# Bayesian Average Error Based Approach to Sample Size Calculations for Hypothesis Testing

Eric M Reyes and Sujit K Ghosh

Department of Statistics, North Carolina State University, NC 27695, USA

Last Updated: September 23, 2010

**Abstract**

Under the classic statistical framework, sample size calculations for a hypothesis test of interest are made to adequately maintain pre-specified Type-I and Type-II error rates. These methods often suffer from several practical limitations, including the need to posit values for the parameters of interest without accounting for the uncertainty in these estimates. We propose a framework for hypothesis testing and sample size determination using Bayesian average errors by extending the classical framework. We consider the approach of rejecting the null hypothesis, in favor of the alternative, when a test statistic exceeds a cutoff. We choose the cutoff to minimize a weighted sum of Bayesian average errors and choose the sample size to bound the total error for the hypothesis test. We then apply this methodology to determine the sample size required for several designs common in medical studies.

Keywords: Bayes factor; Bayesian average errors; Hypothesis testing; Sample size determination.

# 1 Introduction

Sample size determination is critical in designing medical studies to test a hypothesis of interest. Failure to consider sample size calculations prior to a study has resulted in studies that lacked the power to detect clinically important effects (Friedman, Furberg, and DeMets 1998). Adcock (1997) presents a review of both classical and Bayesian approaches to sample size determination and Inoue, Berry, and Parmigiani (2005) establish a general framework that connects the classical and Bayesian perspectives.

Under the classical framework, sample size calculations are aimed at adequately maintaining pre-specified Type-I and Type-II error rates. There are some drawbacks to these classical methods. For a simple null and simple alternative hypothesis, Type-I and Type-II error rates are straightforward to obtain; however, for a hypothesis of the form

$$H_0 : \theta = 0 \quad \text{vs.} \quad H_1 : \theta \neq 0$$

where $\theta$ may denote the true average difference between treatment effects, the Type-II error depends on the unknown values of the parameter $\theta$ in the alternative hypothesis. Calculation of a Type-II error rate thus often requires the user to posit a value for the parameter under the alternative. Positing suitable values under a given hypothesis becomes more difficult when the null hypothesis is composite (e.g., hypotheses for non-inferiority tests). Sample size calculations under the classical framework are often based on a pivot quantity; however, Adcock (1997) points out that the existence of a pivot quantity is not guaranteed, even in common settings. The nuisance parameters in composite (null or alternative) hypotheses need to be estimated, eliminated by conditioning on a suitable statistic, or assumed known to compute the power (or Type-II error rate) at a desired value of the parameter of interest. Such a situation is very common in practice, especially when comparing two or more treatments by measuring their success (or failure) rates (for an example, see Section 3.4). These parameters, however, are rarely known in practice with a high degree of precision (M'Lan, Joseph, and Wolfson 2008), and suitable conditioning statistics may not be available in general (e.g., in non-inferiority tests for comparing two binomial proportions) (Röhmel and Mansmann 1999). Additionally, classical methods often rely heavily on asymptotic (normal) approximations when testing two composite hypotheses, which may be questionable in many common situations (M'Lan et al. 2008). While there have been some recent attempts to overcome these limitations (Fox et al. 2007), they remain an obstacle in common practice and often result in larger sample sizes than necessary. A general method for sample size calculations is not available that can be applied to a broad range of applications across a broad class of sampling distributions.

Such limitations have motivated several Bayesian solutions to sample size determination. The majority of Bayesian solutions are aimed at sample size calculations for interval estimation; M'Lan et al. (2008) give a thorough review and then extend such methods, with an emphasis on estimating binomial proportions. Some work on using decision theory to approach sample size determination has been done, as included in the review by Pezeshk (2003). Weiss (1997) introduced the use of Bayes factors in sample size determination with a focus on hypothesis testing as opposed to interval estimation. De Santis (2004) approached sample size determination using Bayes factors through an "evidential approach." This has since been extended to the use of alternative Bayes factors (De Santis 2007).

We propose a framework for hypothesis testing and sample size determination using Bayesian average errors that extends the classical framework from a Bayesian perspective by automatically determining the required cutoff value for a test statistic that maintains a target level of the total error rate. This framework does not suffer from the same limitations as those methods developed under the classical framework and provides a general approach to handling simple and complex hypotheses.

We develop our framework for hypothesis testing in Section 2, with a detailed discussion of this framework in the context of a one-sample normal density with known variance. Section 3 builds on the methodology to provide a rule for sample size determination; this rule is then applied to several examples common in clinical trials. We also compare our method with the classical approach. Section 4 presents an application of our framework to the design of a study evaluating the safety of a medication. We conclude our paper with a discussion of possible extensions to this framework.

# 2 Methodology

## 2.1 Notation

Suppose we observe data $X \sim f(x|\theta)$, where $f(x|\theta)$ denotes the conditional density of the data vector $X$ given the parameter $\theta \in \Theta$ and $\Theta$ denotes the parameter space. The data vector $X$ often consists of independent observations, but such an assumption is not needed to develop our framework. Define a prior density for $\theta$ on $\Theta$ by $\pi(\theta)$ with respect to some dominating measure. Consider the problem of comparing two competing hypotheses:

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta_1, \tag{2.1}$$

where $\Theta_0 \cap \Theta_1 = \emptyset$ and $\Theta_0 \cup \Theta_1 \subseteq \Theta$. Assume that $Pr(\theta \in \Theta_j) = \int_{\Theta_j} \pi(\theta) d\theta > 0$ for $j = 0, 1$.

Using the notation of Robert (2001), define $\pi_j(\theta)$ to be the prior with support $\Theta_j$ for $j = 0, 1$. That is,

$$\pi_j(\theta) = \frac{\mathbb{I}(\theta \in \Theta_j)\pi(\theta)}{\int_{\Theta_j} \pi(\theta) d\theta}$$

where $\mathbb{I}(u)$ takes the value 1 if the event $u$ occurs and 0 otherwise. Define the constrained marginal density of $X$ under $H_j$ for $j = 0, 1$ as

$$m_j(x) = \int_{\Theta_j} f(x|\theta)\pi_j(\theta) d\theta.$$

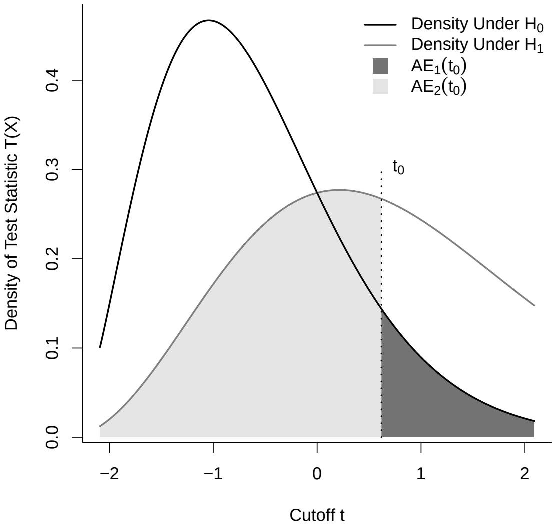


Figure 1: Definition of Average Bayes Type-I and Type-II Error for a given decision rule "reject $H_0$ if $T(X) > t_0$."

## 2.2 Bayesian Average Errors

Let $T(X)$ denote a "test statistic" measuring the evidence favoring the alternative hypothesis; that is, we reject the null hypothesis (in favor of the alternative) if $T(X) > t$ for some value $t$. Define the *Average Bayes Type-I Error* ($\mathrm{AE}_1$) for this decision rule as

$$\mathrm{AE}_1(t) = Pr(T(X) > t|\theta \in \Theta_0).$$

Similarly, define the *Average Bayes Type-II Error* ($\mathrm{AE}_2$) for this test as

$$\mathrm{AE}_2(t) = Pr(T(X) \leq t|\theta \in \Theta_1).$$

Figure 1 gives a graphical representation of $\mathrm{AE}_1$ and $\mathrm{AE}_2$ for a specified cutoff $t = t_0$. These errors are distinguished from the corresponding classical errors because under a Bayesian framework, $Pr(\theta \in \Theta_j)$, as well as the conditional probability $Pr(T(X) > t|\theta \in \Theta_j)$, is well-defined for $j = 0, 1$. More over, calculation of $\mathrm{AE}_2$ does not require selection of a particular value for the parameter under the alternative, as calculation of the classical Type-II error rate often does. Notice that these definitions avoid the need to construct a conditioning statistic to eliminate nuisance parameters or to estimate or specify a particular value of the nuisance parameter when the hypotheses are composite.

These Bayesian average errors can conveniently be expressed in terms of the marginal distributions of Section 2.1. That is,

$$\begin{aligned}\mathrm{AE}_1(t) &= \int \mathbb{I}(T(x) > t) m_0(x) dx \quad \text{and} \\ \mathrm{AE}_2(t) &= \int \mathbb{I}(T(x) \leq t) m_1(x) dx.\end{aligned}$$

We note that $\mathrm{AE}_1$ is a non-increasing function in $t$ while $\mathrm{AE}_2$ is a non-decreasing function. Thus, as the cutoff $t$ is altered, there is a tradeoff between these error rates. While the cutoff value $t$ can be chosen such that $\mathrm{AE}_1$ and/or $\mathrm{AE}_2$ is bounded (Weiss 1997), we propose choosing a cutoff value $t$ such that a weighted combination of these errors is minimized, allowing both error rates to be controlled simultaneously. Define the *Total Weighted Error* (TWE) as

$$\mathrm{TWE}(t, w) = w\mathrm{AE}_1(t) + (1 - w)\mathrm{AE}_2(t),$$

for some $w \in [0, 1]$. The value of $w$ is specified *a priori* and is used to place more emphasis on controlling one type of error over the other; values of $w$ near 1, for example, place more emphasis on controlling $\mathrm{AE}_1$. For a given value of $w \in [0, 1]$, we obtain the optimal cutoff $t_0(w)$ as

$$t_0(w) = \arg\min_t \{\mathrm{TWE}(t, w)\}.$$

That is, we reject $H_0$ if $T(X) > t_0(w)$. We now refine this general framework by considering a specific choice of $T(X)$ for which the existence of the corresponding $t_0(w)$ is guaranteed and can be derived explicitly.

### 2.3 Bayes Factor as Test Statistic

A natural choice for $T(X)$ is the logarithm of the Bayes factor (Weiss 1997). Define the Bayes factor (BF) as

$$\text{BF}(X) = \left(\frac{Pr(\theta \in \Theta_1|X)}{Pr(\theta \in \Theta_0|X)}\right)\left(\frac{Pr(\theta \in \Theta_0)}{Pr(\theta \in \Theta_1)}\right),$$

and consider $T(X) = \log(\text{BF}(X))$. In this section, we provide further justification for the BF to be the "optimal" test statistic. Observe that under this definition, positive values of $T(X)$ favor the alternative hypothesis. Note that $T(X)$ can be expressed in terms of the marginal distributions of Section 2.1. That is,

$$T(X) = \log(\text{BF}(X)) = \log(m_1(X)) - \log(m_0(X)).$$

**Corollary 1.** *Let $T(X) = \log(BF(X))$ for observed data $X \sim f(X|\theta)$, with prior $\pi(\theta)$ such that for $j = 0, 1$, $Pr(\theta \in \Theta_j) > 0$ for the hypothesis test*

$$H_0 : \theta \in \Theta_0 \quad vs. \quad H_1 : \theta \in \Theta_1.$$

*Then, for a given value of $w \in (0, 1)$, $t_0(w) = \log\left(\frac{w}{1-w}\right)$ minimizes TWE.*

This is a consequence of a more general result shown in Appendix A, which in turn establishes the optimality of BF as the best test statistic in minimizing TWE. Corollary 1 provides a one-to-one correspondence between a cutoff value and the emphasis placed on controlling $\text{AE}_1$. For example, Kass and Raftery (1995) suggested that a cutoff of 5 for the log BF indicates decisive evidence in favor of the alternative; this cutoff corresponds to $w = 0.99$, placing extreme emphasis on controlling $\text{AE}_1$.

We illustrate this framework on a common example.

## 2.4 Example 1: One Sample Normal with Known Variance, Two-Sided Test

Following Weiss (1997) and Adcock (1997), we consider, in detail, the testing of the mean of a normal distribution with known variance. Consider $X = (x_1, x_2, \ldots, x_n)$ such that $x_1, x_2, \ldots, x_n|\theta \overset{\text{iid.}}{\sim} N(\theta, \sigma_0^2)$, where $\sigma_0^2$ is known and $\theta \in \Theta = \mathbb{R}$. The hypothesis test of interest, as defined in (2.1), is completely specified through $\Theta_0$ and $\Theta_1$. Let $\Theta_0 = \{\theta_0\}$ for some $\theta_0 \in \mathbb{R}$, and let $\Theta_1 = \{\theta \in \mathbb{R} : \theta \neq \theta_0\}$. Consider the prior

$$\pi(\theta) = u\mathbb{I}(\theta = \theta_0) + (1-u)\mathbb{I}(\theta \neq \theta_0)\tau^{-1}\phi\left(\frac{\theta - \mu}{\tau}\right),$$

where $u = Pr(\theta = \theta_0) \in (0,1)$ and $\phi(x)$ represents the standard Normal density. Note that this formulation of the prior satisfies the assumption that the prior probability of each hypothesis be nonzero. Taking $u = 0.5$ reflects the belief that the two hypotheses are *a priori*, equally likely.

We now calculate $m_0(X)$ and $m_1(X)$, which will allow computation of $T(X)$ and the average errors. The sample mean $\bar{X} = n^{-1}\sum_{i=1}^{n} x_i$ is sufficient; thus, it is enough to work with its distribution: $\bar{X}|\theta \sim N(\theta, \sigma_0^2/n)$. The densities $\pi_0(\theta)$ and $\pi_1(\theta)$ are readily available from the prior density as

$$\begin{aligned} \pi_0(\theta) &= \mathbb{I}(\theta = \theta_0) \quad \text{and} \\ \pi_1(\theta) &= \mathbb{I}(\theta \neq \theta_0)\tau^{-1}\phi\left(\frac{\theta - \mu}{\tau}\right), \end{aligned}$$

which correspond to the priors proposed by Weiss (1997). It is straightforward to show that

$$\begin{aligned} m_0(X) &= \frac{\sqrt{n}}{\sigma_0}\phi\left(\frac{\sqrt{n}(\bar{X} - \theta_0)}{\sigma_0}\right) \quad \text{and} \\ m_1(X) &= \frac{1}{\sqrt{n^{-1}\sigma_0^2 + \tau^2}}\phi\left(\frac{\bar{X} - \mu}{\sqrt{n^{-1}\sigma_0^2 + \tau^2}}\right). \end{aligned}$$

We note that as these marginal densities do not depend on the prior parameter $u$, our computations of $\mathrm{AE}_1$ and $\mathrm{AE}_2$ will also be free of the choice of $u$.

$T(X) = \log(BF(X))$ is then given as

$$\begin{aligned} T(X) &= \frac{1}{2}\log\left(\frac{n^{-1}\sigma_0^2}{n^{-1}\sigma_0^2+\tau^2}\right) + \frac{(\bar{X}-\theta_0)^2}{2n^{-1}\sigma_0^2} - \frac{(\bar{X}-\mu)^2}{2(n^{-1}\sigma_0^2+\tau^2)} \\ &= B(\bar{X}+C)^2 - A \end{aligned}$$

where

$$\begin{aligned} A &= \left(\frac{1}{2}\right)\left[\frac{(\mu-\theta_0)^2}{\tau^2} + \log\left(1+\frac{n\tau^2}{\sigma_0^2}\right)\right], \\ B &= \frac{\tau^2}{2n^{-1}\sigma_0^2\left(n^{-1}\sigma_0^2+\tau^2\right)}, \quad \text{and} \\ C &= \left(\frac{\sigma_0^2}{n\tau^2}\right)(\mu-\theta_0) - \theta_0. \end{aligned}$$

Note that under either hypothesis, $T(X) > -A$ with probability 1. Thus, for all $t \leq -A$, $Pr(T(X) \leq t|\theta \in \Theta_j) = 0$ for $j = 0, 1$. We then have that for all $t > -A$,

$$\begin{aligned} \mathrm{AE}_1(t) =& 1 - \Phi\left(\sqrt{\frac{t+A}{Bs_0^2}} - \frac{C+\theta_0}{s_0}\right) + \Phi\left(-\sqrt{\frac{t+A}{Bs_0^2}} - \frac{C+\theta_0}{s_0}\right) \quad \text{and} \\ \mathrm{AE}_2(t) =& \Phi\left(\sqrt{\frac{t+A}{Bs_1^2}} - \frac{C+\mu}{s_1}\right) - \Phi\left(-\sqrt{\frac{t+A}{Bs_1^2}} - \frac{C+\mu}{s_1}\right), \end{aligned}$$

where $\Phi(x)$ represents the distribution function of a standard Normal, $s_0^2 = \sigma_0^2/n$, and $s_1^2 = \sigma_0^2/n + \tau^2$.

We observe that as $\tau^2 \to \infty$, both $\mathrm{AE}_1(t)$ and $\mathrm{AE}_2(t)$ approach 0. That is, extremely vague priors for the alternative hypothesis give small errors at any sample size. This was also noted in the approach of Weiss (1997); as he suggested, careful consideration must be given to the choice of $\tau^2$. For instance, we may use $\tau = 2\sigma_0$.

Finally, we observe that as $n \to \infty$, both $\mathrm{AE}_1(t)$ and $\mathrm{AE}_2(t)$ approach 0. Thus, we can make the sum of these two errors arbitrarily small by choosing a large enough sample size.

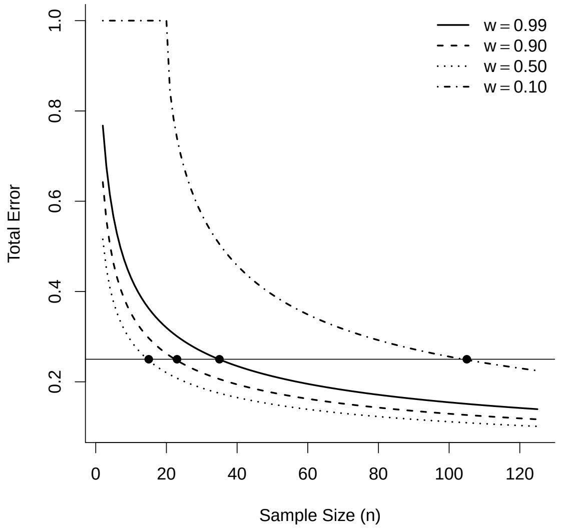


Figure 2: Total Error as the sample size increases for a two-sided test of the mean for a single normal distribution with known variance (Example 1). The selected sample size is the smallest $n$ such that, for the given weight $w$, the Total Error falls below the bound $\alpha = 0.25$.

This motivates our rule for sample size determination.

## 3 Sample Size Determination

Using the methodology presented in Section 2, we define the following criteria for calculating the required sample size for a hypothesis test of interest.

> Given the weight $w \in (0, 1)$ and $\alpha \in (0, 1)$, find the minimum $n > 1$ such that $\text{TE}(t_0(w)) \leq \alpha$, where $\text{TE}(t) = \text{AE}_1(t) + \text{AE}_2(t)$.

This sample size rule for various weights is presented graphically in Figure 2. It is to be noted that $\alpha$ does not represent the Type-I error rate, as is tradition in the classical framework. Rather, $\alpha$ denotes the bound on the sum of the Type-I and Type-II error rates. We now apply this criteria to several examples.

Table 1: Sample size requirements and resulting average errors for several weights for a two-sided test of the mean for a one-sample normal distribution with known variance (Example 1). The two scenarios differ in the choice of the prior mean under the alternative $\mu$. The remaining parameters were the same between the two scenarios: $u = 0.5$, $\theta_0 = 0$, $\sigma_0 = 2$, and $\tau = 2\sigma_0 = 4$. The total error bound $\alpha$ was set at 0.25 for both scenarios.

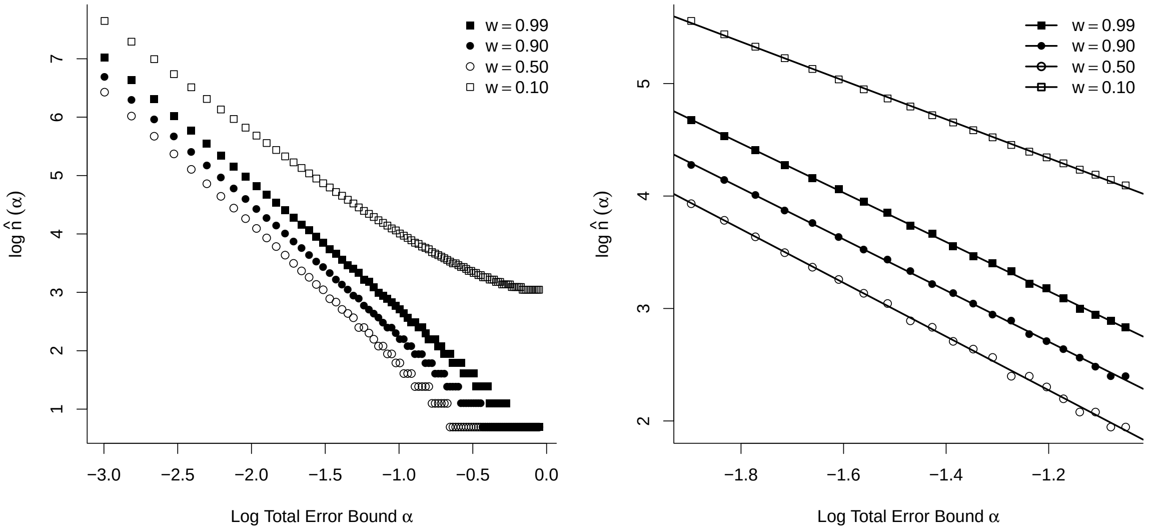
|  | $w$ | $n$ | $\mathrm{AE}_1$ | $\mathrm{AE}_2$ |
|---|---|---|---|---|
| $\mu = 0$ | 0.99 | 147 | 0.0001 | 0.2499 |
|  | 0.95 | 105 | 0.0011 | 0.2488 |
|  | 0.90 | 90 | 0.0027 | 0.2469 |
|  | 0.50 | 59 | 0.0413 | 0.2078 |
|  | 0.10 | 417 | 0.1996 | 0.0500 |
|  | 0.05 | 1566 | 0.2255 | 0.0244 |
| $\mu = 2$ | 0.99 | 55 | 0.0001 | 0.2498 |
|  | 0.95 | 39 | 0.0011 | 0.2488 |
|  | 0.90 | 34 | 0.0028 | 0.2447 |
|  | 0.50 | 22 | 0.0420 | 0.2053 |
|  | 0.10 | 153 | 0.2000 | 0.0500 |
|  | 0.05 | 576 | 0.2255 | 0.0244 |

## 3.1 Example 1 Continued

We return to Example 1, described in Section 2.4. Consider the test defined by $\theta_0 = 0$ with known variance $\sigma_0^2 = 4$. We consider two specifications of the prior parameters. For the first specification, we set $\mu = \theta_0 = 0$; for the second, we set $\mu = \theta_0 + \sigma_0 = 2$. Under both scenarios we take $\tau = 2\sigma_0 = 4$ and $u = 0.5$, though recall that the value of BF does not depend on $u$.

Table 1 reports the required sample size and Bayesian average errors for several weights under both specifications. We note that $w = 0.5$ gives the lowest sample sizes, which turns out to be a general result (see Section 3.6). We also note that weights near 0 tend to require larger samples than weights near 1. As expected, larger sample sizes are also required under the first specification, when the prior mean takes the value of the parameter $\theta$ under $H_0$.

We also investigated the relationship between the total error bound $\alpha$ and the required sample size for a fixed weight. For each of four weights $w$, we calculated the sample size for a



(a) $\alpha \in (0, 1)$

(b) $\alpha \in (0.15, 0.35)$

Figure 3: Log sample size required as a function of the total error bound $\alpha$ for a two-sided test of the mean for a one-sample normal distribution with known variance (Example 1). Panel (b) shows the observed relationship with the least squares fit overlayed. The following parameters were used: $u = 0.5$, $\theta_0 = 0$, $\sigma_0 = 2$, $\mu = \theta_0 = 0$, and $\tau = 2\sigma_0 = 4$.

fine grid of $\alpha \in (0, 1)$. Figure 3 shows the observed relationship under the first specification of prior parameters described above. While it is unclear what the underlying relationship is, we observe that for $\alpha \in [0.15, 0.35]$, which encompasses a reasonable set of choices most likely to be used in practice, there is a strong linear relationship between $\log n$ and $\log \alpha$ (Figure 3). This linear relationship for $\alpha \in [0.15, 0.35]$ was observed in all examples.

## 3.2 Example 2: One Sample Binomial, One-Sided Test

Consider collecting a single sample for a binary response. The data can be summarized as $X|\theta \sim \text{Bin}(n, \theta)$, where $\theta \in \Theta = [0, 1]$. Consider the one-sided hypothesis where $\Theta_0 = \{\theta \in [0, 1] : \theta \leq \theta_0\}$ for some $\theta_0 \in \Theta$, and $\Theta_1 = \{\theta \in [0, 1] : \theta > \theta_0\}$. We consider the prior

$$\pi(\theta) = p_{(a,b)}(\theta),$$

where $p_{(a,b)}(\theta)$ represents a Beta$(a, b)$ density. Observe that the prior probability of $H_0$ is determined by the choice of parameters in the prior distribution $a$ and $b$. That is,

$$Pr(\theta \in \Theta_0) = P_{(a,b)}(\theta_0),$$

where $P_{(a,b)}$ represents the Beta$(a, b)$ distribution function.

The densities $\pi_0(\theta)$ and $\pi_1(\theta)$ are then given as

$$\pi_0(\theta) = \mathbb{I}(\theta \leq \theta_0)\frac{p_{(a,b)}(\theta)}{P_{(a,b)}(\theta_0)} \quad \text{and}$$
$$\pi_1(\theta) = \mathbb{I}(\theta > \theta_0)\frac{p_{(a,b)}(\theta)}{1 - P_{(a,b)}(\theta_0)}.$$

It can then be shown that

$$m_0(X) = \binom{n}{x}\left(\frac{\text{be}(a + x, b + n - x)P_{(a+x,b+n-x)}(\theta_0)}{\text{be}(a, b)P_{(a,b)}(\theta_0)}\right)$$

and

$$m_1(X) = \binom{n}{x}\left(\frac{\text{be}(a + x, b + n - x)(1 - P_{(a+x,b+n-x)}(\theta_0))}{\text{be}(a, b)(1 - P_{(a,b)}(\theta_0))}\right),$$

where $\text{be}(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$ denotes the beta function and $\Gamma(x)$ denotes the gamma function. While an expression for $\text{AE}_1$ is not available as explicitly as it was in Example 1, it is straightforward to calculate as

$$\text{AE}_1(t) = \sum_{k=0}^{n} \mathbb{I}(T(k) > t)m_0(k).$$

$\text{AE}_2$ can be similarly computed.

Table 2 reports the required sample size and Bayesian average errors for several weights under three different hypotheses: $\theta_0 = 0.25$, $\theta_0 = 0.50$, $\theta_0 = 0.75$. In all three scenarios, we considered a uniform prior distribution ($a = b = 1$). The results are similar to that of Example 1. We note that $\text{AE}_1$ is a non-increasing function in $w$, while $\text{AE}_2$ is a non-

Table 2: Sample size requirements and resulting average errors for several weights for a one-sided test of the event rate for a one-sample binomial distribution (Example 2). The three scenarios differ in the choice of the critical value in the null hypothesis $\theta_0$. The prior parameters were the same between the three scenarios: $a = b = 1$. The total error bound $\alpha$ was set at 0.25 for all scenarios.

| | $w$ | $n$ | $\text{AE}_1$ | $\text{AE}_2$ |
|---|---|---|---|---|
| $\theta_0 = 0.25$ | 0.99 | 45 | 0.0002 | 0.2465 |
| | 0.95 | 25 | 0.0035 | 0.2319 |
| | 0.90 | 18 | 0.0085 | 0.2309 |
| | 0.50 | 8 | 0.1005 | 0.1446 |
| | 0.10 | 28 | 0.2288 | 0.0188 |
| | 0.05 | 60 | 0.2312 | 0.0060 |
| $\theta_0 = 0.50$ | 0.99 | 92 | 0.0003 | 0.2476 |
| | 0.95 | 41 | 0.0039 | 0.2420 |
| | 0.90 | 27 | 0.0116 | 0.2259 |
| | 0.50 | 9 | 0.1230 | 0.1230 |
| | 0.10 | 27 | 0.2259 | 0.0116 |
| | 0.05 | 41 | 0.2420 | 0.0039 |
| $\theta_0 = 0.75$ | 0.99 | 184 | 0.0004 | 0.2444 |
| | 0.95 | 60 | 0.0060 | 0.2312 |
| | 0.90 | 28 | 0.0188 | 0.2288 |
| | 0.50 | 8 | 0.1446 | 0.1005 |
| | 0.10 | 18 | 0.2309 | 0.0085 |
| | 0.05 | 25 | 0.2319 | 0.0035 |

decreasing function, and $w = 0.5$ results in the smallest sample size.

### 3.3 Example 3: Two Sample Normal with Known Variance, Two-Sided Test

Clinical trials often involve comparing two independent treatment groups with a continuous response. Consider $X = (x_{1,1}, x_{1,2}, \ldots, x_{1,n}, x_{2,1}, x_{2,2}, \ldots, x_{2,n})$ such that $x_{k,1}, x_{k,2}, \ldots, x_{k,n} | \theta \overset{\text{ind.}}{\sim} N(\theta_k, \sigma_0^2)$, for $k = 1, 2$ where $\sigma_0^2$ is known and $\theta = (\theta_1, \theta_2) \in \Theta = \mathbb{R}^2$. Let $\Theta_0 = \{\theta \in \Theta :$

$\theta_1 = \theta_2\}$, and let $\Theta_1 = \{\theta \in \Theta : \theta_1 \neq \theta_2\}$. Consider the prior

$$\begin{aligned}\pi(\theta) =& u\mathbb{I}(\theta_1 = \theta_2 = \eta)\tau_0^{-1}\phi\left(\frac{\eta - \mu_0}{\tau_0}\right) + \\ & (1-u)\mathbb{I}(\theta_1 \neq \theta_2)\tau_1^{-1}\phi\left(\frac{\theta_1 - \mu_1}{\tau_1}\right)\tau_2^{-1}\phi\left(\frac{\theta_2 - \mu_2}{\tau_2}\right).\end{aligned}$$

The statistic $(\bar{X}_1, \bar{X}_2)$ is sufficient. Using the joint likelihood of the sample means, it can be shown that the marginal density of $(\bar{X}_1, \bar{X}_2)$ under $H_0$ is given by

$$(\bar{X}_1, \bar{X}_2)|\theta \in \Theta_0 \sim N\left(\begin{pmatrix}\mu_0 \\ \mu_0\end{pmatrix}, \begin{pmatrix}\sigma_0^2 n^{-1} + \tau_0^2 & \tau_0^2 \\ \tau_0^2 & \sigma_0^2 n^{-1} + \tau_0^2\end{pmatrix}\right),$$

and the marginal density under $H_1$ is given by

$$m_1(X) = (\sigma_0^2 n^{-1} + \tau_1^2)^{-1/2}\phi\left(\frac{\bar{X}_1 - \mu_1}{(\sigma_0^2 n^{-1} + \tau_1^2)^{1/2}}\right)(\sigma_0^2 n^{-1} + \tau_2^2)^{-1/2}\phi\left(\frac{\bar{X}_2 - \mu_2}{(\sigma_0^2 n^{-1} + \tau_2^2)^{1/2}}\right).$$

That is, under $H_0$, $(\bar{X}_1, \bar{X}_2)$ follows a bivariate normal density and under $H_1$ the sample means are independently normal.

Table 3 reports the required sample size and Bayesian average errors for several weights under two specifications of the prior parameters. Under the first specification, we set all prior parameters for the means equal ($\mu_0 = \mu_1 = \mu_2 = 0$); for the second, we set $\mu_0 = 0$, $\mu_1 = -2$, and $\mu_2 = 2$. For both scenarios, we set $u = 0.5$, $\sigma_0 = 2$ and consider $\tau_0 = \tau_1 = \tau_2 = 2\sigma_0 = 4$.

## 3.4 Example 4: Two Sample Binomial, Two-Sided Test

Another common design in many clinical studies is a two arm trial to compare the event rate between two independent treatment groups. Let $X = (x_1, x_2)$ where $x_k|\theta_k \overset{\text{ind.}}{\sim} \text{Bin}(n, \theta_k)$ for $k = 1, 2$ where $\theta = (\theta_1, \theta_2)$ and $\Theta = [0, 1]^2$. Let $\Theta_0 = \{\theta \in \Theta : \theta_1 = \theta_2\}$ and $\Theta_1 = \{\theta \in \Theta :$

Table 3: Sample size requirements and resulting average errors for several weights for a two-sided hypothesis for two independent means when the variance is known (Example 3). In both scenarios, $u = 0.5$ and $\sigma_0 = 2$. The total error bound $\alpha$ was set at 0.25 for both scenarios.

| Prior Parameters | | | | | | Results | | | |
|---|---|---|---|---|---|---|---|---|---|
| $\mu_0$ | $\tau_0$ | $\mu_1$ | $\tau_1$ | $\mu_2$ | $\tau_2$ | $w$ | $n$ | AE$_1$ | AE$_2$ |
| 0 | 4 | 0 | 4 | 0 | 4 | 0.99 | 74 | 0.0000 | 0.2464 |
| 0 | 4 | 0 | 4 | 0 | 4 | 0.95 | 55 | 0.0004 | 0.2352 |
| 0 | 4 | 0 | 4 | 0 | 4 | 0.90 | 50 | 0.0040 | 0.2448 |
| 0 | 4 | 0 | 4 | 0 | 4 | 0.50 | 34 | 0.0576 | 0.1920 |
| 0 | 4 | 0 | 4 | 0 | 4 | 0.10 | 349 | 0.2240 | 0.0240 |
| | | | | | | | | | |
| 0 | 4 | -2 | 4 | 2 | 4 | 0.99 | 26 | 0.0000 | 0.2328 |
| 0 | 4 | -2 | 4 | 2 | 4 | 0.95 | 18 | 0.0004 | 0.2392 |
| 0 | 4 | -2 | 4 | 2 | 4 | 0.90 | 16 | 0.0016 | 0.2312 |
| 0 | 4 | -2 | 4 | 2 | 4 | 0.50 | 15 | 0.0380 | 0.1952 |
| 0 | 4 | -2 | 4 | 2 | 4 | 0.10 | 70 | 0.1744 | 0.0752 |

$\theta_1 \neq \theta_2\}$. Consider the prior

$$\begin{aligned}\pi(\theta) =& u\mathbb{I}(\theta_1 = \theta_2 = \eta)p_{(a_0,b_0)}(\eta)+\\ &(1-u)\mathbb{I}(\theta_1 \neq \theta_2)p_{(a_1,b_1)}(\theta_1)p_{(a_2,b_2)}(\theta_2).\end{aligned}$$

The marginal densities are given by

$$m_0(X) = \binom{n}{x_1}\binom{n}{x_2}\left(\frac{\text{be}(a_0 + x_1 + x_2, b_0 + 2n - x_1 - x_2)}{\text{be}(a_0, b_0)}\right)$$

and

$$m_1(X) = \binom{n}{x_1}\binom{n}{x_2}\left(\frac{\text{be}(a_1 + x_1, b_1 + n - x_1)}{\text{be}(a_1, b_1)}\right)\left(\frac{\text{be}(a_2 + x_2, b_2 + n - x_2)}{\text{be}(a_2, b_2)}\right).$$

Table 4 reports the required sample size and Bayesian average errors for several weights under two specifications of the prior parameters. Under the first specification, $a_0 = b_0 = a_1 = b_1 = a_2 = b_2 = 1$; under the second, $a_0 = b_0 = 1$, $a_1 = 5/16$, $b_1 = 15/16$, $a_2 = 15/16$,

Table 4: Sample size requirements and resulting average errors for several weights for a two-sided hypothesis for two independent proportions (Example 4). In both scenarios, $u = 0.5$. The total error bound $\alpha$ was set at 0.25 for both scenarios.

| Prior Parameters | | | | | | Results | | | |
|---|---|---|---|---|---|---|---|---|---|
| $a_0$ | $b_0$ | $a_1$ | $b_1$ | $a_2$ | $b_2$ | $w$ | $n$ | AE$_1$ | AE$_2$ |
| 1 | 1 | 1 | 1 | 1 | 1 | 0.99 | 285 | 0.0001 | 0.2498 |
| 1 | 1 | 1 | 1 | 1 | 1 | 0.95 | 202 | 0.0011 | 0.2482 |
| 1 | 1 | 1 | 1 | 1 | 1 | 0.90 | 172 | 0.0028 | 0.2467 |
| 1 | 1 | 1 | 1 | 1 | 1 | 0.50 | 111 | 0.0429 | 0.2065 |
| 1 | 1 | 1 | 1 | 1 | 1 | 0.10 | 827 | 0.2018 | 0.0479 |
| 1 | 1 | 15/16 | 5/16 | 5/16 | 15/16 | 0.99 | 52 | 0.0001 | 0.2485 |
| 1 | 1 | 15/16 | 5/16 | 5/16 | 15/16 | 0.95 | 37 | 0.0012 | 0.2487 |
| 1 | 1 | 15/16 | 5/16 | 5/16 | 15/16 | 0.90 | 32 | 0.0028 | 0.2452 |
| 1 | 1 | 15/16 | 5/16 | 5/16 | 15/16 | 0.50 | 20 | 0.0554 | 0.1916 |
| 1 | 1 | 15/16 | 5/16 | 5/16 | 15/16 | 0.10 | 136 | 0.2019 | 0.0472 |

and $b_2 = 5/16$. For both specifications $u = 0.5$. The prior parameters under the latter specification were chosen such that under the null hypothesis, $E[\eta] = 0.5$, and under the alternative hypothesis $E[\theta_1] = 0.25$ and $E[\theta_2] = 0.75$. Note that for both specifications $Var[\eta] = Var[\theta_1] = Var[\theta_2] = 1/12$.

### 3.5 Comparison with Classical Approach

Consider the first specification from Example 1, in which the prior mean is taken to be the value of the parameter under the null hypothesis. We compare the resulting sample sizes from applying our method with those obtained from the classical approach described in Chow, Shao, and Wang (2003):

$$n_d = \frac{\sigma_0^2 \left(Z_{1-\alpha^*/2} + Z_{1-\beta^*}\right)^2}{d^2},$$

where $\alpha^*$ and $\beta^*$ represent the Type-I and Type-II errors to maintain and $d = \theta_1 - \theta_0$ is the minimum difference to detect with power $1 - \beta^*$. To compare our method with that

Table 5: Comparison with classical method. Sample sizes corresponding to Example 1 were obtained using the classical frequentist method for various "minimum differences" to detect. The differences are given in terms of the known standard deviation; the desired Type-I and Type-II error for the frequentist method were set at 0.05 and 0.2, respectively. $n_{\text{BAE}}$ corresponds to the sample size required using the method we propose based on Bayesian average errors, with a weight of $w = 0.5$, a total error bound $\alpha = 0.25$ and the following prior parameters: $u = 0.5$, $\theta_0 = 0$, $\sigma_0 = 2$, $\mu = \theta_0 = 0$, and $\tau = 2\sigma_0 = 4$. The resulting sample size attained an $\text{AE}_1 = 0.041$ and $\text{AE}_2 = 0.206$.

| $n_{\text{BAE}}$ | $n_{0.1\sigma_0}$ | $n_{0.2\sigma_0}$ | $n_{0.3\sigma_0}$ | $n_{0.4\sigma_0}$ | $n_{0.5\sigma_0}$ | $n_{\sigma_0}$ |
|---|---|---|---|---|---|---|
| 15 | 785 | 197 | 88 | 50 | 32 | 8 |

of the classical, we take $\alpha^*$ and $\beta^*$ to be 0.05 and 0.2, respectively, and use a total error bound of $\alpha = 0.25$ with a weight $w = 0.5$ for our method. We use the same prior parameters as in the first specification of Example 1. Table 5 compares the resulting sample sizes for various choices of $d$. Our method compares favorably for this specification. Observe that our sample size was derived assuming that the value of the distribution of the mean under the alternative hypothesis was centered about the null value $\theta_0$.

## 3.6 Choice of Weight $w$

It is clear that the resulting sample size and average errors are dependent upon the choice of the weight $w$. While the choice of $w$ may be study dependent or left up to the preference of the user, we offer two general guidelines.

First, we note that in each example the smallest sample size was obtained when $w = 0.5$. This is a general result, which is obtained as a direct consequence of Corollary 1.

**Corollary 2.** *Let $T(X) = \log(BF(X))$ for observed data $X \sim f(X|\theta)$, with prior $\pi(\theta)$ such that for $j = 0, 1$, $Pr(\theta \in \Theta_j) > 0$ for the hypothesis test*

$$H_0 : \theta \in \Theta_0 \quad vs. \quad H_1 : \theta \in \Theta_1.$$

*Then, $t_0 = 0$ minimizes $TE$ where $TE(t) = AE_1(t) + AE_2(t)$.*

This result is easily seen by recognizing that $\text{TE}(t)$ can be expressed as $\text{TWE}(t, 0.5)$, and applying Corollary 1 we have that choosing a weight of $w = 0.5$ gives an optimal cutoff of $t_0 = 0$, minimizing TE. The formulation of the rule for sample size determination thus results in the smallest sample size when $w = 0.5$. If the prior probability of each hypothesis is equivalent, then $w = 0.5$ is also equivalent to rejecting the null hypothesis when the posterior probability of $H_1$ is greater than the posterior probability of $H_0$. Choosing $w = 0.5$ seems a good rule of thumb if there is no strongly preferred bound on $\text{AE}_1$ or $\text{AE}_2$.

When there is a preferred bound on $\text{AE}_1$, for example, one can choose a weight $w$ that ensures the bound is satisfied. Observe that

$$w\text{AE}_1(t_0(w)) \leq \text{TWE}(t_0(w), w) \leq 1 - w, \tag{3.1}$$

where the second inequality comes from recognizing that TWE can be written as

$$\text{TWE}(t, w) = (1 - w) - (1 - w) \int \mathbb{I}(T(X) > t) \left( \text{BF}(x) - \frac{w}{1 - w} \right) m_0(x) dx.$$

From (3.1), we then see that choosing $w = 1/(1 + \alpha^*)$ ensures that $\text{AE}_1(t_0(w)) \leq \alpha^*$. For example, to guarantee that $\text{AE}_1 < 0.05$, we set $w = 1/1.05 = 0.953$. From our examples, we see that this is not a tight bound. That is, a weight of $w = 0.95$ often gave $\text{AE}_1$ values much lower than 0.05. In fact, in the examples we considered $w = 0.5$ was often sufficient for attaining an $\text{AE}_1$ less than 0.05.

# 4 Application to a Safety Study

Statins are a class of medications commonly prescribed to lower cholesterol. A recent clinical trial was conducted to determine the efficacy of rosuvastatin therapy for lowering cholesterol in children with familial hypercholesterolemia (Avis et al. 2010). The results of the study indicated that treatment with a 20mg dose of rosuvastatin was effective in lowering cholesterol compared to placebo. In addition, there was no indication that the drug resulted in more adverse events than placebo; however, the study was not powered on these secondary safety endpoints.

Suppose an investigator wishes to conduct two follow-up studies to assess the safety of rosuvastatin in children. The first study is aimed at investigating if children treated with rosuvastatin (20mg dose) are at higher risk for adverse events compared to a placebo group. Avis et al. (2010) reported that 54% and 55% of children experienced adverse events in the placebo and rosuvastatin group, respectively. Using the prior from Section 3.4, the investigator chooses the prior parameters so that under the null hypothesis the mean response rate is 0.545 with a variance of 0.125; under the alternative hypothesis, the mean response rate is 0.54 (0.55) with a variance of 0.125 for the placebo (rosuvastatin) group. The prior probability of the null hypothesis $u$ is taken to be 0.5. Following the specifications of the original trial, the investigator considers a total error bound of 0.15. Using a weight of 0.5, the required sample size is 243 subjects for each treatment arm, yielding an $\mathrm{AE}_1 = 0.021$ and $\mathrm{AE}_2 = 0.129$.

In the second study, the investigator wishes to undertake a single arm study of children treated with the 20mg dose of rosuvastatin to determine if the treatment impairs renal function. The investigator considers the change in Glomerular Filtration Rate (GFR) from baseline through 12 weeks of treatment as the response. A drop in GFR can indicate renal abnormalities; so, the investigator wishes to test the hypothesis that on average, the change in GFR is different from 0. As the variance is unknown, he considers the following prior:

$$\begin{aligned}\pi(\theta) =& u\mathbb{I}(\theta = \theta_0) IG_{(a,b)}(\sigma^2) + \\ & (1-u)\mathbb{I}(\theta \neq \theta_0)(c\sigma)^{-1}\phi\left(\frac{\theta-\mu}{c\sigma}\right) IG_{(a,b)}(\sigma^2),\end{aligned}$$

where $IG_{(a,b)}(x)$ indicates the InverseGamma density function with shape parameter $a$ and rate parameter $b$. Using the results of Avis et al. (2010) as a guide, the investigator chooses to consider $a = 48$ and $b = 9604$, giving a mode of 196 for the population variance. Also, the investigator chooses to set the prior mean $\mu = 0$ and scaling constant $c = 0.5$ indicating the prior standard deviation is one half that of the population standard deviation. Taking the total error bound to be 0.15 and using a weight of 0.5, the required sample size is 802, giving an $\text{AE}_1 = 0.035$ and $\text{AE}_2 = 0.114$.

# 5 Discussion

Using Bayesian average errors, we have presented a general approach to hypothesis testing and sample size determination that is broadly applicable. This method does not suffer from limitations such as powering the study on a single value of the parameter under the alternative or applying normal approximations.

While we have written all the computations in terms of the marginal distributions under each hypothesis, the errors can also be expressed in terms of the prior, posterior, and marginal (of $X$) distributions. When these distributions are not available in closed form, numerical methods may be required. However, we have found that under reasonable priors, the marginal distributions are analytically tractable for several designs common to medical studies involving binomial and normal distributions.

There is an increased computational burden as the sample size increases for complex hypotheses. However, the observed relationship between $\log n$ and $\log \alpha$ may prove useful in estimating a reasonable sample size and decreasing computation time. This relationship has been observed in each example we have considered. More work is needed to better understand the relationship between the total error bound and the resulting sample size.

Though the Bayes factor appears the optimal choice of $T(X)$ in some sense (see Appendix A), it is known that the Bayes factor is sensitive to the choice of the prior (Robert 2001). For many applications, sensitivity to the choice of the prior is often overcome through noninformative reference priors (Robert 2001); however, these priors are often improper. Recall that an assumption of our method ($Pr(\theta \in \Theta_j) > 0$ for $j = 0, 1$) does not permit the use of improper priors. One possible extension of our method is choosing $T(X)$ to be an alternative Bayes factor, which has been shown to be robust to the choice of the prior (De Santis 2007).

The errors we consider are conditional on the hypothesis. Some argue this is unnatural in the Bayesian framework. Lee and Zelen (2000) considered an approach to hypothesis testing using errors conditional on the result of the trial. It would be of interest to consider modifying our approach to define the total weighted error as the weighted sum of the errors in Lee and Zelen (2000).

We have considered the choice of $w$ and $\alpha$ to be fixed throughout our discussion. The bound $\alpha$ may be determined by a regulating agency. Most likely, $\alpha$ and $w$ will be predicated on the available funds for the study. Given a total error $\alpha$ one can afford, the best choice of $w$ remains an open question - though we have given some guidance in Section 3.6.

The R package `BAEssd` implements these methods for several examples involving binary and normal responses, which is available on the author's website.

# References

Adcock, C. J. (1997). Sample size determination: a review. *Journal of the Royal Statistical Society. Series D (The Statistician)* **46,** 261–283.

Avis, H. J., Hutten, B. A., Gagné, C., Langslet, G., McCrindle, B. W., Wiegman, A., Hsia, J., Kastelein, J. J. P., and Stein, E. A. (2010). Efficacy and safety of rosuvastatin therapy for children with familial hypercholesterolemia. *Journal of the American College of Cardiology* **55,** 1121–1126.

Chow, S. C., Shao, J., and Wang, H. (2003). *Sample Size Calculations in Clinical Research.* Marcel Dekker Inc, 1 edition.

De Santis, F. (2004). Statistical evidence and sample size determination for bayesian hypothesis testing. *Journal of Statistical Planning and Inference* **124,** 121–144.

De Santis, F. (2007). Alternative bayes factors: Sample size determination and discriminatory power assessment. *Test* **16,** 504–522.

Fox, D. R., Ben-Haim, Y., Hayes, K. R., McCarthy, M. A., Wintle, B., and Dunstan, P. (2007). An info-gap approach to power and sample size calculations. *Enviornmetrics* **18,** 189–203.

Friedman, L. M., Furberg, C. D., and DeMets, D. L. (1998). *Fundamentals of clinical trials.* Springer, 3 edition.

Inoue, L. Y. T., Berry, D. A., and Parmigiani, G. (2005). Relationship between bayesian and frequentis sample size determination. *American Statistician* **59,** 79–87.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90,** 773–795.

Lee, S. J. and Zelen, M. (2000). Clinical trials and sample size considerations: another perspective. *Statistical Science* **15,** 95–103.

M'Lan, C. E., Joseph, L., and Wolfson, D. B. (2008). Bayesian sample size determination for binomial proportions. *Bayesian Analysis* **3,** 269–296.

Pezeshk, H. (2003). Bayesian techniques for sample size determination in clinical trials: a short review. *Statistical Methods in Medical Research* **12,** 489–504.

Robert, C. P. (2001). *The bayesian choice.* Springer, 2 edition.

Röhmel, J. and Mansmann, U. (1999). Unconditional non-asymptotic one-sided tests for independent binomial proportions when the interest lies in showing non-inferiority and/or superiority. *Biometrical Journal* **41,** 149–170.

Weiss, R. (1997). Bayesian sample size calculations for hypothesis testing. *Journal of the Royal Statistical Society. Series D (The Statistician)* **46,** 185–191.

# Appendix A: Optimal Test Function

**Theorem 1.** *Consider testing the hypothesis as described in Section 2.1. Let $BF(X)$ denote the Bayes factor and let*

$$\varphi(X) : x \to [0,1]$$

*represent a randomized test for the hypothesis. Then, for a given value of $w \in (0,1)$, $\hat{\varphi}(X)$ minimizes TWE where*

$$\hat{\varphi}(X) = \mathbb{I}(BF(X) > \frac{w}{1-w}).$$

*Proof.* Observe that $\mathrm{AE}_1$ and $\mathrm{AE}_2$ for the randomized test $\varphi(X)$ are defined as

$$\begin{aligned}\mathrm{AE}_1 &= E\left[\varphi(X)|\theta \in \Theta_0\right] \quad \text{and} \\ \mathrm{AE}_2 &= E\left[1-\varphi(X)|\theta \in \Theta_1\right].\end{aligned}$$

As $m_j(x)$ represents the marginal distribution of $X$ under $H_j$, we have that

$$\begin{aligned}\mathrm{AE}_1(\varphi) &= \int \varphi(x) m_0(x) dx \quad \text{and} \\ \mathrm{AE}_2(\varphi) &= \int \left(1-\varphi(x)\right) m_1(x) dx.\end{aligned}$$

Now, observe that for a given $w \in (0,1)$ we have that

$$\begin{aligned}\mathrm{TWE}(\varphi) &= w\mathrm{AE}_1(\varphi) + (1-w)\mathrm{AE}_2(\varphi) \\ &= w\int \varphi(x) m_0(x) dx + (1-w)\int \left(1-\varphi(x)\right) m_1(x) dx \\ &= w\int \varphi(x) m_0(x) dx + (1-w) - (1-w)\int \varphi(x)\mathrm{BF}(x) m_0(x) dx \\ &= (1-w) - (1-w)\int \varphi(x)\left(\mathrm{BF}(x) - \frac{w}{1-w}\right) m_0(x) dx.\end{aligned}$$

As $w$ is fixed, minimizing TWE with respect to $\phi(\cdot)$ is equivalent to maximizing

$$\int \varphi(x) \left( \mathrm{BF}(x) - \frac{w}{1-w} \right) m_0(x) dx.$$

Recall that $\hat{\varphi}(X)$ is given by

$$\hat{\varphi}(X) = \mathbb{I}(\mathrm{BF}(X) > \frac{w}{1-w}).$$

Let $X = x_1$ where $x_1$ is in the support of $X$ under $H_0$, which ensures $m_0(x_1) > 0$ and hence $\mathrm{BF}(x_1)$ is well defined. Now, observe that

$$\varphi(x_1) \left( \mathrm{BF}(x_1) - \frac{w}{1-w} \right) \leq \hat{\varphi}(x_1) \left( \mathrm{BF}(x_1) - \frac{w}{1-w} \right).$$

Since the choice of $x_1$ was arbitrary, we have that this result holds for all $x$ such that $m_0(x) > 0$. Thus, we then have that

$$\int \varphi(x) \left( \mathrm{BF}(x) - \frac{w}{1-w} \right) m_0(x) dx \leq \int \hat{\varphi}(x) \left( \mathrm{BF}(x) - \frac{w}{1-w} \right) m_0(x) dx.$$

That is, the integral of interest is maximized for any choice of $\varphi(X)$ when $\varphi(X) = \hat{\varphi}(X)$. Thus, $\hat{\varphi}(X)$ minimizes $\mathrm{TWE}(\varphi)$. □