Controlling the Conflict Between Frequentist and Bayesian Estimators

P. Brutti[1] , F. De Santis[2] and S. Gubbiotti[2]

[1] *Dipartimento di Scienze Economiche ed Aziendali, LUISS Guido Carli Roma*

[2] *Dipartimento di Scienze Statistiche, Sapienza Università di Roma*

Abstract

In the presence of prior information on an unknown parameter of a statistical model, Bayesian and frequentist estimates based on the same observed data do not coincide. However, in many standard parametric problems, their discrepancy tends to be reduced as the sample size increases. In this paper we consider the pre-experimental design problem of selecting sample sizes that guarantee large probabilities of observing a small discrepancy between Bayesian and frequentist point estimates of a parameter. We propose a Bayesian predictive approach and we illustrate some examples using the normal model. We argue that these examples may be discussed even in introductory-level courses in Bayesian inference.

Keywords: Sample Size Determination, Clinical Trials, Discrepancy between estimators, Predictive Approach.

1 Introduction

Bayesian statistics offers the theoretical framework for combining experimental and extra-experimental information on phenomena under study. As a consequence, Bayesian procedures for inference on an unknown parameter of a statistical model take into account both experimental data and information on the parameter incorporated in the so-called prior distribution.

In the presence of pre-experimental information, frequentist and Bayesian procedures, such as point or interval estimates based on the same observed sample -in general- do not coincide. However, in many standard parametric problems, the

discrepancy between frequentist and Bayesian procedures is rather limited when sampling information dominates the prior distribution. Furthermore, this conflict tends to disappear as the sample size increases. Therefore, for sufficiently large sample sizes, frequentist procedures may provide good approximations of Bayesian methods.

A paradigmatic example is the estimation problem for the expected value of a normal random variable. In this case (see Section 3.1 for technical details), given n observations from independent and identically distributed (i.i.d.) normal random variables, the standard Bayesian estimate of θ is a linear combination of the sampling mean, $\bar{x}_n$, and of a prior guess on the parameter, μ_A:

$$\omega_n \bar{x}_n + (1 - \omega_n)\mu_A, \qquad \omega_n \in (0, 1), \tag{1}$$

where the value of ω_n in the above formula tends to one as n diverges. Therefore, for a sufficiently large sample size, $\omega_n \bar{x}_n + (1 - \omega_n)\mu_A \simeq \bar{x}_n$, that is the Bayesian estimate (1) is well approximated by the sample mean.

In most of introductory books on Bayesian inference [see, for instance, Berger (1985), Bernardo and Smith (1994), Gelman et al. (2004), Lee (2004), O'Hagan and Forster (2004), Robert (2001)], the progressive reduction of conflict between Bayesian and frequentist procedures is typically showed only as a limiting result. However, before observing the data, any measure of conflict between estimators is a random variable and one might desire to choose the sample size so to have a large probability to observe a small discrepancy between Bayesian and frequentist procedures. This sample size determination (SSD in the following) problem is the topic of the present paper, that is organized as follows. In Section 2.1, we introduce a measure of discrepancy, D_n, between a frequentist and a Bayesian estimator. In Section 2.2 we define sample size determination criteria based on the predictive distribution of D_n. The basic idea is to choose the minimal sample size necessary to have a large probability that D_n is sufficiently small. We derive the explicit expression of D_n, its predictive cumulative distribution function (cdf) and expected value for the normal model with conjugate priors, assuming both known (Section 3.1) and unknown (Section 3.2) variance. We illustrate some examples for these basic uniparametric models that may be discussed even in introductory-level courses in Bayesian statistical inference. In Section 4 we apply the methodology to an illustrative example based on a superiority clinical trial.

2 Methodology

2.1 Discrepancy between estimators

Let $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)$ be a random sample from a probability distribution $f_n(\cdot|\theta)$, where θ is an unknown real-valued parameter that belongs to the parameter space, Θ. Following the Bayesian inferential approach, we assume that θ is a random variable. For simplicity, assume that $\Theta \subseteq \mathbb{R}$. Let $\mathbf{x}_n = (x_1, x_2, \ldots, x_n)$ be an observed sample, $\pi_A(\cdot)$ the prior density function of θ, $f_n(\mathbf{x}_n|\theta)$ its likelihood function and

$$\pi(\theta|\mathbf{x}_n) = \frac{f_n(\mathbf{x}_n|\theta)\pi_A(\theta)}{\int_\Theta f_n(\mathbf{x}_n|\theta)\pi_A(\theta)d\theta}$$

its posterior distribution. We will refer to π_A as to the *analysis-prior*. It models pre-experimental knowledge/uncertainty on θ taken into account in posterior analysis. For instance, if θ represents the unknown effect of a new clinical treatment, π_A may model all the pre-trial information we have on the parameter, based on subjective opinion of experts and/or historical data.

We denote a Bayesian estimator of θ as $\hat{\theta}_B(\mathbf{X}_n)$ whereas $\hat{\theta}_F(\mathbf{X}_n)$ is a generic consistent frequentist estimator. In this article we consider the posterior expectation of the parameter θ, $\mathrm{E}(\theta|\mathbf{X}_n) = \int_\Theta \theta\pi(\theta|\mathbf{x}_n)d\theta$, as $\hat{\theta}_B$ and the maximum likelihood estimator (MLE) as $\hat{\theta}_F$. Nevertheless, all the following can be extended to other Bayesian and frequentist point estimators.

Let $D_n(\mathbf{X}_n)$ be a measure of discrepancy between $\hat{\theta}_B$ and $\hat{\theta}_F$. Specifically, we consider the standard squared difference between estimators:

$$D_n(\mathbf{X}_n) = [\hat{\theta}_B(\mathbf{X}_n) - \hat{\theta}_F(\mathbf{X}_n)]^2. \tag{2}$$

Before observing the data, $\hat{\theta}_B$, $\hat{\theta}_F$ and D_n are random variables (functions of $\mathbf{X}_n$). With no loss of generality we assume that, as n tends to infinity, D_n converges in probability to zero. Therefore, as the sample size increases, Bayesian and frequentist point estimators tends to become closer and closer.

2.2 Sample size determination

Turning to SSD, we want to determine the minimal sample size such that the probability of observing a small discrepancy between $\hat{\theta}_B$ and $\hat{\theta}_F$ is sufficiently large. As for any other design problem in Bayesian inference, this probability can be evaluated using two alternative distributions for the data [see, for instance, Chaloner

and Verdinelli (1995)]. The *conditional* approach prescribes the use of the sampling distribution $f_n(\cdot|\theta)$, with $\theta = \mu_D$, a "design value" for the unknown parameter. This method takes into account a single value for θ at the design phase of the analysis and leads to sample sizes that are optimal conditionally on this chosen value μ_D. The *predictive* approach implies the use of the predictive distribution

$$m_D(\mathbf{x}_n) = \int_\Theta f_n(\mathbf{x}_n|\theta)\pi_D(\theta)d\theta,$$

where π_D (*design-prior*) is a density function that accounts for a range of plausible design values of θ. The corresponding sample sizes are optimal conditionally on the chosen design-prior distribution π_D. Note that, if π_D is a point-mass prior on μ_D, the predictive and the conditional approach coincide. Hence, we adopt here notation of the predictive method, which includes the conditional approach as a special case.

At this point it is important to note that we are now using two distinct priors: π_D for determination of the marginal distribution of the data and π_A for determination of the posterior distribution of θ. Several previous articles consider distinct design and analysis-priors as we do here. Among these, see Tsutakawa (1972), Etzioni and Kadane (1993), Joseph, du Berger and Belisle (1997), O'Hagan and Stevens (2001), Wang and Gelfand (2002), Sahu and Smith (2006), De Santis (2006) and Sambucini (2008). The necessity of making this distinction is self-evident, for instance, in clinical experiments. As an example, let us consider a superiority trial, whose goal is to prove superiority of a new treatment over a standard therapy. In this case, superiority is conjectured and the design-prior must reflect the optimism of the researcher. However, he/she may decide to be neutral in reporting posterior results, as often required by regulatory agencies. Under these circumstances, we would choose a design-prior (π_D) centered on a value greater than a significative effect-level we wish to assess. At the same time, we would assume an analysis-prior (π_A) centered on zero (to express neutrality) and relatively noninformative, so to let the data drive the analysis.

Before moving to SSD criteria, a technical remark is in order. In fact, the marginal distribution m_D exists only if π_D is a proper proper distribution. Conversely, π_A can also be a standard noninformative and improper prior.

The SSD criteria that we consider in this article are based on the simple idea of selecting the smallest n so that the predictive distribution of D_n is sufficiently concentrated on small values. Let

$$p_n(d) = \mathbb{P}(D_n \leq d), \qquad d > 0 \tag{3}$$

be the cumulative distribution function of D_n. For any pair of chosen values $d > 0$ and $\gamma \in (0, 1)$, the optimal sample size $n_p^*(d, \gamma)$ is the minimum n such that $p_n(d)$ is larger than γ:

$$n_p^*(d, \gamma) = \min\{n \in \mathbb{N} : p_n(d) > \gamma\}, \qquad \gamma \in (0, 1). \tag{4}$$

Alternatively, let

$$e_n = \mathbb{E}(D_n) \tag{5}$$

be the expected values of D_n computed with respect to the marginal distribution m_D. For a given value of $d > 0$, the optimal sample size n_e^* is:

$$n_e^*(d) = \min\{n \in \mathbb{N} : e_n \leq d\}. \tag{6}$$

3 Results for the normal model

3.1 Known variance.

Let $\mathbf{X}_n$ be a random sample from a $N(\theta, \sigma^2)$ distribution. The MLE of θ is $\hat{\theta}_F = \bar{x}_n$. Assume for θ a normal analysis-prior density $\pi_A(\theta) = N(\theta|\mu_A, \sigma^2/n_A)$, where, in general, $N(\cdot|a, b)$ denotes the density function of a normal random variable of parameters (a, b) and where n_A is given the standard interpretation of "prior sample size". The normal density for θ is said *conjugate* to the model f_n since the resulting posterior distribution is still a normal density. In fact, from standard results on conjugate analysis for the normal model [Bernardo and Smith (1994, p. 439)], the posterior distribution of θ is

$$\pi(\theta|\mathbf{x}_n) = N\left(\theta \left| \frac{n\bar{x}_n + n_A\mu_A}{n + n_A}, \sigma^2 \frac{1}{n + n_A}\right.\right).$$

Hence

$$\hat{\theta}_B = \frac{n\bar{x}_n + n_A\mu_A}{n + n_A} = \omega_n\bar{x}_n + (1 - \omega_n)\mu_A, \qquad \omega_n = \frac{n}{n + n_A} = \frac{1}{1 + \frac{I_\pi}{I_n}}, \tag{7}$$

where $I_n = n/\sigma^2$ is the observed information and $I_\pi = n_A/\sigma^2$ is the prior precision (i.e. the inverse of the prior variance). The ratio $I_\pi/I_n = n/n_A$ determines the impact of the prior mean μ_A and of the MLE $\bar{x}_n$ on the weighted average that define $\hat{\theta}_B$. In fact, the larger the ratio I_π/I_n, the closer $\hat{\theta}_B$ to μ_A. Conversely, the smaller the ratio, the closer $\hat{\theta}_B$ to $\hat{\theta}_F$.

Suppose now that

$$\pi_D(\theta) = N(\theta|\mu_D, \sigma^2/n_D), \tag{8}$$

where μ_D, n_D and σ^2 are known constants. In this case, the predictive density function of $\bar{x}_n$ is

$$m_D(\bar{x}_n) = N(\bar{x}_n|\mu_D, \psi_n^2), \qquad \text{where} \qquad \psi_n^2 = b_n\sigma^2 \quad \text{and} \quad b_n = \frac{n+n_D}{nn_D}. \tag{9}$$

We now give the explicit expression for D_n, its predictive expected value (e_n) and cumulative distribution function (p_n).

Result 1. Assume that $X_i|\theta$ has density $N(\cdot|\theta, \sigma^2/n)$, $i = 1, 2, \ldots, n$, and that $\pi_A(\theta) = N(\theta|\mu_A, \sigma^2/n_A)$, Then,

$$D_n = a_n^2(\bar{X}_n - \mu_A)^2, \qquad \text{where} \qquad a_n = \frac{n_A}{n+n_A}. \tag{10}$$

Furthermore, assuming $\pi_D(\theta) = N(\theta|\mu_D, \sigma^2/n_D)$, it follows that

$$\begin{aligned} e_n &= a_n^2[b_n\sigma^2 + \delta^2], \\ p_n(d) &= \Phi\left[b_n^{-1/2}(\delta + a_n^{-1}d^{1/2})\,\sigma^{-1}\right] - \Phi\left[b_n^{-1/2}(\delta - a_n^{-1}d^{1/2})\,\sigma^{-1}\right], \end{aligned}$$

where $\delta = \mu_D - \mu_A$.

Proof. The expression of e_n is determined noting that, under (9), $\mathbb{E}[\bar{X}_n] = \mu_D$ and $\mathbb{E}[\bar{X}_n^2] = \mu_D^2 + \psi_n^2$. Furthermore, the expression of p_n can be determined by noting that $p_n(d) = \mathbb{P}(\mu_A - a_n^{-1}\sqrt{d} \leq \bar{X}_n \leq \mu_A + a_n^{-1}\sqrt{d})$.

Remarks

1. Noting that $b_n = O(1)$ and $a_n = o(n^{-1})$, it follows that $e_n = o(n^{-2})$ and that, as n diverges, D_n converges in probability to zero as fast as n^{-2}.

2. It can be checked (see Appendix A.1) that $D_n = a_n^2\psi_n^2 Y_n$, where Y_n is a non-central chi-square random variable with one degree of freedom and non-centrality parameter $\lambda_n = \delta^2/\psi_n^2$. Therefore, p_n, the predictive cdf of D_n, can also be expressed as follows: $p_n(d) = \mathbb{P}[D_n \leq d] = F_{\lambda_n}\left(d/a_n\psi_n^2\right)$, where F_{λ_n} is the cdf of Y_n.

3. Under the hypotheses of Result 1, but assuming $\pi_A(\theta) = \pi_D(\theta)$, we have that $e_n = \sigma^2 \; n_A/n(n+n_A)$ and $p_n(d) = \Phi\big(b_n^{-1/2}a_n^{-1}d^{1/2} \; \sigma^{-1}\big) - \Phi\big(- b_n^{-1/2}a_n^{-1}d^{1/2} \; \sigma^{-1}\big)$. These formulas are obtained by replacing n_D with n_A and by setting $\delta = 0$ in the corresponding expressions of Result 1.

4. Note that the expressions of both e_n and p_n depend on the prior means only through the absolute difference $|\delta| = |\mu_A - \mu_D|$.

Example 1. We illustrate the method described in this section using some numerical examples. For instance, let us assume a fixed known variance $\sigma^2 = 1$ and a design-prior with parameters $\mu_D = 2$ and $n_D = 20$. Just for illustration, we take $d = 0.2$ and, in order to assess the impact of different prior beliefs on e_n and p_n, we consider a set of values for the analysis-prior parameters.

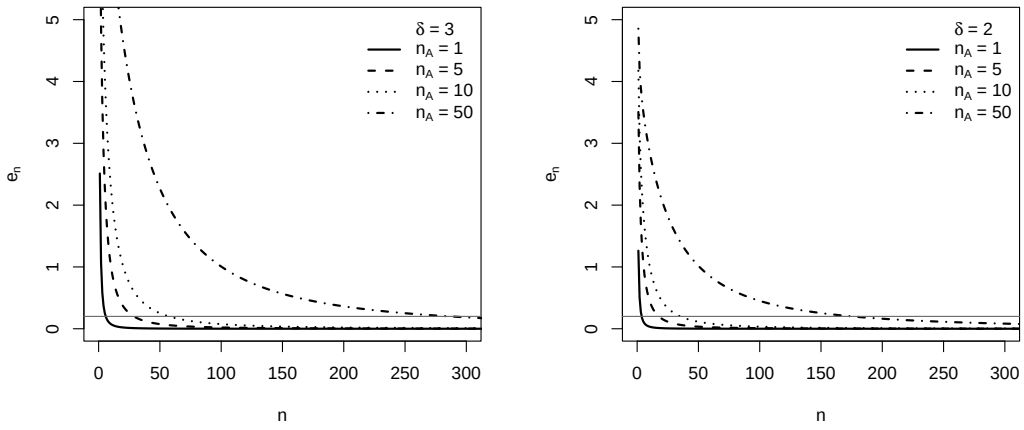

Figure 1: Plots of e_n with respect to n assuming $\mu_D = 2, n_D = 20$, $\sigma^2 = 1$ and for several choices of $|\delta|$ and n_A, $d = 0.2$.

In Figure 1, e_n is plotted with respect to n for $|\delta| = |\mu_A - \mu_D| = 2, 3$ and for several values of n_A. The two basic features to be noticed are the following:

- as $|\delta|$ becomes smaller – i.e. as the analysis-prior becomes more and more consistent with the design guess of the trial – the values of e_n become smaller and smaller for any given n;

- as n_A (the weight of the analysis-prior) increases, the values of the expected discrepancy become larger for any given n.

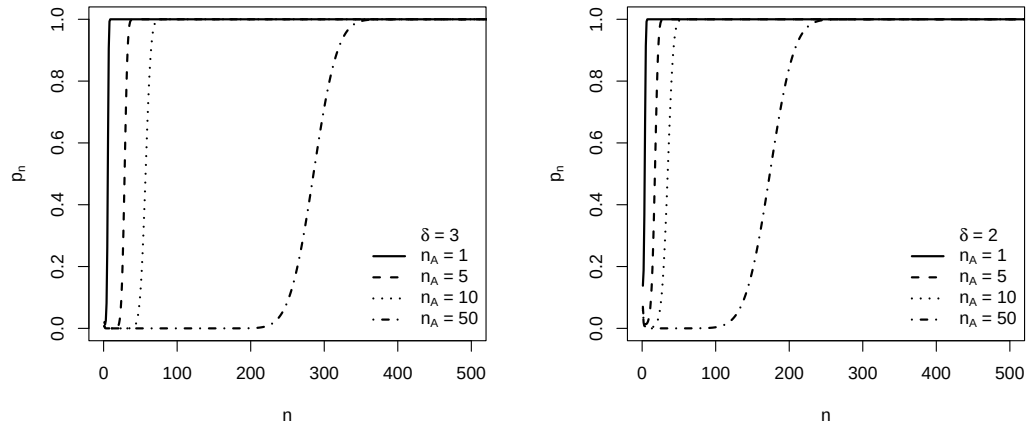

Figure 2: Plots of p_n with respect to n assuming $\mu_D = 2, n_D = 20$, $\sigma^2 = 1$ and for several choices of $|\delta|$ and n_A, $d = 0.2$, $\gamma = 0.9$.

Table 1-(a) reports the optimal sample sizes for different choices of $|\delta|$ and n_A. In general, for any fixed n_A, the smaller $|\delta|$ is, the smaller $n_e^*(d)$ and, for any fixed value of $|\delta|$, the larger n_A is, the larger the corresponding optimal sample size. Note also that the impact of n_A is stronger for larger values of δ. Similar considerations hold

	$n_e^*(d)$					$n_p^*(d,\gamma)$				
	$\|\boldsymbol{\delta}\|$					$\|\boldsymbol{\delta}\|$				
$\boldsymbol{n}_A$	4	3	2	1	0	4	3	2	1	0
1	8	6	4	2	1	9	7	5	3	2
5	40	29	18	7	3	44	33	22	12	5
10	80	57	35	14	3	87	64	43	21	7
50	398	286	175	65	5	430	318	207	97	14

Table 1: Sample sizes for several values of the analysis prior parameters, using the criterion based on e_n and p_n, assuming, $n_D = 20$, $\sigma^2 = 1$, several choices for $|\delta| = |\mu_D - \mu_A|$ and n_A, with $d = 0.2$, $\gamma = 0.9$.

for the plots of p_n in Figure 2 and the optimal sample sizes $n_p^*(d,\gamma)$ in Table 1-(b), obtained with the probability criterion with threshold $\gamma = 0.9$.

3.2 Unknown variance.

Let us now assume that the variance, σ^2, of the normal model is unknown. To extend the previous results, a natural choice is to consider a full conjugate analysis-prior for (θ, σ^2), i.e.: $\pi_A(\theta, \sigma^2) = \pi_A(\theta|\sigma^2)\pi_A(\sigma^2)$, where

$$\pi_A(\theta|\sigma^2) = N(\theta|\mu_A, \sigma^2/n_A) \qquad \text{and} \qquad \pi_A(\sigma^2) = IG(\sigma^2|\alpha_A, \beta_A), \tag{11}$$

and where, in general, $IG(\cdot|a, b)$ denotes the density function of an inverted-gamma random variable of parameter (a, b) (see Bernardo-Smith, 1994). From standard results on conjugate analysis (see, for instance, Bernardo and Smith (1994), p. 440) it follows that the posterior distribution of θ is a three parameters Student density whose expected value given is unchanged with respect to the known-variance case. The MLE of θ is as well unchanged with respect to the known variance case. Therefore, in this case, the expression of D_n is still given by (10).
For the design-prior, assume that $\pi_D(\theta, \sigma^2) = \pi_D(\theta|\sigma^2)\pi_D(\sigma^2)$, where

$$\pi_D(\theta|\sigma^2) = N(\theta|\mu_D, \sigma^2/n_D) \qquad \text{and} \qquad \pi_D(\sigma^2) = IG(\sigma^2|\alpha_D, \beta_D). \tag{12}$$

It follows that, $m_D(\bar{x}_n)$ is a three-parameters Student distribution with parameters $(\mu_D, \eta_n, 2\alpha_D)$, where $\eta_n = n_D n(n_D + n)^{-1}\alpha_D\beta_D^{-1}$ (Bernardo and Smith, 1994). The following result gives the explicit expressions of e_n and p_n for this unknown-variance case.

Result 2. Assume that $X_i|\theta$ has density $N(\cdot|\theta, \sigma^2/n)$, $i = 1, 2, \ldots, n$, and that $\pi_A(\theta, \sigma^2)$ and $\pi_D(\theta, \sigma^2)$ are respectively given by (11) and (12). It follows that the expression of $D_n =$ is still given by (10) and that

$$\begin{aligned} e_n &= a_n^2[b_n\widetilde{\sigma}^2 + \delta^2], \\ p_n(d) &= T_\nu\left[b_n^{-1/2}(\delta + a_n^{-1}d^{1/2})\,\dot{\sigma}^{-1}\right] - T_\nu\left[b_n^{-1/2}(\delta - a_n^{-1}d^{1/2})\,\dot{\sigma}^{-1}\right] \end{aligned}$$

where $\widetilde{\sigma}^2 = E[\sigma^2] = \beta_D(\alpha_D - 1)^{-1}$, $\dot{\sigma} = E[1/\sigma^2]^{-1/2} = \alpha_D^{-1}\beta_D$ and where $T_\nu(\cdot)$ is the cdf of a standard Student t distribution distribution with $\nu = 2\alpha_D$ degrees of freedom and where $\delta = \mu_D - \mu_A$.

Proof. The expression of e_n is determined noting that, under (12), $\mathbb{E}[\bar{X}_n] = \mu_D$ and $\mathbb{E}[\bar{X}_n^2] = \mu_D^2 + \eta_n^{-1}\alpha_D(\alpha_D - 1)^{-1}$ [see, for instance, Bernardo and Smith (1994), p. 122, for the moments of a three-parameters Student random variable]. The expression of p_n can be determined as in Result 1.

Note that the expression of e_n in Result 2 is obtained by the expression of e_n determined for the known variance case (Result 1) by simply replacing σ^2 with its expected value $\tilde{\sigma}^2$, determined with the design-prior (11). Similarly, the expression of p_n in Result 2 is obtained by the expression of p_n for the known variance case by replacing the functions $\Phi(\cdot)$ with $T_\nu(\cdot)$ and σ with $\dot{\sigma}$. For a general discussion on the relationships between optimal designs for the normal model with known and unknown variance see Verdinelli (2000).

Example 2. Let us consider again the setup of Example 1 in the previous section, but let us now assume (θ, σ^2) unknown with design- and analysis-prior as in Result 2. In Table 2 we consider three sets of values for the design-prior hyperparameters (α_D, β_D), yielding a prior mode equal to 1 (i.e. the value assigned to σ^2 in Example 1), with different values for the prior variance (0.33, 0.05, 0.02 respectively). When e_n is considered the resulting sample sizes are not affected by the different choices of the hyperparameters (see the comments after the proof of Result 2). Conversely, using p_n, smaller values of the prior variance results in slightly lowered value of the optimal sample sizes, uniformly with respect to $|\delta|$. Moreover, comparing Table 2 with the last row of Table 1 (corresponding to $n_A = 50$), we notice that the sample sizes are exactly the same for the criterion based on e_n and almost overlapping also for the one based on p_n.

		$n_e^*(d)$					$n_p^*(d, \gamma)$				
		$\|\boldsymbol{\delta}\|$					$\|\boldsymbol{\delta}\|$				
α_D	β_D	4	3	2	1	0	4	3	2	1	0
5	4	398	286	175	65	5	425	314	202	92	11
20	19	”	”	”	”	”	429	317	206	96	13
50	49	”	”	”	”	”	430	318	207	97	14

Table 2: Optimal sample sizes for several values of $|\delta|$, using the criterion based on e_n and p_n, assuming $n_A = 50$, $d = 0.2$, $\gamma = 0.9$ and hyperparameters α_D and β_D such that the prior mode for σ equals 1

Example 3. In the preceding examples we have seen that, in general, for any fixed sample size n, the conflict between $\hat{\theta}_F$ and $\hat{\theta}_B$ increases with the prior sample size, n_A. Hence, we can use the predictive analysis of the previous sections to establish, for given sample sizes, an upper bound for the analysis-prior variance in order to keep p_n (the probability of small values of the discrepancy between estimators)

sufficiently large (or, similarly, to have e_n sufficiently small). This kind of analysis allows one to identify values of n_A that correspond to priors that, for a given sample size, are relatively non-informative.
In Figure 3 e_n and p_n are plotted as functions of n_A for $n = 10$ and for several values of $|\delta|$. In the left panel we notice that, generally, e_n increases with the prior sample size n_A. This behavior is particularly evident in correspondence of larger values of $|\delta|$, while for $\delta = 0$, e_n increases very slowly with n_A. Table 3 reports the values of n_A needed reach $d = 0.2$. The right panel of Figure 3 shows that p_n decreases as n_A increases: the reduction is dramatic for $\delta = 4$, whereas it is smoother and smoother as μ_A approaches μ_D. In Table 3 we report the maximum values of n_A such that the probability exceeds the probability threshold $\gamma = 0.9$.

Finally, when we fix a larger value for the sample size, for instance $n = 100$, the affordable prior sample size n_A uniformly increases for any given value of δ, as shown in Table 3.

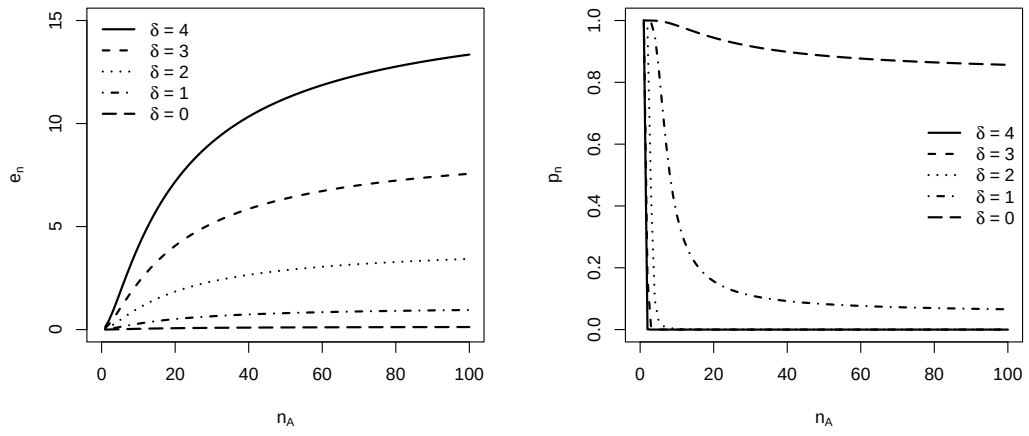

Figure 3: Plots of e_n and p_n with respect to n_A for several choices of $|\delta|$ ($\alpha_D = 5, \beta_D = 4$).

4 Illustrative example: a superiority trial

Results of the previous section are now employed in the context of a clinical trial designed to prove superiority of a new treatment towards a standard therapy for the same disease. The unknown parameter of interest is the log-odds ratio $\theta = \log \theta_1(1 - \theta_2)/\theta_2(1 - \theta_1)$, where θ_i is the probability of death under therapy

		n_A				
		$\lvert\boldsymbol{\delta}\rvert$				
	n	4	3	2	1	0
e_n	10	1	1	2	7	-
	100	12	17	28	76	-
p_n	10	1	1	2	4	39
	100	11	15	24	54	-

Table 3: Maximum values of n_A such that e_n reaches $d = 0.2$ and p_n reaches the probability threshold $\gamma = 0.9$, respectively, for a fixed sample size $n = 10, 100$ and several values of $|\delta|$.

i, $i = 1, 2$. Let $\bar{x}_{n_i}$ denote the sample proportion of events (death) under treatment i. Following, for instance, Spiegelhalter et al. (2004), the MLE of θ, $\hat{\theta}_F = \log \bar{x}_{n_1}(1-\bar{x}_{n_2})/\bar{x}_{n_2}(1-\bar{x}_{n_1})$, is asymptotically normal with parameters $(\theta, 4/n)$, where n (effective sample size) has now the interpretation of the overall number of deaths under the two treatments. Negative values of the estimates provide evidence if favor of treatment 1. Results of Section 3 can now be used for approximate inference on θ, by replacing $\bar{x}_n$ with $\hat{\theta}_F$ in all the formulas.

As a specific example, we consider the set-up of the GREAT trial, analyzed in detail in Spiegelhalter et al. (2004, pp.69-72) The goal of the experiment was to compare the effects of two alternative treatments for myocardial infarction. The evidence from the trial was in favor of the new theraphy (treatment 1=anistreplase) with respect to a standard therapy. In fact, $\hat{\theta}_F$ was equal to -0.74, corresponding to a 52% reduction in odds of death.

Assume now that a new experiment is planned to prove superiority of anistreplase and to estimate θ. We use the optimistic result of the GREAT trial to elicit the design prior π_D. Hence, we set $\mu_D = -0.74$ and $n_D = 30.5$. Under this optimistic scenario, we expect to observe values of $\hat{\theta}_F$ indicating superiority of anistreplase.

The optimistic outcomes of the GREAT trial (that we use here to set π_D) were in conflict with historical data available at the time of the experiment. These data have been used by Spiegelhalter et al. (2004) to elicit a normal analysis-prior density with parameters $\mu_A = -0.26$ and $n_A = 236.7$, which is quite more sceptical than π_D towards the new treatment. Using this prior, we expect to observe values of $\hat{\theta}_B$ that are, at least for moderate sample sizes, in conflict with those of $\hat{\theta}_F$. However, as n increases, the effect of the prior tends to disappear and the conflict between $\hat{\theta}_F$ and

n_A	236.7	236.7/2	236.7/5	236.7/10
n_e^*	99	56	28	18
n_p^*	130	73	33	20

Table 4: Sample sizes for the GREAT example.

$\hat{\theta}_B$ to be reduced. Figure 4 reports the plot of e_n as a function of the sample size

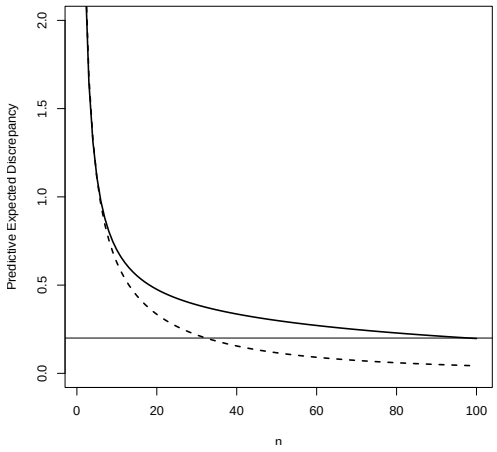

Figure 4: Plots of e_n for the GREAT example using the clinical prior [$\mu_A = -0.26$, $n_A = 236.7$ (solid line)] and the sceptical prior [$\mu_A = 0$, $n_A = 32.3$ (dashed line)].

(solid line). As expected, values of e_n decrease as n increases. Choosing, for instance, $d = 0.2$, yields $n_e^*(d) = 99$. This large value of the sample size highly depends on the analysis-prior sample size, n_A. In fact, if we reduce the analysis-prior precision (n_A), the corresponding optimal sample sizes decrease considerably, as shown in Table 4. For sensitivity analysis, Spiegelhater et al. (2004) consider an alternative analysis-prior, more sceptical about large treatment effects than the previous clinical prior. To formalize scepticism about the new treatment, the parameters are chosen so that the resulting prior is centered on zero and assigns 95% of its probability mass on an interval of the parameter space ranging from 50% of reduction in odds of death using anistreplase, to a 100% of increase. The resulting parameters are $\mu_A = 0$ and $n_A = 32.3$. The plot of e_n is reported in Figure 4 (dashed line). It is interesting to note that, despite the expected value of this prior ($\mu_A = 0$) is substantially more sceptical than the expected value of the clinical prior ($\mu_A = -.26$), the values of e_n are uniformly smaller than those of the first prior, due to the strong effect of the

large value of n_A in the clinical prior. The optimal sample size is now $n_e^*(d) = 33$. For completeness, we also report, in Figure 5, the plots of $p_n(d)$ for the two priors,

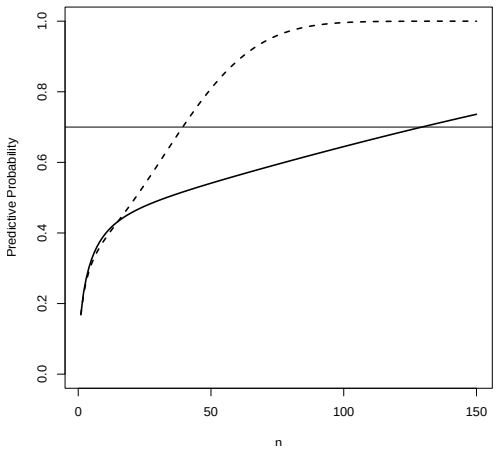

Figure 5: Plots of $p_n(d)$ for the GREAT example using the clinical prior [$\mu_A = -026$, $n_A = 236.7$ (solid line)] and the sceptical prior [$\mu_A = 0$, $n_A = 32.3$ (dashed line)], $d = 0.2$.

assuming $d = 0.2$ and $\gamma = 0.7$. The strong effect of prior precision is now even more self-evident: the optimal sample size is $n_p^*(d, \gamma) = 130$ for the clinical prior and $n_p^*(d, \gamma) = 33$ for the sceptical prior.

5 Conclusion

It is well known that, in many standard parametric problems, Bayesian and frequentist methods often provides approximately equal estimates as long as sampling information dominates the prior and that their difference tends to be reduced for increasing sample sizes. In this paper we have formalized the pre-experimental problem of selecting the minimal sample size sufficient to keep the conflict between Bayesian and frequentist point estimates sufficiently small. Of course, we have made several choices to deal with this problem that could be questioned. Here is a non-exhaustive list of critical points that may deserve further discussion and development.

i) *Alternative measures of conflict.* We have selected a specific measure of divergence between estimator, namely the quadratic difference (2). Of course, alternative choices can be considered. For instance, we can use the absolute

difference between estimators

$$D'_n(\mathbf{X}_n) = |\hat{\theta}_B(\mathbf{X}_n) - \hat{\theta}_F(\mathbf{X}_n)|.$$

It can be shown that, for the normal models of Section 3, closed-form expressions for D'_n, e_n and p_n can be determined.

ii) *Predictive approach.* For pre-experimental sample size computations we used the predictive distribution $m_D(\cdot)$ instead of the sampling distribution $f_n(\cdot|\theta)|_{\theta=\mu_D}$, in order to account for possible uncertainty on the design-value μ_D. In general, using a predictive distribution in the place of the sampling distribution implies larger sample sizes.

iii) *Two-priors approach.* Within the predictive approach to SSD, we prefer considering two priors, one for design and one for analysis, rather than using the same priors. In fact, we believe that accounting for pre-experimental information is different from modelling design uncertainty.

iv) *Extensions.* In this paper we have considered the basic normal model and an inferential problem that can be easily illustrated even in introductory courses on Bayesian statistics. We have focused on the normal model with conjugate priors. Extensions to more complex and useful models do not present difficulties in principle but may require analytical and computational efforts.

v) *Relationships with other SSD methods.* We believe that predictive control of the conflict between estimators is an important tool for evaluating the impact of prior assumptions on posterior analysis and to understand to what extent Bayesian procedure can be approximated by frequentist procedure. We also believe that the SSD criteria of Section 2.2 should be used in concert with other criteria specifically aimed at guaranteeing good performance of inferential procedures and based on, for instance, the variance of point estimators, length and location of interval estimators, measures of evidence in testing.

REFERENCES

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, II ed., Springer-Verlag, New York.

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

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

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

Gelman A, Carlin, J.B., Stern, H.S., Rubin, D.B. (2004). *Bayesian Data Analysis*, II ed., Chapman Hall/CRC.

Joseph, L., du Berger, R. and Belisle, P. (1997). Bayesian and mixed Bayesian/likelihood criteria for sample size determination. *Statistics in Medicine*, 16, 769-781.

Lee, P.M. (2004). *Bayesian Statistics: An Introduction*, III ed., Arnold, London.

O'Hagan, A. and Forster, J. J. (2004). *Bayesian Inference*, II ed., vol. 2B of "Kendall's Advanced Theory of Statistics". Arnold, London.

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

Robert, P.C., (2001). *from Decision-Theoretic Motivations to Computational Implementation, II ed. Springer-Verlag, New York.*

Sahu, S. K. and Smith, T. M. F. (2006). A Bayesian method of sample size determination with practical applications . *Journal of the Royal Statistical Society*, Ser. A, 169, no. 2, 235-253.

Sambucini, V. (2008). A Bayesian predictive two-stage design for phase II clinical trials. *Statistics in Medicine*, 27, no. 8, 1199-1224.

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

Tsutakawa, R.K. (1972). Design of experiment for bioassay. *Journal of the American Statistical Association*, 67, n. 339, 584-590.

Verdinelli, I. (2000). A note on Bayesian design for the normal linear model with unknown error variance. *Biometrika* (2000) 87,1, 222-227.

Wang, F., and Gelfand, A.E. (2002). A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science*, 17, n. 2, 193-208.

APPENDIX

A1.

We have seen in Section 2.1 that, from standard results on conjugate analysis for the normal model (see, for instance, Bernardo and Smith, 1994), the predictive distribution of the sample mean, $\bar{X}_n$ is $N(\mu_D, \psi_n^2)$ where $\psi_n^2 = \sigma^2(n+n_D)(nn_D)^{-1}$. Therefore, $\bar{X}_n - \mu_A \sim N(\mu_D - \mu_A, \psi_n^2)$ and $Y_n = (\bar{X}_n - \mu_A)^2\psi_n^{-2}$ is a a chi-square random variable with non-centrality parameter λ_n:

$$Y_n = \frac{(\bar{X}_n - \mu_A)^2}{\psi_n^2} \sim \chi_1^2(\lambda_n), \qquad \lambda_n = \frac{(\mu_D - \mu_A)^2}{\psi_n^2}.$$

Noting that $D_n = a_n\psi_n^2 Y_n$, it follows that the marginal cdf of D_n is

$$p_n(d) = \mathbb{P}[D_n < d] = \mathbb{P}[a_n\psi_n^2 Y_n < d] = F_{\lambda_n}\left(\frac{d}{a_n\psi_n^2}\right),$$

where F_{λ_n} is the cdf of a non-central chi-square with one degree of freedom and non-centrality parameter λ_n.