METRON - *International Journal of Statistics*
2003, vol. LXI, n. 2, pp. 243-266

II BOZZE

File: MQ.2003.10 tex
23.settembre .2003

ALESSIO FARCOMENI

Unified conditional frequentist and Bayesian testing: computations in practice and sample size determination

Summary - Hypothesis testing is one of the areas of inference in which Bayesian and frequentist methods tend to disagree. Berger *et al.* (1994) proposed a modification of the usual Bayesian testing, which made it acceptable from a conditional frequentist point of view. In this article, we propose a simple numerical method to avoid heavy computations when applying their test. We motivate this by showing the case of testing the parameter of an exponential random variable. We are confident that this simplification will encourage people in using the test in applications. Finally, a procedure for choosing the sample size for this test is proposed.

Key Words - Hypothesis testing; Bayesian methods; Conditional inference; Frequentist methods; Simulation; Sample size determination.

1. Introduction

Hypothesis testing is one of the areas of inference in which Bayesians and frequentists tend to disagree. The classical frequentist approach builds a rejection region from fixed probability of incorrect rejection (Type I error), usually denoted by α. After the test, one reports this pre-experimental error measure; sometimes together with the dual, the probability of incorrect acceptance (Type II error), usually denoted by β. This practice has been criticized for ignoring experimental evidence provided by the data for or against the null hypothesis. A common alternative is the p-value, but unfortunately it can be misleading as a measure of evidence. See for instance Berger and Delampady (1987), and Royall (1997) for a discussion on the concept of statistical evidence. More recently, Sellke *et al.* (2001) discussed many issues involving the use of the p-values in statistical analysis. Kiefer (1977) has fulfilled the need for data dependent error probabilities, through conditioning on a well chosen statistic. Suppose $S(x)$ is some statistic where larger values indicate data with

Received January 2003 and revised July 2003.

greater evidential support for (or against) the null hypothesis. Conditional error probabilities are given by:

$$\begin{aligned} \alpha(s) &= \Pr\ (\text{Type I error } | S(x) = s) \\ \beta(s) &= \Pr\ (\text{Type II error } | S(x) = s) \end{aligned} \tag{1}$$

For a discussion on the conditional frequentist approach, see Brownie and Kiefer (1977). For a review on conditioning, see for instance Frosini (1999). On the other side, Bayesian methods are usually criticized for lack of objectivity, but naturally report and base the decision on measures of evidence, like posterior probabilities of hypotheses, and the Bayes factor.

Berger *et al.* (1994) devised a conditioning statistic $S(x)$, such that, outside a non-decision region (NDR), the conditional frequentist test of a precise hypothesis and the usual Bayesian test would agree both in the action to be taken (acceptance or rejection of H_0) and in the reported error probabilities. i.e., Bayesians and (conditional) frequentists will always make the same decision, and report the same numbers; even if these are assigned different interpretations. The NDR is simply a subset of the sample space that does not allow one to make a choice between the hypotheses. In a sense, it's the "price" to pay for the agreement between the methods.

The procedure is easily applied: one needs only to compute the lower and upper bounds of the NDR, r and a; and then reject H_0 if the test statistic is smaller than r, or retain H_0 if it is bigger than a. We will describe this later in more detail, together with considerations on the error probabilities.

In general, exact determination of r and a involves tedious computation, as we will argue when showing the extension of the procedure to testing the mean of an exponential distribution. In this paper, we show a simple method to obtain the thresholds of the NDR via simulation. This simulation method draws independent samples from distributions ready at hand, thus making the procedure straightforward. We hope the simulation method will stimulate people to use the modified test. Reporting the thresholds, even when using a full-Bayesian or unconditional-frequentist approach, would suggest if and when the two methodologies are converging; thus enriching the analysis and making it more widely acceptable.

The remaining of the paper is as follows: in Section 2 we will review the modified Bayesian-frequentist test. Section 3 will extend the procedure to testing the parameter of an exponential distribution, both in the simple and composite alternative case. Section 4 will outline the simulation method for computing the bounds of the NDR. In Section 5 we also propose two methods of choice of the optimal sample size for the modified test. We follow De Santis (2003), and show that the modified test avoids some known problems associated with that methodology.

2. The modified Bayesian-frequentist test

Let $f(x|\theta)$ be the density of the data, depending on the unknown parameter of interest $\theta \in \Theta$; define the likelihood function $l(\theta) = f(x|\theta)$ (x is the observed data, and thus fixed). Let $\pi_1(\theta)$ be the prior distribution assigned on the space Θ.

We use a 0-1 loss function (see Bernardo and Smith (1994)) with equal prior probabilities for the two hypotheses. Berger *et al.* (1994) builds the procedure for generalized 0-1 losses and/or unequal prior probabilities. In a further subsection we will comment on the extension to the case of unequal prior probabilities. We will always test a precise null hypothesis $\theta = \theta_0$. The alternative will be either precise or two-sided (no one-sided problem is considered).

In the Bayesian framework, the Bayes factor is commonly used for testing (see for instance Kass and Raftery (1995)). The marginal distribution under the null will be $m_0(x) = l(\theta_0)$, while the marginal under the alternative hypothesis is $m_1(x) = l(\theta_1)$ for a precise alternative and $m_1(x) = \int\limits_{\Theta} l(\theta)\pi_1(\theta)\, d\theta$ for the composite alternative. Define the Bayes factor to be: $B(x) = \frac{m_0(x)}{m_1(x)}$. The Bayes factor compares the two marginals, to determine which hypothesis is more supported by the data.

It is straightforward to see that the posterior probabilities of the two hypotheses are:

$$\Pr(H_0|x) = \alpha^*(x) = \frac{B(x)}{1+B(x)}; \tag{2}$$

and of course

$$\Pr(H_1|x) = \beta^*(x) = \frac{1}{1+B(x)}. \tag{3}$$

Here we define the posterior error probability as the posterior probability of the rejected hypothesis.

The usual Bayesian test consists in computing the Bayes factor $B(x)$ and then:

$$\begin{cases} B(x) \le 1, & \text{reject } \ H_0, \ \text{report } \alpha^*(x) \\ B(x) \ge 1, & \text{retain } \ H_0, \ \text{report } \beta^*(x) \end{cases}$$

We will now introduce the modified test, as it is proposed in Berger *et al.* (1997). Let $F_i(s) = \Pr(B(x) < s|H_i),\ i = 0, 1$ be the CDF of the Bayes factor under H_i, which we suppose continuous and invertible. Define the function

$$\psi(s) = F_0^{-1}(1 - F_1(s)), \tag{4}$$

and the statistic

$$S(x) = \min\{B(x), \psi^{-1}(B(x))\} \tag{5}$$

Define the lower and upper bound of the NDR as:

$$\begin{aligned} r &= \min(1, \psi^{-1}(1)) \\ a &= \max(1, \psi(1)) \end{aligned} \tag{5'}$$

Then, the unified Bayesian-frequentist test will be done as follows:

$$\begin{cases} B(x) \leq r, & \text{reject } H_0 \\ r < B(x) < a, & \text{do nothing,} \\ B(x) \geq a, & \text{retain } H_0 \end{cases} \tag{6}$$

where $B(x)$ is the Bayes factor.

Berger *et al.* (1994) shows that frequentist error probabilities conditioned on $S(x)$ will be equal to the posterior probabilities when both H_0 and H_1 are simple, outside the NDR.

Berger *et al.* (1997) shows that when H_1 is composite, the conditional Type I error probability $\alpha(s)$ will be equal to the posterior probability of H_1 $\alpha^*(x)$, outside the rejection region. The posterior probability of H_0 will be equal to an opportune expected value of the Type II conditional error probability: $\int_\Theta \beta(\theta|s)\pi_1(\theta|s)\, d\theta = \beta^*(x)$, where $\pi_1(\theta|s)$ is the posterior p.d.f. under H_1 on the observed value s of $S(X)$.

The equality of $\alpha(s)$ and $\alpha^*(x)$ is the key point in these two results: Type I error is usually perceived in classical statistics as the primary error measure. In this sense, this agreement is crucial in the acceptance of the procedure.

Surprisingly, it is never necessary to actually compute the conditioning statistic $S(x)$.

To summarize, the motivation of introducing a non-decision region in the Bayesian test is two-fold. On one hand, we want conditional frequentist error probabilities, at least of the Type I error, to coincide with (Bayesian) posterior error probabilities. On the other hand, from a Bayesian point of view, it is not acceptable to reject an hypothesis with a posterior error probability greater than 50%. It is straigthforward to see that the smallest set in which the two things won't happen is the NDR.

Berger *et al.* (1994) introduced the modified test in the more general case, for the simple hypotheses case. Berger *et al.* (1997) extended the procedure to the composite alternative with 0-1 loss, equal prior probabilities. They also showed the technique for the one-sided and two-sided normal testing, for comparing normal means, with known and unknown variance. Dass and Berger (1999) provided results in the case of composite null hypothesis. Berger *et al.* (1998) proposed an efficient way to apply the modified test in the case of sequential testing of nested hypotheses.

2.1. Unequal prior probabilities of the hypotheses

It is not always the case that prior probabilities of the hypotheses can be assumed to be equal. Following Berger *et al.* (1994), it is possible to generalize the test as follows: let p_{H_0} be the prior probability of H_0. Obviously, $p_{H_1} = 1 - p_{H_0}$ will be the prior probability assigned to H_1. Let $\eta = \frac{p_{H_1}}{p_{H_0}}$. It is straightforward to see that the posterior probabilities will be

$$\alpha_\eta(B(x)) = B(x)/(\eta + B(x))$$

for H_0 and

$$\beta_\eta(B(x)) = 1/(\eta + B(x))$$

for H_1. Berger *et al.* (1994) show that the generalized thresholds are:

$$\begin{cases} r_\eta = \eta \text{ and } a_\eta = F_0^{-1}(1 - \eta F_1(\eta)) \text{ if } \ F_0(\eta) \leq 1 - \eta F_1(\eta), \\ r_\eta = F_1^{-1}(1/\eta[1 - \eta F_0(\eta)]) \text{ and } a_\eta = \eta \text{ if } \ F_0(\eta) > 1 - \eta F_1(\eta). \end{cases}$$

Then, the test will be as usual:

$$\begin{cases} B(x) \leq r_\eta, & \text{reject } H_0 \text{ with error probability } \alpha_\eta(B(x)) \\ r_\eta < B(x) < a_\eta, & \text{no decision} \\ B(x) \geq a_\eta, & \text{retain } H_0 \text{ with error probability } \beta_\eta(B(x)), \end{cases}$$

Berger *et al.* (1994) show that the posterior error probabilities, outside the no decision region, are exactly equal to the error probabilities conditioned on the statistic $S(x)$ defined in (5). Note that the choice of η, the ratio of the prior probabilities, has a strong effect on the posterior error probabilities and the thresholds. We will discuss further results on the case of unequal prior probabilities of the hypotheses in Section 5.

3. Testing the parameter of an exponential distribution

In this section we will show the modified test applied to testing the parameter of an exponential distribution.

3.1. Simple hypotheses

Let $(x_1, \ldots, x_n)$ be a vector of i.i.d. random variables, conditional to an unknown parameter θ; such that $f(x_i|\theta) = \theta e^{-x_i\theta}$, $x_i \geq 0$, $\theta \geq 0$.

Let us test the simple versus simple hypotheses, let s be the sufficient statistic $s = \sum_{i=1}^{n} x_i$, and $\lambda = \theta_0/\theta_1$, with of course $\lambda \neq 1$ for identifiability reasons.

It is straightforward to show that $B(s) = \lambda^n e^{-(\theta_0-\theta_1)s}$. We will now need to compute $\psi(1)$ and maybe $\psi^{-1}(1)$ as defined in (4), i.e., we will need the two cdfs of the Bayes factor $F_0(s)$ and $F_1(s)$, and their inverses; to determine r and a as in (6). Let s_b, for b fixed in R, be the solution to the equation

$$B(s) = b, \tag{7}$$

i.e.[1] $s_b = \dfrac{\ln(\frac{\lambda^n}{b})}{\theta_0 - \theta_1}$.

We have that $2s\theta$ is a chi-square random variable with $2n$ degrees of freedom, hence

$$F_0(b) = \begin{cases} H\left(2\lambda \ln\left(\dfrac{\lambda^n}{b}\right)/(\lambda-1)\right) & \text{if } \lambda < 1 \\ 1 - H\left(2\lambda \ln\left(\dfrac{\lambda^n}{b}\right)/(\lambda-1)\right) & \text{if } \lambda > 1 \end{cases} \tag{8}$$

while

$$F_1(b) = \begin{cases} H\left(2 \ln\left(\dfrac{\lambda^n}{b}\right)/(\lambda-1)\right) & \text{if } \lambda < 1 \\ 1 - H\left(2 \ln\left(\dfrac{\lambda^n}{b}\right)/(\lambda-1)\right) & \text{if } \lambda > 1\,, \end{cases} \tag{9}$$

where $H(x)$ is the CDF of a chi square with $2n$ d.f.

On the other hand,

$$F_0^{-1}(k) = \begin{cases} \lambda^n \exp\left\{-(\lambda-1)\dfrac{H^{-1}(k)}{2\lambda}\right\} & \text{if } \lambda < 1 \\ \lambda^n \exp\left\{-(\lambda-1)\dfrac{H^{-1}(1-k)}{2\lambda}\right\} & \text{if } \lambda > 1 \end{cases}$$

and

$$F_1^{-1}(k) = \begin{cases} \lambda^n \exp\left\{-(\lambda-1)\dfrac{H^{-1}(k)}{2}\right\} & \text{if } \lambda < 1 \\ \lambda^n \exp\left\{-(\lambda-1)\dfrac{H^{-1}(1-k)}{2}\right\} & \text{if } \lambda > 1\,. \end{cases}$$

It is immediate to see that, $\forall\ \lambda \neq 1$, $\psi(\cdot) = F_0^{-1}(1 - F_1(\cdot))$ is:

$$\psi(b) = \lambda^n \exp\left\{-\frac{1}{2}\frac{(\lambda-1)}{\lambda}H^{-1}\left(1 - H\left(\frac{2\ln\left(\frac{\lambda^n}{b}\right)}{\lambda-1}\right)\right)\right\},$$

[1]Note that, if $\theta_0 > \theta_1$, the equation is solved $\forall\ 0 < b < \lambda^n$. If $\theta_0 < \theta_1$, the equation is solved $\forall b > \lambda^n$, but in that case $\lambda < 1$ and $\lim\limits_{n\to+\infty} \lambda^n = 0$.

and

$$\psi^{-1}(b) = \lambda^n \exp\left\{-\frac{1}{2}(\lambda-1)H^{-1}\left(1 - H\left(\frac{2\lambda \ln\left(\frac{\lambda^n}{b}\right)}{\lambda - 1}\right)\right)\right\}.$$

In practice, to find r and a, one only needs to evaluate this last two functions in $b = 1$; then compute r and a as in (5′) and do the test as in (6).

The next theorem shows that only cases in which $\lambda > 1$, i.e. $\theta_0 > \theta_1$, are to be considered in this case; since the case in which $\lambda < 1$ is symmetric and can be derived from the first.

Theorem 1. *Let $\psi(s, l)$ be $\psi(s)$ computed when $\lambda = l$, and $F_i(s, l)$ the CDF of $B(s)$ conditional on H_i being true and $\lambda = l$. Then, $\psi(1, \lambda) = (\psi^{-1}(1, 1/\lambda))^{-1}$, and $F_0(1, \lambda) = 1 - F_1(1, 1/\lambda)$*

The theorem implies that, when $\lambda < 1$, if r and a are computed for $1/\lambda$, the NDR is: $\{x \in \mathcal{X}/\ 1/a < B(x) < 1/r\}$. For a proof of the theorem, see Appendix.

Example 1. Let us illustrate the results in Theorem 1 with an example. Let $n = 10$ and the system of hypotheses be:

$$\begin{cases} H_0 : \theta_0 = 9 \\ H_1 : \theta_1 = 3\,. \end{cases}$$

In this case, $\lambda = 3$, so $r = 1$ and $a = 1.83$. If we want to test the hypotheses:

$$\begin{cases} H_0 : \theta_0 = 0.1 \\ H_1 : \theta_1 = 0.3\,, \end{cases}$$

with a sample of $n = 10$ elements; $\lambda = 1/3$ so, by Theorem 1, $r = 1/1.83 = 0.546$ and $a = 1$.

Table 1 shows values of r and a for some n and λ. It is interesting to notice that, for these values of λ, the size of the NDR is practically constant with respect to n.

TABLE 1. *Values of r and a for simple-simple testing of an exponential parameter.*

n	$\lambda = 1.5$		2		3	
1	$r = 1$	$a = 1.23$	$r = 1$	$a = 1.41$	$r = 1$	$a = 1.69$
5	$r = 1$	$a = 1.28$	$r = 1$	$a = 1.49$	$r = 1$	$a = 1.82$
10	$r = 1$	$a = 1.28$	$r = 1$	$a = 1.50$	$r = 1$	$a = 1.83$
20	$r = 1$	$a = 1.29$	$r = 1$	$a = 1.51$	$r = 1$	$a = 1.84$
30	$r = 1$	$a = 1.29$	$r = 1$	$a = 1.51$	$r = 1$	$a = 1.84$
50	$r = 1$	$a = 1.29$	$r = 1$	$a = 1.51$	$r = 1$	$a = 1.84$
60	$r = 1$	$a = 1.29$	$r = 1$	$a = 1.51$	$r = 1$	$a = 1.86$

3.1.1 –Probability of the NDR

In many applications, the size of the NDR is non decreasing with n. On the other hand, Berger *et al.* (1997) note that, under certain conditions, the probability of falling into the NDR is rapidly decreasing with n:

$$\Pr(r < B(x) < a|H_i) \cong e^{-nI} \xrightarrow{n} 0, \quad i = 0, 1$$

where $I = -\log \inf_{0\leq t\leq 1} \int m_0^t(x) m_1^{1-t}(x)\, dx$.

In this subsection we will numerically analyze the probability of falling in the non decision region in the exponential case with simple hypotheses.

Using (8) and (9), for the chosen values of λ ($\psi(1) > 1$ in all cases) we have:

$$\begin{aligned}\Pr(1 < B(x) < a|H_0) &= F_0(a) - F_1(1) = 1 - F_0(1) - F_1(1) = \\ &= H\left(\frac{2n\lambda \ln\lambda}{\lambda - 1}\right) + H\left(\frac{2n \ln\lambda}{\lambda - 1}\right) - 1,\end{aligned}$$

where $H(\cdot)$ is the CDF of a chi-square with $2n$ degrees of freedom. To compute $\Pr(1 < B(x) < a|H_1)$ it is necessary to proceed numerically, since a closed form expression for $F_1(a)$ is not available.

Figure 1 shows the behavior of the probability of the NDR under H_0. This probability is almost zero for $n > 20$. Even though the curves for different values of λ are very close, the bigger λ, the smaller the probability of non deciding; since it is easier to discriminate between the hypotheses.

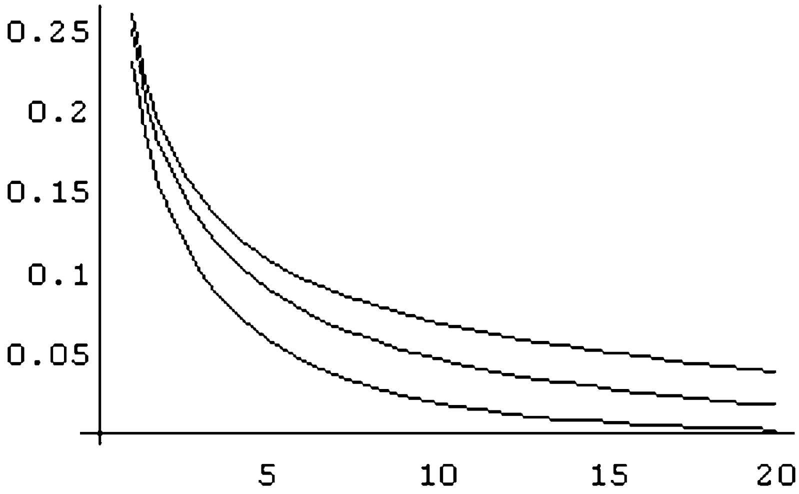

Figure 1. $\Pr(1 < B(x) < a|H_0)$ for the exponential test, simple hypotheses, various λ

Figure 2 shows the probability of falling in the NDR under H_1, as a function of n, for $\lambda = 1.5$. The probability of non deciding when H_0 is false is much smaller than when it is true. This is intuitive, since the NDR is a part of the usual "acceptance" region ($a > 1$).

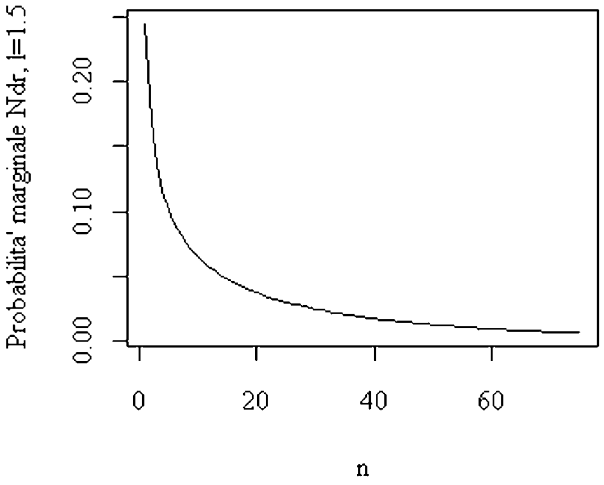

Figure 2. Pr($1 < B(x) < a|H_1$), for the exponential test, simple hypotheses, $\lambda = 1.5$

Figure 3 shows the marginal probability of falling in the NDR as a function of n.

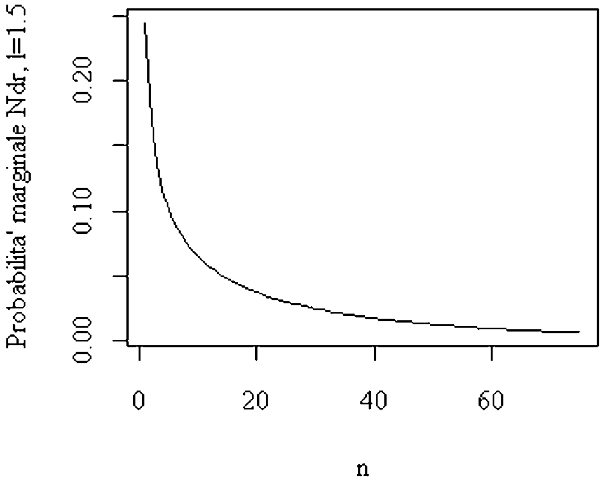

Figure 3. Pr($1 < B(x) < a$), for the exponential test, simple hypotheses, $\lambda = 1.5$

3.2. Composite alternative hypothesis

Let us now apply the modified Bayesian test to the same model as in 3.1, with a composite alternative hypothesis.

We will need to choose a prior for $\theta \in \Theta_1$. To simplify, we will use the conjugate prior[2], a $gamma(\delta, \lambda)$. It is straightforward to see that under H_1, the sum of the values of the observations s is a $gamma - gamma(\delta, \lambda, n)$:

$$m_1(s) = \frac{\lambda^{\delta} s^{n-1} \Gamma(n+\delta)}{\Gamma(\delta)\Gamma(n)(\lambda+s)^{n+\delta}}.$$

For definition and properties of the gamma-gamma see for instance Piccinato (1996) or Bernardo and Smith (1994). The Bayes factor is then:

$$B(s) = \frac{\theta_0^n e^{-\theta_0 * s} \Gamma(\delta)(\lambda+s)^{n+\delta}}{\lambda^{\delta}\Gamma(n+\delta)}.$$

3.2.1 –Choice of prior parameters

Prior parameters λ and δ will be chosen using prior information.

One possibility is to make the prior mean to be equal to the expected value of $x_i|H_0$. The condition is $\frac{1}{\theta_0} = \frac{\delta}{\lambda}$, i.e.,

$$\lambda = \theta_0 \delta. \tag{10}$$

Modifications to this condition are straightforward.

Another sensible approach is to choose how much strength to give to the prior information with respect to one observation. If $1/k$ is this proportion, one needs only fix the remaining parameter, δ, such that prior variance is k times the variance of $x_i|H_0$. I.e., $\frac{k}{\theta_0} = \frac{\delta}{\lambda^2}$, that is, given condition (10), $\delta = \frac{1}{k}$. It is customary to set k to one (see for instance Kass and Wasserman (1995)), but any positive value is possible. Usually, one would set $k \geq 1$.

As an example, figure 4 shows the Bayes factor as a function of x with $\theta_0 = 1,\ \delta = \lambda = 2/3,\ n = 10$.

[2]Note that, in this section, λ is something completely different from what indicated previously with the same symbol.

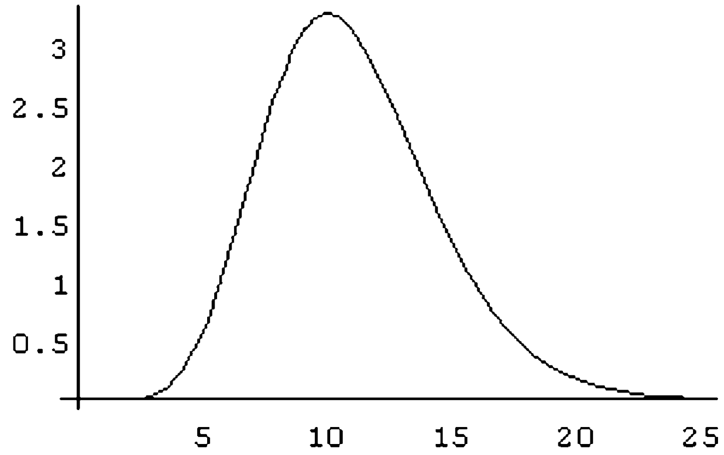

Figure 4. $B(s)$ in the exponential test, composite alternative.

3.2.2 –Determination of the Thresholds

Exact determination of the thresholds r and a is as follows: Let $s_{b,1}$ and $s_{b,2}$ be the two admissible solutions of the equation $B(s) = b$, $s_{b,1} \leq s_{b,2}$. If there is only one, or none solution, the missing is set to zero in the following formulas.

Since $F_0(a) = \Pr(B(s) < a| \ H_0) = \Pr(s < s_{a,1}|H_0) + \Pr(s > s_{a,2}|H_0)$ we have that: $F_0(a) = H(2s_{a,1}\theta_0) + 1 - H(2s_{a,2}\theta_0)$; and, $F_1(b) = GG(s_{b,1}) + 1 - GG(s_{b,2})$, where $GG(s)$ is the CDF of a $gamma - gamma(\delta, \lambda, n)$. Hence, solving the equation $F_0(a) = 1 - F_1(1)$ is equivalent to solve: $H(2\theta_0 s_{a,1}) + 1 - H(2\theta_0 s_{a,2}) = GG(s_{1,2}) - GG(s_{1,1})$,

We need to compute $\psi(1)$ by solving the system:

$$\begin{cases} B(s) = 1 \\ B(s) = \psi(1) \\ H(2\theta_0 s_{\psi(1),1}) + 1 - H(2\theta_0 s_{\psi(1),2}) = GG(s_{1,2}) - GG(s_{1,1}) \end{cases} \tag{11}$$

If $\psi(1) > 1$, fix $r = 1$ and $a = \psi(1)$. Otherwise, one needs to compute $\psi^{-1}(1)$ by solving the system

$$\begin{cases} B(s) = 1 \\ B(s) = r \\ H(2\theta_0 s_{1,1}) + 1 - H(2\theta_0 s_{1,2}) = GG(s_{r,2}) - GG(s_{r,1}), \end{cases}$$

and fix $r = \psi^{-1}(1)$, $a = 1$.

Note that all the computations are much more complex than in the case where both hypotheses are simple. In general, it is evident that there are two drawbacks to exact calculation of the thresholds. First of all, it is very rare to find a closed form expression for $s_{b,j}$, especially when it is not unique, and so one must proceed numerically. Moreover, the number of solutions $s_{b,j}$ varies case by case, and so the way they enter into the computation of $F_i(\cdot)$. This means that a unifying general procedure cannot be given, coded, and used in practice.

Note also that this problem makes it really hard to study the robustness with respect to the prior. Each time the form of the prior is changed, the researcher must start over from scratch.

For this reason, a general approximate method, based on simulation, is proposed in Section 4, that will make it straightforward to get the thresholds without having to go into tedious computations and programming, independently for the model assumed for the data and for the parameters.

Table 2 shows the exact values of r and a, when $\theta_0 = 1$, for some choices of n, δ, λ. Note that the size of the NDR is increasing with n, even though it is always reasonably small. Note moreover that there is strong dependence on the choice of the prior parameters. This is a well known feature of the modified test in the composite alternative case, noted also by Berger *et al.* (1997) in the case of normal distributed random variables. For fixed n, the size of the NDR is increasing as $\lambda = \delta$ decrease. This is because prior variance is increasing as prior parameters simultaneously decrease, making the prior distribution more "flat" and making it harder to discriminate between the hypotheses.

TABLE 2. *Exact values of r and a for the exponential test, composite hypothesis,* $\lambda = \delta$.

n	$\delta = 1$		2/3		1/2	
5	$r = 1$	$a = 2.03$	$r = 1$	$a = 2.31$	$r = 1$	$a = 2.53$
10	$r = 1$	$a = 2.42$	$r = 1$	$a = 2.73$	$r = 1$	$a = 3.01$
15	$r = 1$	$a = 2.70$	$r = 1$	$a = 3.01$	$r = 1$	$a = 3.28$
20	$r = 1$	$a = 2.89$	$r = 1$	$a = 3.18$	$r = 1$	$a = 3.43$
30	$r = 1$	$a = 3.16$	$r = 1$	$a = 3.51$	$r = 1$	$a = 3.73$
40	$r = 1$	$a = 3.42$	$r = 1$	$a = 3.61$	$r = 1$	$a = 4.11$
50	$r = 1$	$a = 3.57$	$r = 1$	$a = 3.85$	$r = 1$	$a = 4.17$

We also have dependence on the value of θ_0. Table 3 shows values of $\psi(1)$ for different θ_0, with $n = 20$ and $\delta = \lambda = 1$.

We have that $\psi(1)$ is slowly increasing as θ_0 goes far from 1. The decrease of a between $\theta_0 = 5$ and $\theta_0 = 10$ is because the equation $B(s) = 1$ has only one solution for(3) $\theta_0 > 8.3$, and two for $\theta < 8.3$

(3)To be precise, the equation $B(s) = 1$ has only one solution for $\theta_0 > \sqrt[n]{n!}$, about 8.3 when $n = 20$. Remember that $\delta = \lambda = 1$.

TABLE 3. *$\psi(1)$ as θ_0 changes, $n = 20$, $\lambda = \delta = 1$.*

θ_0	a
10	4.60
5	4.94
2	3.46
0.5	3.07
1/5	3.88
1/10	4.66

4. COMPUTING THE THRESHOLDS

4.1. *Exact Thresholds*

In this subsection we outline the general method, implicitly suggested by Berger *et al.* (1997), to compute exactly r and a.

One needs to compute $\psi(\cdot) = F_0^{-1}(1 - F_1(1))$. Since usually closed form expression for the Bayes factor CDFs $F_i(\cdot)$ are not available, one needs to consider the equation $B(x) = z$, as a function of z. In general this equation will have zero, one or two solutions; or even more. Let $(x_{z,1}, \ldots, x_{z,k})$ be the k solutions to the equation. Let $x_{z,0} = -\infty$ and $x_{z,k+1} = +\infty$. If k is even: $F_i(z) = \int_{x_{z,0}}^{x_{z,1}} m_i(y)\ dy + \ldots + \int_{x_{z,k}}^{x_{z,k+1}} m_i(y)\ dy$, while if k is odd: $F_i(z) = \int_{x_{z,0}}^{x_{z,1}} m_i(y)\ dy + \ldots + \int_{x_{z,k-1}}^{x_{z,k}} m_i(y)\ dy$.

Now we have an expression for $F_i(\cdot)$, but it is unlikely that a closed form expression will be achieved, and so for the inverse $F_i^{-1}(\cdot)$. Thus, $\psi(1)$ will be solution to the equation: $F_0(x) = 1 - F_1(1)$; which in almost all cases will be solved numerically.

4.2. *Simulated thresholds*

Simulation can avoid lots of computation and programming. It is immediate to see that, if $(x_1, \ldots, x_b)$ is an i.i.d. sample from $m_i(x)$, $(z_1 = B(x_1), \ldots, z_b = B(x_b))$ will be an i.i.d. sample from $F_i(z)$. One can then use the empirical distribution $\hat{F}_i(z) = \frac{1}{b}\sum 1_{\{z_i<z\}}$ to estimate $F_i(z)$. Let $\hat{\psi}(1)$ be the solution to the equation

$$\hat{F}_0(z) = 1 - \hat{F}_1(1). \tag{13}$$

By Glivenko-Cantelli, you have $\hat{F}_i(z) \xrightarrow{p} F_i(z)$ in probability. It is easy to use this result to prove that $\lim_{b\to+\infty} \hat{\psi}(1) = \psi(1)$. In the same way, if $\hat{\psi}^{-1}(1)$ is solution to the equation:

$$\hat{F}_0(1) = 1 - \hat{F}_1(z), \tag{14}$$

$\hat{\psi}^{-1}(1)$ gets closer and closer to $\psi^{-1}(1)$ as b increases.

One will almost always be able to sample from $m_0(x) = f(x|\theta_0)$, while sampling from $m_1(x)$ in the composite alternative case may be a harder problem. MCMC methods can be used to get a dependent sample from $F_1(\cdot)$. If the sample has ergodic properties, convergence in probability will still hold.

It is also possible that the value $F_i(1)$ is known. Of course, substituting the real value for the estimate in the equations (13) and (14) will speed up the convergence and provide more accurate results. Since in most cases it is enough to solve equation (13), and get $F_1(1)$ by one single numerical integration; MCMC methods will be rarely needed.

4.2.1 –An example

Let $(x_1, \dots, x_n)$ be an i.i.d vector with $x_i|\theta, \sigma^2 \sim N(\theta, \sigma^2)$, known σ. Suppose $H_1 : \theta \neq \theta_0$, for some θ_0.

Let $\pi_1(\theta)$ be a conjugate prior, i.e., $\theta \sim N(\theta_0, \xi\sigma^2)$, for some $\xi > 0$.

Table 4 confronts exact and simulated threshold a (Berger *et al.* (1997) proved in the normal case r is always one) We took $b = 100.000$, but since both $m_0(\cdot)$ and $m_1(\cdot)$ are normal distributions, getting all eighth thresholds took less than 10 seconds with a *for* loop; using R, on a 1Ghz Pentium II.

TABLE 4. $\psi(1) = a$ *for the normal test, composite hypotheses.*

$n\xi$	Exact	Simulated
1	1.317	1.319
10	2.321	2.315
15	2.576	2.575
20	2.768	2.758
25	2.922	2.898
30	3.051	3.035
40	3.260	3.233
50	3.425	3.375

The approximation of the simulation is more than satisfying. Note that (marginal) probability of the Bayes factor being in the interval between the exact and simulated threshold will always be negligible.

5. PART II: OPTIMAL SAMPLE SIZE DETERMINATION

In this section, we will show how to choose the sample size for the modified test of Berger *et al.* (1994). We follow a technique proposed by De Santis (2003). For a review of methods, see Adcock (1997) and Weiss (1997).

Along these lines, Verdinelli (1996) and then De Santis (2003) suggest to use the following method: Let k_1 and k_2 be two non negative thresholds, chosen by the statistician, such that: if $B(x) < k_1 \leq 1$, not only we believe H_1 is true, but also that there is strong data evidence in favor of it; if $B(x) > k_2 \geq 1$, not only we believe H_0 is true, but also that there is strong data evidence in favor of it. Let $p_n(k_1, k_2) = \Pr(B(x) < k_1|H_1) * \Pi_1 + \Pr(B(x) > k_2|H_0) * \Pi_0$ be the marginal probability of observing strong evidence in favor of the true hypothesis. The analysis follows the following steps:

1. *Pre-experimental phase.* Fix $\rho \in (0, 1)$, and choose n_ρ such that: $n_\rho = \min\{n/\ p_n(k_1, k_2) \geq \rho\}$;
2. *Experimental phase.* Observe a sample of n_ρ elements.
3. *Post-experimental phase.* Do the test in this way:

$$\begin{cases} B(x) \leq k_1, & \text{reject } H_0 \text{ and report } \alpha^*(B(x)) \\ k_1 < B(x) < k_2, & \text{no decision} \\ B(x) \geq k_2, & \text{retain } H_0 \text{ and report } \beta^*(B(x)), \end{cases}$$

where x is the observed sample.

The obvious problem here is the choice of sensible values for k_1 and k_2. Jeffreys (1961) suggests a possible scale of k_1 and k_2 according to the strength of evidence the researcher wants to see in the results to confirm his decision. He notes, however, that no universal answer can be given with these criteria. A formal approach is taken in De Santis (2003), but goes beyond the purposes of this article.

We now show that one possibility is setting $k_1 = r$ and $k_2 = a$, thus using the modified test and choosing the optimal sample size for it. In practice, we choose a value for n so that, with high probability, the Bayesian test is equivalent to a conditional frequentist test. Apart from overcoming the problem of choosing the thresholds k_1 and k_2, this has some nice properties. We show that using these thresholds implies a symmetry between the two hypotheses. Since we are assuming a 0-1 loss function, this symmetry is sensible from a decision theory point of view. It is also straightforward to see that the modified test in (6) is minimax in the class of tests defined by all varying k_1 and k_2.

5.1. Optimal sample size for the modified test

Let $p_n(a, r)$ be the pre-experimental probability of making the correct decision: $p_n(a, r) = \frac{1}{2}\Pr(D_0(a; n)|H_0) + \frac{1}{2}\Pr(D_1(r; n)|H_1)$, where $D_0(a; n) = \{x/\ B(x) > a\}$ $(D_1(r; n) = \{x/\ B(x) < r\})$, is the subset of the sample space in which H_0 is retained (resp., rejected).

By construction, $\Pr(D_1(r; n)|H_1) = \Pr(D_0(a; n)|H_0) = p_n(a; r)$, and so

$$p_n(a; r) = \begin{cases} F_1(1) & \text{if } \psi(1) \geq 1 \\ 1 - F_0(1) & \text{if } \psi(1) < 1. \end{cases} \tag{15}$$

Note that $p_n(a; r)$ in this form is not explicitly dependent on the values of r and a. This is useful, as in this case r and a change with n. This fact is also the main difference with the method proposed in De Santis (2003): in our case, as the thresholds change with n, we can conclude that the sample size implies the thresholds for testing.

The fact that $\Pr(D_0(a; n)|H_0)$ is equal to $\Pr(D_1(r; n)|H_1)$ $\forall\ n \in \mathbb{N}$ is a key to the minimax optimality stated above, and also implies that there is symmetry between the hypotheses for n fixed: the pre-experimental probability of retaining H_0 when it is true is the same as the probability of rejecting H_0 when it is false.

To summarize, this is the algorithm we are proposing for testing:

1. *Pre-experimental Phase.*
   (a) *Choice of the optimal sample size.* Fix $\rho \in (0, 1)$, and choose n_ρ such that: $n_\rho = \min\{n/\ p_n(a, r) \geq \rho\}$, where $p_n(a, r)$ is defined as in (15);
   (b) *Determination of the thresholds.* Compute a and r correspondent to $n = n_\rho$.
2. *Experimental Phase.* Observe a sample of size n_ρ.
3. *Post-experimental Phase.* Compute the Bayes factor and do the test:

$$\begin{cases} B(x) \leq r, & \text{reject } H_0 \text{ and report } \alpha^*(B(x)) \\ r < B(x) < a, & \text{no decision} \\ B(x) \geq a, & \text{retain } H_0 \text{ and report } \beta^*(B(x)). \end{cases}$$

Note that we expect to end up this algorithm outside the NDR, making the right decision, in $\rho\%$ of our tests.

We think here that it could be sensible use this method of choice of the sample size also when doing a full Bayesian test, i.e., when disregarding the NDR. In this sense, one would choose a sample size so that, with high probability, the Bayesian test would be equivalent to a conditional frequentist one.

We complete the discussion with some examples showing that in usual applications the optimal sample size is reasonably small for sensible values of ρ.

5.2. Optimal Sample Size in Applications

5.2.1 –Normal random sample with known variance

Suppose we are dealing with normal random variables, and that the the alternative is composite. Suppose[4] the prior under the alternative hypothesis is such that $\theta \sim N(\theta_1, \xi\sigma^2)$, with $\theta_1 \in \mathcal{R}$ and $\xi > 0$. Let $\Delta = (\theta_0 - \theta_1)/\sqrt{\xi\sigma^2}$.

Since $\psi(1) > 1$, from (15) we have: $p_n(a; r) = F_1(1)$.

[4] See Berger *et al.* (1997) for a detailed insight on the normal setting.

Let $u_1^{\pm} = \frac{\Delta}{\sqrt{n\xi}} \pm \sqrt{\frac{n\xi+1}{n\xi}(\Delta^2 + \ln(n\xi + 1))}$, then

$$F_1(1) = \Phi\left(\frac{\Delta\sqrt{n\xi+1} - \sqrt{\Delta^2 + \ln(1+n\xi)}}{\sqrt{n\xi}}\right) + \Phi\left(-\frac{\Delta\sqrt{n\xi+1} + \sqrt{\Delta^2 + \ln(1+n\xi)}}{\sqrt{n\xi}}\right). \tag{16}$$

Table 5 shows $n_\rho\xi$ and associated threshold a for some values of ρ and Δ. The optimal sample size n_ρ is just $n_\rho = \left\lceil \frac{n_\rho\xi}{\xi} \right\rceil$ for fixed $\xi \neq 0$. Notice that these sample sizes are all reasonable, as usually $\xi \geq 1$ (in practice, people tend to down weight prior information using "flat" priors, with respect to the density function of one observation, see Kass and Wasserman (1995) for a discussion on this issue).

TABLE 5. *$n_\rho\xi$ and associated a, normal random variables with known σ.*

/	$\Delta = 0$		$\Delta = 1$		$\Delta = 2$	
ρ	$n_\rho\xi$	a	$n_\rho\xi$	a	$n_\rho\xi$	a
0.50	4	1.822	1	1.665	1	1.777
0.75	36	3.183	13	3.152	2	2.306
0.85	138	4.214	51	4.204	4	3.036
0.90	376	5.036	138	5.025	10	4.206
0.95	1924	6.433	708	6.430	40	5.990

As we could expect, n_ρ is inversely proportional to ξ, since as ξ grows the prior is less concentrated and so more and more observations are needed to make a statement on H_0.

Figure 5 shows $p_n(a; r)$ with respect to $n_\rho\xi$, for different values of Δ. As one would expect, as Δ increases the prior distribution is centered farther and farther from θ_0, and so it is easier to discriminate.

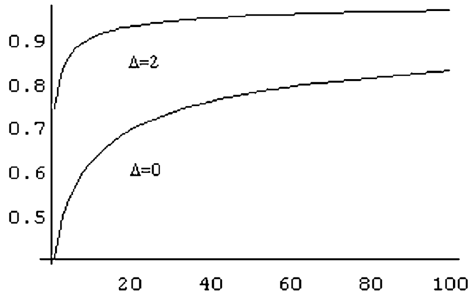

Figure 5. $p_n(a; r)$, normal test.

5.2.2 –Exponential case, simple hypotheses

WLOG we can suppose $\lambda = \frac{\theta_0}{\theta_1} > 1$. Using (15), (8) and (9); we have:

$$p_n(a;r) = \begin{cases} 1 - H(2n\ln(\lambda)/(\lambda-1)) & \text{if } \psi(1) \geq 1 \\ H(2n\lambda\ln(\lambda)/(\lambda-1)) & \text{if } \psi(1) < 1, \end{cases}$$

where $H(\cdot)$ is the CDF of a chi-square random variable with $2n$ d.f.

Table 6 shows n_ρ for some values of λ and ρ. Note that all optimal sample sizes are relatively small, thus allowing the researcher to be parsimonious. Note also that as λ gets farther and farther from the value 1, it is easier and easier to discriminate between the two hypotheses and the optimal sample size is thus smaller.

TABLE 6. *n_p for some λ.*

λ	$\rho = 0.5$	0.75	0.85	0.95
$1,25^{\pm 1}$	4	43	93	222
$1,5^{\pm 1}$	2	15	30	70
$1,75^{\pm 1}$	2	9	17	38
$2^{\pm 1}$	2	6	11	25
$3^{\pm 1}$	1	3	5	11

Figure 6 shows $p_n(a;r)$ with respect to n.

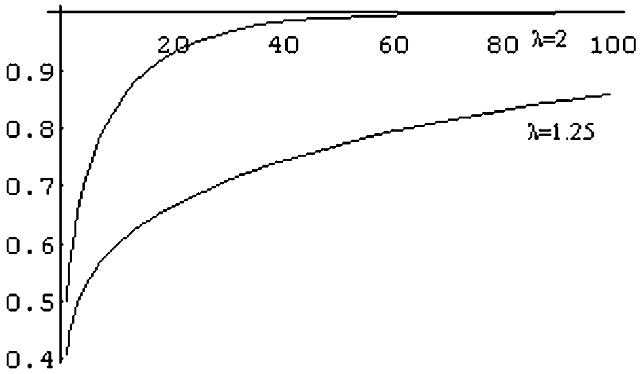

Figure 6. $p_n(a;r)$ for the exponential test, simple hypothese.

5.2.3 –Exponential case, composite alternative

Using (15), it can be shown that:

$$p_n(a;r) = \begin{cases} GG(s_{1,1}) + 1 - GG(s_{1,2}) & \text{if } \psi(1) \geq 1 \\ H(2s_{1,2}\theta_0) - H(2s_{1,1}\theta_0) & \text{if } \psi(1) < 1, \end{cases}$$

where $GG(\cdot)$ is the CDF of a Gamma-Gamma random variable with parameters (δ, λ, n) and $s_{1,i}$, $i = 1, 2$, are the solutions of the equation $B(s) = 1$. Table 7 shows optimal n_ρ for some δ, with $\theta_0 = 1$ and $\lambda = \delta$. The sample sizes needed to have reasonable probabilities of strong and correct evidence are relatively small and of course increase with ρ..

TABLE 7. *n_ρ for some δ, with $\theta_0 = 1$ and $\lambda = \delta\theta_0$.*

δ	$\rho = 0.5$	0.75	0.85
1	4	31	120
2/3	3	19	65
1/2	1	14	51

5.3. Unequal Prior Probabilities of the Two Hypotheses

We will now extend the method of choice of the optimal sample size to the generalized unified test, i.e., to the case in which the prior probabilities of the two hypotheses are not equal. Remember that we call $\eta = \frac{p_{H_1}}{p_{H_0}}$, the ratio of the two prior probabilities.

Since

$$F_0(a_\eta) = \max(F_0(\eta), 1 - \eta F_1(\eta))$$
$$F_1(r_\eta) = \min(F_1(\eta), \frac{1}{\eta}[1 - F_0(\eta)]),$$

we have that

$$Pr(D_0(a_\eta; n)|H_0) = \begin{cases} 1 - F_0(\eta) & \text{if } F_0(\eta) > 1 - \eta F_1(\eta) \\ \eta F_1(\eta) & \text{if } F_0(\eta) \leq 1 - \eta F_1(\eta) \end{cases}$$

and

$$Pr(D_1(r_\eta; n)|H_1) = \begin{cases} 1/\eta[1 - F_0(\eta)] & \text{if } F_0(\eta) > 1 - \eta F_1(\eta) \\ F_1(\eta) & \text{if } F_0(\eta) \leq 1 - \eta F_1(\eta) \end{cases}$$

The $\Pr(D_0(a_\eta; n)|H_0)$ is always η times $\Pr(D_1(r_\eta; n)|H_1)$. Hence we have the equality

$$p_{H_0} \Pr(D_0(a_\eta; n)|H_0) = p_{H_1} \Pr(D_1(r_\eta; n)|H_1).$$

We no longer have the equality between the probabilities of correct rejection, but we now have equality between the joint probabilities of an hypothesis being true and the rejection of the other, after the experiment.

It is straightforward to see that

$$p_n(r_\eta; a_\eta) = \begin{cases} 2p_{H_1} F_1(\eta) & \text{if } F_0(\eta) \leq 1 - \eta F_1(\eta) \\ 2p_{H_0}[1 - F_0(\eta)] & \text{if } F_0(\eta) > 1 - \eta F_1(\eta) \end{cases} \qquad (17)$$

Once again, the probability in (17) does not explicitly depend on r_η or a_η, but only on n.

What we said about the case in which η is equal to 1, easily generalizes to $\eta \neq 1$. There is only an important feature to comment on: the probability $p_n(r_\eta; a_\eta)$ has an horizontal asymptote at level $2\min(p_{H_1}, p_{H_0})$, and, when the two probabilities are not equal, it is strictly smaller than 1. Hence, in this case, the probability of falling in the no decision region does not go to 0 as n grows. Note that the maximum probability of correct decision that can be attained can be very low. For instance, if $p_{H_1} = 0.8$, there is no sample size for which $p_n(r_\eta; a_\eta) \geq 0.4$, a very small threshold.

Example 2. Consider the data used in Lee and Zelen (2000) regarding clinical trials from the Eastern Cooperative Oncology Group (ECOG), on various types of cancer. The data is very general and the observed values x are a standardized measure of the difference between a control group and a group who took a new treatment. The parameter θ measures the general efficiency of the treatment. Lee and Zelen (2000) assume x is normal with mean θ and variance σ^2.

We are testing a null hypothesis that $\theta = 0$ against a double sided alternative.

Lee and Zelen (2000) determine prior information on the basis of 87 clinical trials, and estimate that $0.28 < p_{H_1} < 0.32$. For the time being, we will assume $p_{H_1} = 0.3$, and thus $\eta = 0.429$. The upper bound for the probability of correct decision is 0.6. A sensible strategy is to get as close as possible to this upper bound, with a reasonable sample size.

Lee and Zelen (2000) assume the prior to be normal, with mean 0. We suggest to take ξ, the ratio between prior and sample variance, to 1. Following Kass and Wasserman (1995), this implies that we give to the prior information the same weight that we give to one observation. Recall that with normal distributions, $\psi(1) > 1$, so we have $p_n(1; a_{0.429,1}) = 0.6F_1(0.429)$.

Table 8 shows the optimal sample sizes for different levels ρ. Note that the procedure is not robust with respect to the choice of p_{H_1}. This is particularly evident for large values of ρ, as it is easily seen in figure 7. Note that we need few observations to get to a level $\rho = 0.5$, while when $p_{H_1} = 0.3$, for instance, we need $1263 - 761 = 502$ more observations to have an increase of 1% in the probability of correct decision, from 0.55 to 0.56.

TABLE 8. *n_ρ, various ρ.*

ρ	$p_{H_1} = 0.32$	$p_{H_1} = 0.30$	$p_{H_1} = 0.28$
0.30	7	9	12
0.40	19	28	43
0.50	76	152	440
0.55	220	761	23895
0.56	291	1263	$+\infty$
0.60	1427	$+\infty$	

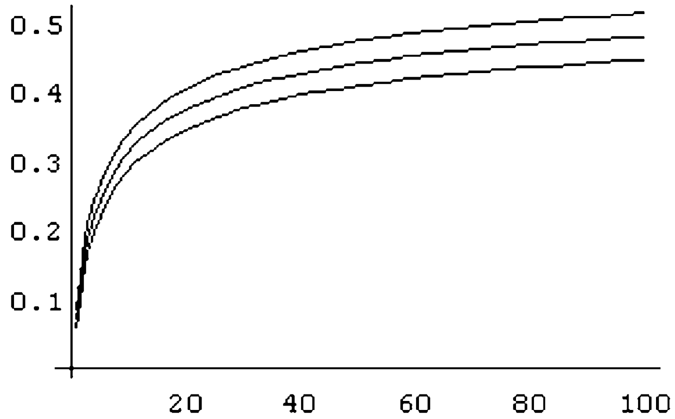

Figure 7. $p_n(a_{1,0.3/0.7}; 1)$, various p_{H_1}

5.4. An alternative procedure for the choice of sample size

Dass and Berger (1999) propose to fix n such that

$$\begin{cases} \Pr(\beta^*(B(x)) < \alpha_0 | H_0) \geq p_0 \\ \Pr(\alpha^*(B(x)) < \alpha_1 | H_1) \geq p_1, \end{cases}$$

for some α_0, α_1, p_0 and p_1. The difference with the previous method is that Π_0 and Π_1 don't come into the analysis, and that the two hypotheses can have very different weight. In fact, with probability greater than or equal to p_0 we expect to retain H_0 with error probability $\beta^*(B(x)) < \alpha_0$ when H_0 is true; and to reject H_0 with error probability $\alpha^*(B(x)) < \alpha_1$ $p_1\%$ of the times when H_0 is false. It is then possible to choose one between p_0 or p_1 so that the behavior of the Bayes factor under the corresponding hypothesis is not contributing to the determination of the sample size. To avoid this, Dass and Berger (1999) recommend to fix $\alpha_1 = \alpha_0$ and $p_0 = p_1$.

Note that this procedure is applicable to the modified test only if the samples in the NDR are not admissible to the system (18), i.e., they are not such that $\beta^*(B(x)) < \alpha_0$ or $\alpha^*(B(x)) < \alpha_1$. To clarify, supposing $\psi(1) > 1$, it makes no sense to have $1/(a+1) < \beta^*(B(x)) < \alpha_0$. This in fact maps to $\alpha_0^{-1} - 1 < B(x) < a$, which means we are allowing to end up in the NDR.

It is straightforward then to see that the additional constraints $\alpha_0 \leq 1/(a+1)$ and $\alpha_1 \geq r/(r+1)$ are needed. If $\alpha_0 = 1/(a+1)$ and $\alpha_1 = r/(r+1)$, the optimal sample size is the same as in the previous subsection.

This approach suggests how to overcome a possible drawback of the procedure proposed in the previous subsection. In fact, one between a and r

is always equal to 1. The idea is that k_1 and k_2 are bounds for "strong", or at least "moderate", evidence expressed by the Bayes factor against one of the two hypotheses. In this sense, it may be sensible to request that k_1 and k_2 must be different than 1. A possible solution is to fix $\alpha_0 = \alpha_1 = \min(r/(r+1), 1/(1+a))$. The procedure will be more conservative (the optimal sample sizes will be sensibly higher), but, with high probability, we will end up outside the NDR and with strong (moderate) evidence in favor of the true hypothesis. With the setting of the previous section, this is equivalent to fixing $D_0(\max(a, 1/r); n)$ and $D_1(\min(r, 1/a); n)$. It is straightforward to see that $p_n(\max(a, 1/r); \min(r, 1/a))$ is equal to:

$$p_n(\max(a, 1/r); \min(r, 1/a)) = \begin{cases} \frac{F_1(1/a)+F_1(1)}{2} & \text{if } \psi(1) \geq 1 \\ 1 - \frac{F_0(1/r)+F_0(1)}{2} & \text{if } \psi(1) < 1 \end{cases},$$

so the optimal sample size is $n = \arg\min(n \in \mathbb{N}/ \; p_n(\max(a, 1/r); \min(r, 1/a)) > \alpha)$.

It is immediate to see that $p_n(\max(a, 1/r); \min(r, 1/a)) < p_n(a; r)$, $\forall\, n \in \mathbb{N}$, so bigger values of the optimal sample size will be given.

Example 3. [Fisher and Van Belle, (1993)] The data is taken from Fisher and Van Belle (1993), and is about the weight at birth of $n = 15$ babies born in King County in 1997, and dead by SIDS (sudden death syndrome). Suppose the weight at birth is distributed like a normal with $\sigma = 800g$. The average weight of all babies born in King County in 1977 was $3300g$. Hence, we are testing to see if $H_0 : \theta = 3300$. Berger *et al.* (1997) suggest to pick $\Delta = 0$ and $\xi = 2$. They explain this choice by the fact that the prior $N(0, 2\sigma^2)$ is approximately a $Cauchy(0, \sigma^2)$, which is the reference prior suggested by Jeffreys (1961). From Table 4 we see that $a = 3.051$. Since the sample average is $\bar{x} = 3199.8$, and so $u = (3199.8 - 3300)\sqrt{15}/800 = -0.485$; and $B(-0.485) = 4.968 > 3.051$, we retain the null hypothesis and conclude that there is no relationship between sudden death and weight at birth. The error probability is $\beta^*(B(x)) = 0.168$.

We have $p_{15}(1, 3.051) = 0.7351$. To have a probability of strong and correct evidence higher than 0.75, we would have needed $36/2 = 18$ observations, only 3 more. For a level of 0.85 the optimal sample size is 69, with $a = 4.214$.

The alternative approach yields: $p_{15}(3.051; 3.051^{-1}) = 0.6882$, sensibly lower. For a level of 0.75 you need $n = 30$, since $p_{30}(3.563; 3.563^{-1}) = 0.754$.

In conclusion, the modified test proposed by Berger *et al.* (1994) can be easily used if computing the bounds of the NDR via simulation. Reporting the thresholds together with the Bayes factor can say whether the procedure is equivalent to the application of frequentist methods. On the other side, this method can suggest the optimal sample size to choose for testing.

Appendix

Proof of Theorem 1. Let $1/\lambda = \nu$, and assume $\lambda > 1$. We have that

$$2\lambda \ln\left(\frac{\lambda^n}{b}\right)/(\lambda - 1) = 2\ln(\nu^n b)/(\nu - 1), \tag{18}$$

Since,

$$\begin{aligned} 2\lambda \ln\left(\frac{\lambda^n}{b}\right)/(\lambda - 1) &= 2\ln(1/\nu^n b)/(1 - \nu) \\ &= 2\ln(\nu^n b)/(\nu - 1) \end{aligned}$$

In the same way it is possible to prove that

$$2\ln\left(\frac{\lambda^n}{b}\right)/(\lambda - 1) = 2\nu\ln(\nu^n b)/(\nu - 1) \tag{19}$$

Hence,

$$\begin{aligned} \psi(1,\lambda) &= \lambda^n \exp\left\{-\frac{(\lambda-1)H^{-1}\left(1 - H\left(\frac{2n\ln\lambda}{\lambda-1}\right)\right)}{2\lambda}\right\} = \\ &= \nu^{-n}\exp\left\{(\nu - 1)H^{-1}\left(1 - H\left(\frac{2n\nu\ln\nu}{\nu-1}\right)\right)/2\right\} = \\ &= \frac{1}{\psi^{-1}(1,\nu)}. \end{aligned}$$

Moreover, applying (18) again,

$$F_0(1,\lambda) = 1 - H(2\lambda\ln(\lambda^n)/(\lambda - 1))1 - H(2\ln(\nu^n)/(\nu - 1)) = 1 - F_1(1,\nu).$$

Acknowledgments

The author is grateful to Prof. Fulvio De Santis for the great help and support and to Prof. Ludovico Piccinato for the suggestions for improvements.

References

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

Berger J.O., Boukai, B., and Wang, Y. (1997) Unified frequentist and Bayesian testing of a precise hypotesis, *Statistical Science*, 12 (3), 133–160.

Berger, J.O., Boukai, B. and Wang, Y. (1998) Simultaneous Bayesian-Frequentist sequential testing of nested hypothesis, *Biometrika*, 79–92.

Berger, J.O., Brown, L.D., and Wolpert, R.L. (1994) A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing, *Annals of Statistics*, 22 (4), 1787–1807.

Berger, J.O. and Delampady, M. (1987) Testing Precise Hypotheses, *Statist. Sci.*, 3, 317–352.

Bernardo, J.M. and Smith, A.F.M. (1994) *Bayesian theory*, Wiley and Sons, Applied probability and statistics.

Brownie, C. and Kiefer, J. (1977) The ideas of conditional confidence in the simplest setting, *Comm. Statist. Theory Methods*, 6, 691–751.

Dass, S.C. and Berger, J.O. (1999) *Unified Bayesian and Conditional Frequentist Testing of Composite Hypotheses*, University of Michigan.

De Santis, F. (2003) Statistical Evidence and Sample Size Determination for Bayesian Hypothesis testing, *Journal of Statistical Planning and Inference*, to Appear.

Fisher, L.D. and Van Belle, G. (1993) *Biostatistics: A Methodology for the Health Sciences*, Wiley, New York.

Frosini, B. V. (1999) Conditioning, Information and frequentist properties, *Statistica Applicata*, 11, 165–184.

Jeffreys, H. (1961) *Theory of probability*, Oxford University Press, New York.

Kass R.E. and Raftery A. E. (1995) Bayes Factors, *J.A.S.A.*, 90, 773–795.

Kass, R.E. and Wasserman, L. (1995) A reference Bayesian test for nested hypotheses and its relationship to the Schwartz criterion, *J.A.S.A.*, 90, 928–939.

Kiefer, J. (1977) Conditional confidence statements and confidence estimators (with discussion), *J.A.S.A.*, 72, 789–827.

Lee, S.J. and Zelen, M. (2000) Clinical trials and sample size considerations: another perspective, *Statistical science*, 15(2), 95–110.

Piccinato, L. (1996) *Metodi per le decisioni statistiche*, Springer, Collana di statistica.

Royall, R. (1997) *Statistical evidence: a likelihood paradigm*, Chapman and Hall, Monographs on Statistics and Applied Probability, 71.

Sellke, T. and Bayarri, M.J., and Berger, J.O. (2001) Calibration of P-values for testing precise null hypotheses, *The American Statistician*, 55, 62–71.

Verdinelli, I. (1996) *Bayesian designs of experiments for the linear model*, Dept. of Statistics, Carnegie Mellon University.

Weiss, R. (1997) Bayesian sample size calculations for hypotesis testing, *The Statistician*, 46, 185–191.

ALESSIO FARCOMENI
Dipartimento di Statistica
Probabilità e Statistiche Applicate
Università degli Studi "La Sapienza"
Piazzale Aldo Moro, 5
00185 Roma (Italia)
alessio.farcomeni@uniroma1.it