ and $\lambda = 0.9$. Similarly to what we have seen for the Bayesian conditional power, the sample sizes obtained under the sceptical analysis prior are uniformly larger than those obtained under the more enthusiastic distributions. As regard the impact of the design priors, it is straightforward to see that the stronger the degree of uncertainty on the appropriate design value expressed by $\pi^D(\theta)$, the larger the required sample size. For instance, for a fixed prior mode of the design prior, n_B^P increases as n^D get smaller (see **Table 3(b)**, where $\theta^D = 0.4$). However, let us note that more evident changes in the sample size can be appreciated when we compare the effects of design priors based on different prior modes (see the results in **Table 3(a)**, where the design priors represent very distant design scenarios).

These Bayesian predictive SSD procedures, which include the conditional ones as a special case, have been exploited in Ref. [8] to construct single-arm two-stage design for phase II of clinical trials based on binary data. In Ref. [14], instead, an extension to the randomized case has been presented, while in Ref. [15] the same procedures have been implemented by adding the possibility of taking into account uncertainty in the historical response rate.

6. Conclusions

Especially in clinical research, the pre-experimental power analysis is one of the most commonly used methods for sample size calculations. It is tacitly implied that the power function is constructed under a frequentist framework. However, it is possible to introduce Bayesian concepts in the power analysis to provide more flexibility to the sample size determination process.

When the power function is used as a tool to obtain the appropriate sample size, the general idea is to ensure a large probability of correctly rejecting the null hypothesis H_0, when it is actually false because the true θ belongs to H_1. Therefore, the conjecture that the alternative

hypothesis is true represents an essential element of the method. It can be realized by assuming that the true θ is equal to a fixed design value θ^D, suitably selected inside H_1 (conditional approach); alternatively, we can introduce uncertainty on the guessed design value by introducing a design prior distribution that assigns negligible probability to values of θ under H_0 (predictive approach). Moreover, the decision about the rejection of H_0 can be made under a frequentist framework or by performing a Bayesian analysis. In the latter case, it is possible to incorporate in the methodology pre-experimental information possibly available through the specification of an analysis prior distribution. By combining frequentist and Bayesian procedures of analysis, with both the conditional and predictive approaches, we obtain the four power functions described in this chapter. Let us remark that the Bayesian predictive power is the one that allows to add more flexibility to the sample size calculations. At the same time, it let the researcher take into account prior knowledge, as well uncertainty on the design value. However, no design uncertainty can be involved by considering a point-mass design distribution. On the other hand, if no information is available, it is possible to elicit a non-informative analysis prior and let the analysis be based entirely on the data.

Author details

Valeria Sambucini

Address all correspondence to: valeria.sambucini@uniroma1.it

Department of Statistical Sciences, Sapienza Università di Roma, Sapienza, Italy

References

[1]　Ryan TP. Sample Size Determination and Power. Haboken: Wiley; 2013

[2]　Chow SC, Wang H, Shao J. Sample Size Calculations in Clinical Research. 2nd ed. Boca Raton: Chapman and Hall/CRC; 2008

[3]　Julious SA. Sample Sizes for Clinical Trials. Boca Raton: Chapman and Hall/CRC; 2010.

[4]　Wang F, Gelfand AE. A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. Statistical Science. 2002;**17**(2):193-208. DOI: 10.1214/ss/1030550861

[5]　De Santis F. Sample size determination for robust Bayesian analysis. Journal of the American Statistical Association. 2006;**101**(473):278-291. DOI: 10.1198/016214505000000510

[6]　Sahu SK, Smith TMF. A Bayesian method of sample size determination with practical applications. Journal of the Royal Statistical Society: Series A. 2006;**169**:235-253. DOI: 10.1111/j.1467-985X.2006.00408.x

[7] Brutti P, De Santis F, Gubbiotti S. Robust Bayesian sample size determination in clinical trials. Statistics in Medicine. 2008;**27**(13):2290-2306. DOI: 10.1002/sim.3175

[8] Sambucini V. A Bayesian predictive two-stage design for phase II clinical trials. Statistics in Medicine. 2008;**27**(8):1199-1224. DOI: 10.1002/sim.3021

[9] Sambucini V. A Bayesian predictive strategy for an adaptive two-stage design in phase II clinical trials. Statistics in Medicine. 2010;**29**(13):1430-1442. DOI: 10.1002/sim.3800

[10] Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian Approaches to Clinical Trials and Health-Care Evaluation. New York: Wiley; 2004

[11] Gubbiotti S, De Santis F. Classical and Bayesian power functions: Their use in clinical trials. Biomedical Statistics and Clinical Epidemiology. 2008;**2**(3):201-211. DOI: 10.1198/016214505000000510

[12] R Core Team. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. 2016. Available from: http://www.R-project.org

[13] Chernick MR, Liu CY. The saw-toothed behavior of power versus sample size and software solutions: Single binomial proportion using exact methods. The American Statistician. 2002;**56**(2):149-155. DOI: 10.1198/000313002317572835

[14] Cellamare M, Sambucini V. A randomized two-stage design for phase II clinical trials based on a Bayesian predictive approach. Statistics in Medicine. 2015;**34**(6):1059-1078. DOI: 10.1002/sim.6396

[15] Matano F, Sambucini V. Accounting for uncertainty in the historical response rate of the standard treatment in single-arm two-stage designs based on Bayesian power functions. Pharmaceutical Statistics. 2016;**15**(6):517-530. DOI: 10.1002/pst.1788