UNIFIED FREQUENTIST AND BAYESIAN TESTING OF A PRECIST HYPOTHESIS

by

J.O. Berger
Purdue University

and

B. Boukai
Indiana University-Purdue University

Y. Wang
Indiana University-Purdue University

Technical Report #94-25C

October 1994
Revised February 1996

*

UNIFIED FREQUENTIST AND BAYESIAN TESTING OF A PRECISE HYPOTHESIS

J. O. BERGER[1], B. BOUKAI[2] AND Y. WANG[2]

[1] Department of Statistics
Purdue University
West Lafayette

[2] Department of Mathematical Sciences
Indiana University-Purdue University
Indianapolis

January 1996

ABSTRACT. In this paper, we show that the Conditional Frequentist method of testing a precise hypothesis can be made virtually equivalent to Bayesian testing. The conditioning strategy proposed by Berger, Brown and Wolpert (1994), for the simple versus simple case, is generalized to testing a precise null hypothesis versus a composite alternative hypothesis. Using this strategy, both the Conditional Frequentist and the Bayesian will report the same error probabilities upon rejecting or accepting. This is of considerable interest because it is often perceived that Bayesian and frequentist testing are incompatible in this situation. That they are compatible, when conditional frequentist testing is allowed, is a strong indication that the "wrong" frequentist tests are currently being used. The new unified testing procedure is discussed and illustrated in several common testing situations.

1980 *Mathematics Subject Classification* (1985 *Revision*). Primary 62A20; Secondary 62A15.

Key words and phrases. Bayes factor, likelihood ratio, composite hypothesis, conditional test, error probabilities.

Research supported in part by the National Science Foundation under grant DMS-9303556.

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TeX

1. Introduction

The problem of testing statistical hypotheses has been one of the focal points for disagreement between Bayesians and Frequentists. The classical Frequentist approach constructs a *rejection* region and reports associated error probabilities. Incorrect rejection of the null hypothesis H_0, the *Type I error*, has probability α, and incorrect acceptance of H_0, the *Type II error*, has probability β. This traditional (α, β)-Frequentist approach has been criticized for reporting error probabilities that are independent of the given data. Thus a common alternative is to use the P-value as a data-dependent measure of the strength of evidence against the null hypothesis H_0. But the P-value is not a true frequentist measure and has its own shortcomings as a measure of evidence. Edwards, Lindman, and Savage (1963), Berger and Sellke (1987), Berger and Delampady (1987) and Delampady and Berger (1990) have reviewed the practicality of the P-value and explored the dramatic conflict between the P-value and other data-dependent measures of evidence. Indeed, they demonstrate that the P-value can be highly misleading as a measure of the evidence provided by the data against the null hypothesis. Because this point is of central importance in motivating the need for the development here, we digress with an illustration of the problem.

Illustration 1: Suppose that one faces a long series of exploratory tests of possible new drugs for AIDS. We presume that some percentage of this series of drugs are essentially ineffective. (Below, we will imagine this percentage to be 50%, but the same point could be made with any given percentage.) Each drug is tested in an independent experiment, corresponding to a test of no treatment effect based on normal data. For each drug, the P-value is computed, and those with P-values smaller than 0.05 are deemed to be effective. (This is perhaps an unfair caricature of standard practice, but that is not relevant to the point we are trying to make about P-values.)

Suppose a doctor reads the results of the published studies, but feels confused about the meaning of P-values. (Let us even assume here that all studies are published, whether they obtain statistical significance or not; the real situation of publication selection bias only worsens the situation.) So, in hopes of achieving a better understanding, the doctor asks the resident statistician to answer a simple question: "A number of these published studies have P-values that are between 0.04 and 0.05; of those, what fraction of the corresponding drugs are ineffective?"

The statistician cannot provide a firm answer to this question, but can provide useful bounds if the doctor is willing to postulate a prior opinion that a certain percentage of the drugs being originally tested (say, 50%, as mentioned above) were ineffective. In particular, it is then the case that at least 23% of the drugs having P-values between 0.04 and 0.05 are ineffective, and in practice typically 50% or more will be ineffective (see Berger and Sellke, 1987). Relating to this last number, the doctor concludes: "So if I start out believing that a certain percentage of the drugs will be ineffective, then a P-value near 0.05 does not change my opinion much at all; I should still think that about the same percentage of those with a P-value near 0.05 are ineffective." That is an essentially correct interpretation.

We cast this discussion in a frequentist framework to emphasize that this is a fundamental fact about P-values; in situations such as that here, involving testing a precise null hypothesis, a P-value of 0.05 essentially does not provide any evidence against the null hypothesis. Note, however, that the situation is quite different in situations where there is not a precise null hypothesis; then it will often be the case that only about 1 out of 20 of the drugs with a P-value of 0.05 will be ineffective, assuming that the initial percentage of ineffective drugs is again 50% (cf., Casella and Berger, 1987). In a sense though, this only acerbates the problem; it implies that the interpretation of P-values will change drastically from problem to problem, making them highly questionable as a useful tool for statistical communication.

To rectify these deficiencies, there have been many attempts to modify the classical Frequentist approach by incorporating data-dependent procedures which are based on conditioning. Earlier works in this direction are summarized in Kiefer (1977) and in Berger and Wolpert (1988). In a seminal series of papers, Kiefer (1975, 1976, 1977) and Brownie and Kiefer (1977), the Conditional Frequentist approach was formalized. The basic idea behind this approach is to condition on a statistic measuring the evidential strength of the data, and then to provide error probabilities conditional on the observed value of this statistic. Unfortunately, the approach never achieved substantial popularity, in part because of the difficulty of choosing the statistic upon which to condition (cf., the Discussion of Kiefer, 1977).

A prominent alternative approach to testing is the Bayesian approach, which is based on the most extreme form of conditioning, namely conditioning on the given data. There have been many attempts (see, for example, Good, 1992) to suggest compromises between the Bayesian and the Frequentist approaches. However, these compromises have not been adopted by practitioners of statistical analysis, perhaps because they lacked a complete justification from either perspective.

Recently, Berger, Brown and Wolpert (1994) - henceforth, BBW - considered the testing of simple versus simple hypotheses and showed that the Conditional Frequentist method can be made exactly equivalent to the Bayesian method. This was done by finding a conditioning statistic which allows an agreement between the two approaches. The surprising aspect of this result is not that both the Bayesian and the Conditional Frequentist might have the same decision rule for rejecting or accepting the null hypothesis (this is not so uncommon), but rather that they will report the same (conditional) error probabilities upon rejecting or accepting. That is, the error probabilities reported by the Conditional Frequentist using the proposed conditioning strategy are the same as the posterior probabilities of the relevant errors reported by the Bayesian.

The appeal of such a testing procedure is evident. The proposed test and the suggested conditioning strategy do not comprise a compromise between the Bayesian and the Frequentist approaches, but rather indicate that there is a way of testing that is simultaneously frequentist and Bayesian. The advantages of this "unification" include the following:

(i) Data-dependent error probabilities are utilized, overcoming the chief objection to (α, β) – Frequentist testing. And these are true error probabilities, and hence do not suffer the type of misinterpretation that can arise with P-values.

(ii) Many statisticians are comfortable with a procedure only when it has simultaneous Bayesian and frequentist justifications. The testing procedure we propose, for testing a simple null hypothesis versus a composite alternative, is the first we know of that possesses this simultaneous interpretation (for this problem).

(iii) A severe pedagogical problem is the common misinterpretation among practitioners of frequentist error probabilities as posterior probabilities of hypotheses. By using a procedure for which the two are equivalent, this concern is obviated.

(iv) Since the approach is Bayesianly justifiable, one can take advantage of numerous Bayesian simplifications. For instance, the stopping rule (in, say, a clinical trial) does not affect the reported error probabilities; hence one does not need to embark upon the difficult (and controversial) path of judging how to "spend α" for "looks at the data." (A full discussion of sequential aspects of the procedure would be too lengthy. See BBW for discussion in the simple versus simple case; we will report on the sequential situation for composite hypotheses in a later paper.)

Most "Bayesian-Frequentist agreement" articles end up arguing that the "classical" procedures being used today are satisfactory from either viewpoint. It is noteworthy that this is not the case here. In effect, we argue that the Bayesian procedure is correct, in part because it has a very sensible conditional frequentist interpretation; but this procedure is *very* different than what is typically used in practice. Hence we are proposing a serious change in practical statistical methodology.

The general development given later may appear to be somewhat involved technically, but the new tests that result are often quite simple. To illustrate this, as well as some of the comparison issues mentioned above, we end the introduction with an example.

Example 1.

Suppose that $X_1, X_2, \ldots, X_n$ are n i.i.d. random variables from a normal distribution having unknown mean θ and known variance σ^2, (i.e. the $\mathcal{N}(\theta, \sigma^2)$ distribution) and denote by $\bar{X}_n = \sum X_i/n$ their average; thus $\bar{X}_n \sim \mathcal{N}(\theta, \sigma^2/n)$. Based on the observed value $\bar{x}_n$ of $\bar{X}_n$, we are interested in testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Consider the following three testing procedures, defined in terms of the standard statistic $t = \sqrt{n}(\bar{x}_n - \theta_0)/\sigma$.

Classical Frequentist Test:

$$T_C : \begin{cases} \text{if } |t| \geq t_{\alpha/2}, & \text{reject } H_0 \text{ and report error probability } \alpha, \\ \text{if } |t| < t_{\alpha/2}, & \text{accept } H_0 \text{ and report error probability } \beta(\theta), \end{cases}$$

where α and $\beta(\theta)$ are the probabilities of Type I and Type II error and $t_{\alpha/2}$ is the usual critical value. Since $\beta(\theta)$ depends on the unknown θ, it is common to choose a "subjectively important" value of θ (or two) and report β at that (or those) points.

P-value Test:

$$T_P : \begin{cases} \text{if } |t| \geq t_{\alpha/2}, & \text{reject } H_0 \text{ and report the } P\text{-value } p = 2(1 - \Phi(|t|), \\ \text{if } |t| < t_{\alpha/2}, & \text{do not reject } H_0 \text{ and report } p; \end{cases}$$

here, and in the sequel, Φ denotes the standard normal c.d.f. whose p.d.f. is denoted by ϕ. Typically in such a test, $\alpha = 0.05$ is chosen.

A New Conditional Test:

$$T_1^* : \begin{cases} \text{if } B(t) \leq 1, & \text{reject } H_0 \text{ and report error probability } \alpha^* = B(t)/(1 + B(t)), \\ \text{if } 1 < B(t) < a, & \text{make no decision,} \\ \text{if } B(t) \geq a, & \text{accept } H_0 \text{ and report error probability } \beta^* = 1/(1 + B(t)), \end{cases}$$

where $B(t) = \sqrt{1+2n}\ \exp\{-t^2/(2+n^{-1})\}$ and a is a constant defined in (4.7); a good approximation to a is $a \cong \log(5n) - \log\log(1+2n)$. As we will see later, α^* and β^* have a dual interpretation as (i) (conditional) frequentist Type I and Type II error probabilities and (ii) the posterior probabilities of H_0 and H_1, respectively.

To see these three tests in action, suppose $n = 20$, $\theta_0 = 0, \sigma^2 = 1, \alpha = 0.05$ for T_C and T_P, and $\theta = 1$ is deemed to be of interest for Type II error under T_C. Table 1 summarizes the conclusions from each test for various values of t. Note that $t_{\alpha/2} = 1.96$ and $a = 3.26$.

	Values of $\lvert t\rvert$	T_C	T_P	T_1^*
	0	($\beta(1) = 0.006$)	$p = 1$	$\beta^* = 0.135$
acceptance region	1		$p = 0.317$	$\beta^* = 0.203$
	1.18			No Decision Region
	1.96	($\alpha = 0.05$)	$p = 0.05$	$\alpha^* = 0.496$
rejection region	3		$p = 0.0026$	$\alpha^* = 0.074$
	4		$p = 0.0000$	$\alpha^* = 0.0026$

Table 1. Conclusions from the classical, P-value, and conditional tests when $n = 20$ and $\alpha = 0.05$.

The acceptance and rejection regions of all three tests are the same, except that T_1^* makes no decision when $1.18 < |t| < 1.96$. (This agreement is a convenient coincidence for this illustration, but will not happen in general.) The differences between the tests, therefore, are in the "error probabilities" that are reported.

Compare, first, T_C and T_1^*. The error probabilities for T_C are fixed, while those for T_1^* vary with $|t|$. In the rejection region, for instance, T_C always reports $\alpha = 0.05$, while T_1^* reports error probabilities ranging from nearly 1/2 (when $|t| = 1.96$) to $\alpha^* = 0.0026$ (when $|t| = 4$). The variability in the reports for T_1^* is clearly appealing.

Compare, next, T_P and T_1^*. An immediate advantage of T_1^* is that it can "accept" H_0, with specified error probability, while the P-value (or $1-p$) is in no sense an error probability for acceptance (see the articles mentioned at the beginning of the Introduction for discussion). In the rejection region, p does vary with $|t|$, but it is smaller than α^* by a factor of at least 10. Since we will argue that α^* is a sensible conditional error probability, this discrepancy provides further evidence that P-values can be highly misleading (if interpreted as error probabilities). Indeed, in the situation of *Illustration 1*, note that $\alpha^* = 0.496$ (for those drugs where the P-value is 0.05), which would correctly reflect the fact that, typically, about 50% of these drugs will still be ineffective.

A comment is in order about the "no-decision" region in T_1^*. In practice the no-decision region is typically innocuous, corresponding to a region in which virtually no statistician would feel that the evidence is strong enough for a conclusive decision. The no-decision region could be eliminated, but at the expense of introducing some counter-intuitive properties of the test. Indeed, when this is more fully discussed in section 2.4, it will be observed that even unconditional frequentists should probably introduce a no-decision region to avoid paradoxical behavior.

2. Notation and the "simple" hypotheses case

2.1 The Frequentist and Conditional Frequentist approaches

Suppose we observe the realization x of the random variable $X \in \mathcal{X}$ and wish to test the following *"simple"* hypotheses:

$$H_0 : X \sim m_0(x) \text{ versus } H_1 : X \sim m_1(x), \tag{2.1}$$

where m_0 and m_1 are two specified probability density functions (p.d.f.). We denote by

$$B(x) = \frac{m_0(x)}{m_1(x)}. \tag{2.2}$$

the *likelihood ratio of* H_0 *to* H_1 (or equivalently the *Bayes factor in favor of* H_0). Let F_0 and F_1 be the c.d.f.s of $B(X)$ under H_0 and H_1, respectively (under m_0 and m_1, respectively). For simplicity, we assume in the following that their inverses F_0^{-1} and F_1^{-1} exist. The decision to either *reject* or *accept* H_0 will depend on the observed value of $B(x)$, where *small* values of $B(x)$ correspond to rejection of H_0.

For the traditional Frequentist the classical most powerful test of the simple hypotheses (2.4) is determined by some *critical value* c such that,

$$\begin{cases} \textit{if } B(x) \le c, & \textit{reject } H_0 \\ \textit{if } B(x) > c, & \textit{accept } H_0 \text{ .} \end{cases} \tag{2.3}$$

Corresponding to the test (2.3), the Frequentist reports the Type I and Type II errors probabilities as $\alpha = P_0(B(X) \le c) \equiv F_0(c)$ and $\beta = P_1(B(X) > c) \equiv$

$1 - F_1(c)$. For the standard equal-tailed test with $\alpha = \beta$, the critical value c satisfies $F_0(c) = 1 - F_1(c)$.

The Conditional Frequentist approach allows the reporting of data–dependent error probabilities. In this approach, one considers some statistic $S(X)$, where larger values of $S(X)$ indicate data with greater evidentiary strength (for, or against, H_0), and then reports error probabilities conditional on $S(X) = s$, where s denotes the observed value of $S(X)$. For the test (2.3), the resulting conditional error probabilities are given by

$$\begin{aligned} \alpha(s) &= Pr(\textit{Type I error } |S(X) = s) \equiv P_0(B(X) \leq c|S(X) = s) \\ \beta(s) &= Pr(\textit{Type II error } |S(X) = s) \equiv P_1(B(X) > c|S(X) = s). \end{aligned} \tag{2.4}$$

Thus, for the Conditional Frequentist, the test (2.3) of these simple hypotheses becomes

$$\begin{cases} \textit{if } B(x) \leq c, \textit{ reject } H_0 \textit{ and report conditional error probability } \alpha(s) \\ \textit{if } B(x) > c, \textit{ accept } H_0 \textit{ and report conditional error probability } \beta(s). \end{cases} \tag{2.5}$$

Of course, one is always free to report both $\alpha(s)$ and $\beta(s)$, and indeed the entire functions $\alpha(\cdot)$ and $\beta(\cdot)$, if desired.

Example 2.

Suppose $X > 0$ and we wish to test

$$H_0 : X \sim e^{-x} \text{ versus } H_1 : X \sim \frac{1}{2}e^{-x/2}.$$

Then $B(x) = 2e^{-x/2}$. If we choose $c = 1$ in (2.3), the error probabilities of this unconditional test are $\alpha = 0.25$ and $\beta = 0.5$.

An interesting statistic for formation of a conditional test is $S(X) = |B(X) - 1|$. Clearly S is between 0 and 1, and larger values of S correspond to data providing greater evidence for, or against, H_0. Furthermore, $S(X)$ is an ancillary statistic, having a uniform distribution on (0,1) under either hypothesis. (Conditioning on ancillary statistics is, of course, quite common.)

Computing the conditional Type I and Type II errors in (2.4) is easy because $\{X : S(X) = s\}$ is just a two point set. Calculation then yields, as the Conditional Frequentist test (2.5),

$$\begin{cases} \text{if } B(x) \leq 1, & \textit{reject } H_0 \textit{ and report conditional error probability } \alpha(s) = B(x)/2 \\ \text{if } B(x) > 1, & \textit{accept } H_0 \textit{ and report conditional error probability } \beta(s) = 0.5. \end{cases} \tag{2.6}$$

It is interesting that only the conditional Type I error varies with the data.

It is actually quite rare for there to be suitable ancillary statistics upon which to condition, as in Example 2. (For some other situations in which this occurs, see BBW.) Hence we will employ a different (and more Bayesian) strategy for determining a suitable conditioning statistic. We return to the issue of ancillarity in section 5.

2.2 The Bayesian approach

In Bayesian testing of the above hypotheses, one usually specifies the *prior probabilities*, π_0 for H_0 being true and $1-\pi_0$ for H_1 being true. Then the posterior probability (given the data) of H_0 being true is

$$Pr(H_0|x) = \left[1 + \frac{(1-\pi_0)}{\pi_0}\frac{1}{B(x)}\right]^{-1}. \tag{2.7}$$

To a Bayesian, $B(x)$ in (2.2) is the Bayes factor in favor of H_0, which is often viewed as the odds of H_0 to H_1 arising from the data; $\pi_0/(1-\pi_0)$ is the prior odds. Clearly, small observed values of $B(X)$ suggest rejection of H_0.

When no specific prior probabilities of the hypotheses are available, it is intuitively appealing to choose $\pi_0 = \frac{1}{2}$ in (2.7). We will use this default choice in the remainder of the paper (although generalizations to other π_0 are possible, following the approach in BBW). With this default prior probability, the posterior probability in (2.7) becomes

$$\alpha^*(B(x)) \equiv Pr(H_0|x) = \frac{B(x)}{1+B(x)} \tag{2.8}$$

and the posterior probability that H_1 is true is

$$\beta^*(B(x)) \equiv Pr(H_1|x) = \frac{1}{1+B(x)}. \tag{2.9}$$

The standard Bayesian test for this situation can then be written as

$$\mathbf{T}_1 : \begin{cases} \textit{if } B(x) \leq 1, \textit{ reject } H_0 \textit{ and report the posterior probability } \alpha^*(B(x)) \\ \textit{if } B(x) > 1, \textit{ accept } H_0 \textit{ and report the posterior probability } \beta^*(B(x)). \end{cases}$$

(This is, indeed, the optimal Bayesian test if "0−1" loss is used; again, other losses could be considered, following the lines of BBW.)

2.3 The modified Bayesian test.

The formal similarities between the conditional Frequentist test (2.5) and the test $\mathbf{T}_1$ are quite pronounced. In fact, BBW have shown that $\mathbf{T}_1$ can be given a meaningful conditional Frequentist interpretation, when testing simple versus simple hypotheses. They modified the test $\mathbf{T}_1$ to include a *no decision region* and suggested a conditioning strategy under which the conditional Frequentist test will agree with this modified Bayesian test.

For any $b > 0$, let $\psi(b) = F_0^{-1}(1 - F_1(b))$ with $\psi^{-1}(b) \equiv F_1^{-1}(1 - F_0(b))$ and define

$$\begin{array}{lll} r = 1 \quad \text{and} \quad a = \psi(1) & \text{if} \ \psi(1) \geq 1 \\ r = \psi^{-1}(1) \quad \text{and} \quad a = 1 & \text{if} \ \psi(1) < 1. \end{array} \tag{2.10}$$

Consider the test of H_0 versus H_1 given by

$$\mathbf{T}_1^* : \begin{cases} \textit{if } B(x) \leq r, \textit{ reject } H_0 \textit{ and report the conditional error probability } \alpha^*(B(x)); \\ \textit{if } r < B(x) < a, \textit{ make no decision;} \\ \textit{if } B(x) \geq a, \textit{ accept } H_0 \textit{ and report the conditional error probability } \beta^*(B(x)). \end{cases}$$

The "surprise" observed in BBW (see also Wolpert, 1995) is that T_1^* is also a conditional frequentist test, arising from use of the conditioning statistic

$$S(X) = \min\{B(X),\ \psi^{-1}(B(X))\}, \tag{2.11}$$

over the domain $\mathcal{X}^* = \{X : \ 0 \leq S(X) \leq r\}$. (The complement of $\mathcal{X}^*$ is the no-decision region.) Thus, the Conditional Frequentist who uses the acceptance and rejection regions in T_1^*, along with the conditioning statistic in (2.11), will report conditional error probabilities upon accepting or rejecting which are in complete agreement with the Bayesian posterior probabilities. That is, $\alpha(s) = \alpha^*(B)$ and $\beta(s) = \beta^*(B)$.

The main justification for using (2.11) as the conditioning statistic is that it results in all the desirable consequences discussed in the Introduction. In general it is not an ancillary statistic (except under the "symmetry" condition discussed in BBW). We delay further discussion until section 5.

Example 2 (continued).

Simple computation yields $\psi(b) = 2\sqrt{1 - b/2}$, so $\psi(1) = \sqrt{2} > 1$. Hence $r = 1$ and $a = \sqrt{2}$, so that the *no-decision region* is the interval $(1, \sqrt{2})$. The reported error probabilities, upon rejection or acceptance, are again given by (2.8) and (2.9).

2.4 The "No-Decision" Region and Alternate Tests.

The *no decision region* in the new testing procedure can be a source of criticism. Note that, without the *no decision region*, $\mathbf{T}_1^*$ would be the optimal Bayes test $\mathbf{T}_1$ for a Bayesian (who assumes equal prior probabilities of the hypotheses as well as "$0-1$" loss). In a sense, the *no decision region* is the "price" that must be paid in order to have a valid conditional frequentist interpretation for the optimal Bayes test. Thus, the "size" of the *no decision region* is a particularly important feature to study.

We will see considerable numerical evidence that the *no decision region* is typically rather small, containing only moderate $B(x)$ that would rarely be considered to be strong evidence. Furthermore, when the data consists of n iid. observations from m_0 or m_1, the probability content of the *no-decision region* decays exponentially fast to zero (under either hypothesis). To be more precise, from a *large deviation* result (cf. Chernoff, 1972, pp. 44) it follows immediately that, for the test $\mathbf{T}_1^*$,

$$P_i(\text{"}\textit{no decision region}\text{"}) \sim e^{-n\,I} \to 0,$$

for $i = 0, 1$, as $n \to \infty$, where

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

It should also be clear, from (2.10), that the *no decision region* disappears whenever $F_0(1) = 1 - F_1(1)$, in which case $r = a = 1$. This can happen in cases with *Likelihood Ratio Symmetry* (for definition and discussion see BBW).

The *no-decision region* in $\mathbf{T}_1^*$ could be eliminated. An alternative test without such a region, that was proposed in BBW, is

$$T_2^* : \begin{cases} \text{if } B(x) \le c, & \textit{reject } H_0 \textit{ and report the conditional error probability } \alpha^*(B(x)); \\ \text{if } B(x) > c, & \textit{accept } H_0 \textit{ and report the conditional error probability } \beta^*(B(x)); \end{cases}$$

here the "critical value" c is the solution to $F_0(c) = 1 - F_1(c)$ (i.e., the critical value for the classical test with equal error probabilities).

The reason we prefer $\mathbf{T}_1^*$ to $\mathbf{T}_2^*$ is that, from a Bayesian perspective, it is not sensible to accept or reject when the odds favor the opposite action (at least if the hypotheses have equal prior probabilities and the losses of incorrect actions are equal, as we are assuming). Suppose, for instance, that $c = 5$. Then $\mathbf{T}_2^*$ would "reject H_0" when $B(x) = 4$, even though $B(x) = 4$ would typically be interpreted (by a Bayesian) as 4 to 1 evidence in favor of H_0. For a Bayesian, the inclusion of the *no decision region* prevents this counterintuitive behavior from occurring.

Even for a classical frequentist, the inclusion of a *no decision region* helps alleviate some paradoxical behavior of the unconditional test. To see this, consider two traditional (unconditional) statisticians, A and B, who intend, based on the *same* observation x on X, to construct a size α most powerful test (as given in (2.3)) for testing whether $X \sim m_0(x)$ or $X \sim m_1(x)$. Further, suppose that both statisticians are indifferent to the choice of the p.d.f. for the null hypothesis.

- Statistician A chooses the hypotheses to be

$$H_0^A : X \sim m_0(x) \quad vs. \quad H_1^A : X \sim m_1(x),$$

and constructs the size α most powerful test as:

$$\begin{cases} \textit{if } B(x) \leq c_0, & \textit{reject } H_0^A \\ \textit{if } B(x) > c_0, & \textit{accept } H_0^A \end{cases},$$

where the critical value c_0 is determined by the equation $F_0(c_0) = \alpha$.

- Statistician B chooses the hypotheses to be

$$H_0^B : X \sim m_1(x) \quad vs. \quad H_1^B : X \sim m_0(x),$$

and constructs the size α most powerful test as:

$$\begin{cases} \textit{if } B(x) \geq c_1, & \textit{reject } H_0^B \\ \textit{if } B(x) < c_1, & \textit{accept } H_0^B \end{cases},$$

where, in this case, the critical value c_1 is determined by the equation $1 - F_1(c_1) = \alpha$. Here, as in (2.2), $B(x) = m_0(x)/m_1(x)$.

The difficulty arises whenever $c_0 \neq c_1$, in which case the set

$$\{x : \min(c_0, c_1) < B(x) < \max(c_0, c_1)\ \}$$

is not empty. This set is the set of *disagreement* between the two statisticians, where they will reach different conclusions. This is troubling if their initial feelings about the two hypotheses were symmetric, is terms of (say) loss and believability, and if they felt required to use (say) a specified Type I error α.

This conflict can easily be resolved, however, if one is willing to modify the classical test in (2.3) to incorporate the possibility of *no-decision.* With this in mind, let $r_0 \equiv \min(c_0, c_1)$ and $a_0 \equiv \max(c_0, c_1)$; then the modification of the classical test (2.3) for the simple hypotheses (2.1), which includes a *no decision region*, is:

$$\begin{cases} \textit{if } B(x) \le r_0, \textit{ reject } H_0; \\ \textit{if } r_0 < B(x) < a_0, \textit{ make no decision;} \\ \textit{if } B(x) \ge a_0, \textit{ accept } H_0 \ . \end{cases}$$

Another way of saying this is that, if it is desired to treat m_0 and m_1 symmetrically, with error probabilities of Type I and Type II both to equal a specified α, then introduction of a no-decision region is necessary.

Example 2 (continued).

With a predetermined and desired probability α of the Type I error, simple calculations yield $c_0 = 2\sqrt{\alpha}$, and $c_1 = 2(1-\alpha)$. The *disagreement* region between statisticians A and B disappears only with $\alpha = 0.3819$, at which point $c_0 = c_1 = 1.2360$. This, of course, would also be the "critical value" used in the alternative test $\mathbf{T}_2^*$. With $\alpha = 0.25$, the *disagreement* region between the two statisticians is $(r_0,\ a_0) = (1,\ 1.5)$, somewhat larger than the *no decision region* $(1,\ \sqrt{2})$ obtained in $\mathbf{T}_1^*$. Observe that, as α decreases, the *disagreement* region increases in size. For instance, with $\alpha = 0.05$, this region is $(0.4472,\ 1.9)$.

3. Testing a composite hypothesis.

The test $\mathbf{T}_1^*$ can also be used in the composite hypothesis case. Suppose we observe the realization, x, of the random variable $X \in \mathcal{X}$ from a density $f(x|\theta)$, with θ being an unknown element of the parameter space Θ. In the sequel, we let $P_\theta(\cdot)$ denote conditional probability given $\theta \in \Theta$. Consider the problem of testing simple versus composite hypotheses as given by

$$H_0 : \ \theta = \theta_0 \quad \text{versus} \quad H_1 : \ \theta \in \Theta_1, \tag{3.1}$$

where $\theta_0 \notin \Theta_1 \subset \Theta$. Often we will take Θ_1 to be $\Theta_1 = \{\theta \in \Theta : \ \theta \neq \theta_0\}$. As in Section 2.2, we assume the default prior probability $\pi_0 = \frac{1}{2}$ for the simple hypothesis $H_0 : \theta = \theta_0$, while assigning to Θ_1 the prior density $g(\theta)/2$, where g is a proper p.d.f. over Θ_1. We observe, in passing, that testing a point null hypothesis should typically be thought of as an approximation to testing the more realistic hypothesis $H_0 : \ |\theta - \theta_0| < \epsilon$ for some small $\epsilon > 0$ (cf., Berger and Delampady, 1987).

For this case, the Bayes factor in favor of H_0 is exactly as given in (2.2), i.e. $B(x) = m_0(x)/m_1(x)$, but now with $m_0(x) = f(x|\theta_0)$ and

$$m_1(x) = \int_{\Theta_1} f(x|\theta)g(\theta)d\theta. \tag{3.2}$$

Note that m_1 and m_0 are the marginal densities of X conditional on H_1 and H_0 being true, respectively. (For a Frequentist, g might be thought of as a weight function which allows computation of an average likelihood for H_1, namely, $m_1(x)$ in (3.2).) For a Bayesian, the test of (3.1) can thus be reduced to the equivalent test of the *"simple"* hypotheses $H_0 : X \sim m_0(x)$ versus $H_1 : X \sim m_1(x)$. Hence, modulo the no decision region, the modified Bayesian test, $\mathbf{T}_1^*$, is the natural Bayesian test of the hypotheses in (3.1).

For the Conditional Frequentist who wishes to test $H_0 : \theta = \theta_0$ against $H_1 : \theta \in \Theta_1$, the conditional error probabilities arising from (2.4) would be

$$\alpha(s) \equiv P_{\theta_0}(\textit{rejecting } H_0 \ |S(X) = s), \tag{3.3}$$

and

$$\beta(\theta|s) \equiv P_\theta(\textit{accepting } H_0 \ |S(X) = s). \tag{3.4}$$

One should observe that, since H_1 in (3.1) is a composite hypothesis, the conditional probability of type II error is a function of θ, analogous to one minus the power function in classical statistics. In the following Theorem, we show that $\mathbf{T}_1^*$ still defines a type of valid conditional frequentist test for this situation.

Theorem 1. *For the test $\mathbf{T}_1^*$ of the hypotheses (3.1) and the conditioning statistic given in (2.11), $\alpha(s) \equiv \alpha^*(B)$ (defined by (2.8)) and*

$$E^{g(\theta|s)}[\beta(\theta|s)] \equiv \beta^*(B), \tag{3.5}$$

where $g(\theta|s)$ denotes the posterior p.d.f. of θ conditional on H_1 being true and on the observed value of $S(X)$.

The equality of $\alpha(s)$ and $\alpha^*(B)$ in the above theorem was, in a sense, our primary goal: the conditional Type I error probability and the posterior probability of H_0 are equal. Since Type I error is (rightly or wrongly) perceived to be of primary interest in classical statistics, the agreement of the two reports for the suggested procedure is, perhaps, crucial to its acceptance.

The situation for Type II error is more complicated because the frequentist probability of Type II error necessarily depends on the unknown θ, while $\beta^*(B)$, the posterior probability of H_1, is necessarily a fixed number. The relationship in (3.5) between $\beta^*(B)$ and the conditional frequentist Type II error probability, $\beta(\theta|s)$, is however, quite natural: $\beta^*(B)$ can be interpreted as the average of the conditional Type II error probabilities, with the average being with respect to the posterior distribution of θ given s. Intuitively, this averaging is a considerable improvement over the common classical practice of simply picking a plausible value of θ and reporting the power at that value.

Of course, there is nothing to prevent a frequentist from reporting the entire function $\beta(\theta|s)$ (or the conditional power function, $1-\beta(\theta|s)$). Indeed, one might argue that this is beneficial if the prior distribution has been chosen in a "default" fashion (cf. Jeffreys, 1961), since alternative "averages" of $\beta(\theta|s)$ might be desired. In practice, however, the simplicity of just reporting $\beta^*(B)$ will probably be hard to resist.

There is one oddity here from a Bayesian perspective. It is that $\beta^*(B)$ is not the average Type II error with respect to the posterior distribution of θ given H_1 and the data, but is instead the average Type II error with respect to the posterior distribution given H_1 and given $S = s$. The difference between these two posteriors is typically not too great, however. In any case, conditioning on S is, in a sense, the most conditioning that is allowed for a frequentist and, from the Bayesian perspective, the final answer, $\beta^*(B)$, is fine.

4. Some applications

We present several applications to standard testing problems. To simplify the notation, we let, in this section, $\alpha^*(x) \equiv \alpha^*(B(x)) = B(x)/(1+B(x))$ and $\beta^*(x) \equiv \beta^*(B(x)) = 1/(1+B(x))$.

Example 3: (Two-sided Normal Testing)

We consider the same basic setup of Example 1: based on $\overline{X}_n \sim \mathcal{N}(\theta, \sigma^2/n)$, σ^2 known, we wish to test

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta \neq \theta_0, \tag{4.1}$$

for some specified value of θ_0. A natural choice of the conditional prior (given H_1 is true) for θ over the set $\Theta_1 \equiv \{\theta \neq \theta_0\}$ is a conjugate prior. Hence we assume that g in (3.2) is the $\mathcal{N}(\mu, k\sigma^2)$ p.d.f. Here μ and k are assumed to be known. The parameter μ is the conditional prior mean of θ, given H_1 is true. This allows, under H_1, a measurable *shift* of the conditional prior p.d.f. of θ away from H_0. Let $\Delta = (\theta_0 - \mu)/\sqrt{k}\sigma$. When $\Delta = 0$, the prior p.d.f. is symmetric about θ_0. This choice of Δ is often considered as the default choice for applications, and was used in Example 1. Also in Example 1, the default choice of $k = 2$ was made; the resulting $\mathcal{N}(0, 2\sigma^2)$ prior is similar to the Cauchy $(0, \sigma^2)$ default prior recommended by Jeffreys (1961).

As before, we let t denote the standard test statistic, $t = \sqrt{n}(\bar{x}_n - \theta_0)/\sigma$. It is easy to verify that the (conditional) marginal p.d.fs. of t corresponding to H_0 and H_1, respectively, are

$$m_0(t) = \phi(t) \equiv \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{t^2}{2}\right\} \tag{4.2}$$

and

$$m_1(t) = \frac{1}{\sqrt{2\pi}\sqrt{1+kn}} \exp\left\{\frac{-(t+\sqrt{kn}\Delta)^2}{2(1+kn)}\right\}. \tag{4.3}$$

Combining (4.2) and (4.3) in (2.2), it follows immediately that the Bayes factor in favor of H_0 is

$$B(t) = \sqrt{1+kn} \exp\left\{-\frac{kn}{2(1+kn)}(t - \frac{\Delta}{\sqrt{kn}})^2 + \frac{\Delta^2}{2}\right\}. \tag{4.4}$$

It can be shown that, in the present case, $\psi(1) > 1$, so that $r = 1$ and $a = \psi(1) \equiv F_0^{-1}(1 - F_1(1))$ in (2.10). Hence the *no decision region* in $\mathbf{T}_1^*$ is of the form $(1, a)$. Accordingly, letting CEP denote *Conditional Error Probability*, the testing procedure $\mathbf{T}_1^*$ is,

$$\mathbf{T}_1^* : \begin{cases} \textit{if } B(t) \le 1, \textit{ reject } H_0 \textit{ and report the CEP } \alpha^*(t) = \frac{B(t)}{B(t)+1}; \\ \textit{if } 1 < B(t) < a, \textit{ make no decision}; \\ \textit{if } B(t) \ge a, \textit{ accept } H_0 \textit{ and report the CEP } \beta^*(t) = \frac{1}{B(t)+1}. \end{cases} \tag{4.5}$$

In this case, no explicit expression for the critical value a is available, but a can be found using the following set of equations. For any $b > 0$, let $t_b^{\pm}$ be the two solutions of the equation $B(t) = b$; it follows from (4.4) that

$$t_b^{\pm} = \frac{\Delta}{\sqrt{kn}} \pm \sqrt{\frac{1+kn}{kn}(\log(\frac{1+kn}{b^2}) + \Delta^2)}. \tag{4.6}$$

Using (4.6), the value of a is determined by the equation

$$\Phi(-t_a^+) + \Phi(t_a^-) = \Phi(\Delta_k^+) - \Phi(\Delta_k^-), \tag{4.7}$$

where $t_a^\pm$ is given by (4.6) and $\Delta_k^\pm = (\Delta\sqrt{1+kn} \pm \sqrt{\log(1+kn)+\Delta^2})/\sqrt{kn}$. It is clear that $a \equiv a(kn, \Delta)$ depends on Δ and (with a known k) on the sample size n. In the following table we present values of a for several choices of Δ and kn. Note also, that, for the suggested default choices $k = 2$ and $\Delta = 0$, a closed form approximation to a (accurate to within 1%) was given in Example 1.

Table 2. Values of $a(kn, \Delta)$, for the normal two-sided test.

kn	$\lvert\Delta\rvert$=0	1	2	3	4	5
1	1.317	1.655	1.777	1.793	1.780	1.802
2	1.530	1.987	2.301	2.344	2.359	2.367
3	1.691	2.202	2.710	2.768	2.798	2.808
4	1.822	2.369	3.036	3.137	3.165	3.178
5	1.932	2.506	3.306	3.449	3.483	3.500
6	2.028	2.621	3.536	3.727	3.767	3.786
7	2.113	2.722	3.735	3.978	4.023	4.045
8	2.189	2.812	3.910	4.208	4.259	4.282
9	2.258	2.893	4.066	4.420	4.478	4.503
10	2.321	2.966	4.206	4.617	4.683	4.710
15	2.576	3.256	4.744	5.442	5.559	5.593
20	2.768	3.471	5.121	6.085	6.272	6.314
25	2.922	3.642	5.407	6.608	6.882	6.936
30	3.051	3.783	5.637	7.046	7.421	7.490
40	3.260	4.010	5.990	7.749	8.343	8.455
50	3.425	4.188	6.257	8.293	9.116	9.287
60	3.563	4.336	6.470	8.732	9.781	10.026
70	3.681	4.462	6.647	9.096	10.362	10.694
80	3.784	4.571	6.798	9.404	10.878	11.305
90	3.876	4.668	6.929	9.671	11.338	11.868
100	3.958	4.756	7.045	9.903	11.754	12.390

Illustration 2: Fisher and Belle (1993) provide the birthweights in grams of some $n = 15$ cases of SIDS (Sudden Infant Death Syndrome) born in King County in 1977:

2013 3827 3090 3260 4309
3374 3544 2835 3487 3289
3714 2240 2041 3629 3345

With the standing assumption of normality and a supposed known standard deviation of $\sigma = 800$ g, we consider the test of

$$H_0 : \theta = 3300 \quad \text{versus} \quad H_1 : \theta \neq 3300;$$

here 3300 g is the overall average birthweight in King County in 1977 (which can effectively be considered to be known), so that H_0 would correspond to the (believable) hypothesis that SIDS is not related to birthweight.

We apply the test (4.5) with $\Delta = 0$ and the default choice of $k = 2$. From Table 2, we find $a(30, 0) = 3.051$, and simple calculations yield $t = 0.485$ and $B(t) = 4.968$, so that $B(t) > a$. Thus, according to $\mathbf{T}_1^*$, we accept H_0 and report the CEP $\beta^* = 0.201$.

One can, alternatively, write the test $\mathbf{T}_1^*$ in terms of the standard statistic, t, as follows:

$$\mathbf{T}_1^* : \begin{cases} \textit{if } t \le t_1^- \textit{ or } t \ge t_1^+, \textit{ reject } H_0 \textit{ and report the CEP } \alpha^*(t); \\ \textit{if } t_1^- < t < t_a^- \textit{ or } t_a^+ < t < t_1^+, \textit{ make no decision}; \\ \textit{if } t_a^- \le t \le t_a^+, \textit{ accept } H_0 \textit{ and report the CEP } \beta^*(t). \end{cases}$$

Figure 1 below illustrates the effect of the *"shift"* parameter Δ on the *no decision region* corresponding to the test $\mathbf{T}_1^*$. Note the symmetry of the regions when $\Delta = 0$ and that the size of the *no decision region* decreases as Δ increases.

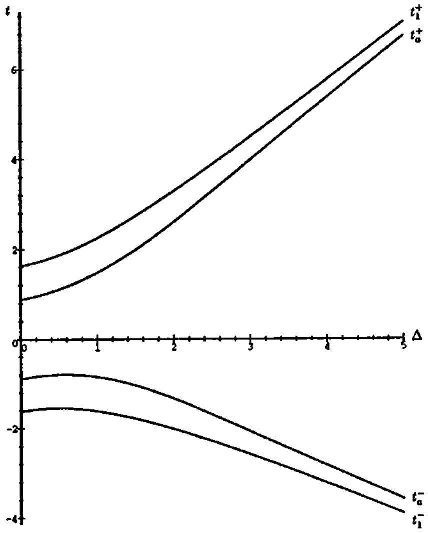

Figure 1. The *no decision region* of $\mathbf{T}_1^*$ as a function of Δ and with $kn = 10$, for the normal two-sided test of Example 3.

Example 4: (One-sided Normal Testing)

We continue with the same basic setup of Example 3, but now we wish to test the hypotheses

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0.$$

The choice of conditional prior (given H_1 is true) for θ over the set $\Theta_1 \equiv \{\theta > \theta_0\}$

is

$$g(\theta) = \frac{2}{\sqrt{k}\sigma}\phi(\frac{\theta-\theta_0}{\sqrt{k}\sigma}), \quad \theta > \theta_0.$$

With this prior p.d.f., the marginal p.d.f. (3.2) (given H_1 is true) of t becomes

$$m_1(t) = \frac{2}{\sqrt{1+kn}}\phi(\frac{t}{\sqrt{(1+kn)}})\Phi(\frac{knt}{\sqrt{1+kn}}).$$

Note that, in this case, $m_0(t)$ remains unchanged. Hence the corresponding Bayes factor can be written as

$$B(t) = \frac{\sqrt{1+kn}}{2}\exp\{\frac{-knt^2}{2(1+kn)}\}\left(\Phi(\frac{knt}{\sqrt{1+kn}})\right)^{-1}.$$

Again, it can be verified that the *no decision region* is of the form $(1, a)$, where a can be determined numerically by the following set of equations:

$$\begin{cases} B(t_1) = 1; \qquad B(t_a) = a; \\ 1 - \Phi(t_a) = 2\int_{-\infty}^{t_1/\sqrt{1+kn}} \Phi(knt)\phi(t)dt. \end{cases}$$

Thus the test $\mathbf{T}_1^*$ (as presented in terms of the standard test statistic t) is

$$\mathbf{T}_1^* : \begin{cases} \textit{if } t \geq t_1, & \textit{reject } H_0 \textit{ and report the CEP } \alpha^*(t); \\ \textit{if } t_a < t < t_1\,, & \textit{make no decision}; \\ \textit{if } t \leq t_a\,, & \textit{accept } H_0 \textit{ and report the CEP } \beta^*(t). \end{cases}$$

Table 3 presents values of a, t_a and t_1 for selected choices of kn. Note that the *no decision region* is somewhat smaller than for the two-sided test.

Table 3. Values of a, t_a and t_1 for the normal one-sided test

kn	a	t_a	t_1
1	1.271	0.183	0.560
2	1.448	0.262	0.731
3	1.580	0.320	0.841
4	1.858	0.367	0.923
5	1.774	0.406	0.987
6	1.851	0.4.6	1.040
7	1.918	0.469	1.085
8	1.979	0.495	1.124
9	2.034	0.519	1.159
10	2.084	0.541	1.190
15	2.285	0.627	1.308

kn	a	t_a	t_1
20	2.436	0.690	1.390
25	2.558	0.740	1.454
30	2.659	0.781	1.505
35	2.747	0.817	1.548
40	2.825	0.847	1.584
50	2.956	0.898	1.645
60	3.066	0.940	1.693
70	3.161	0.976	1.734
80	3.244	1.006	1.768
90	3.318	1.033	1.799
100	3.385	1.057	1.825

Example 5: (ANOVA I)

Consider p independent samples $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{in})$, $(i = 1, \ldots, p)$, of n i.i.d. random variables from the $\mathcal{N}(\mu_i, \sigma^2)$ distribution, with unknown σ^2. We are interested in testing

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_p = 0 \tag{4.8}$$

against the standard alternative H_1 : *not all* μ_i *are equal to* 0. Note that, when $p = 1$, this is the standard two-sided test with unknown σ^2.

We will use a hierarchical prior defined as follows. Let the μ_i, $i = 1, \ldots, p$, be *i.i.d.* with a first-stage $\mathcal{N}(0, \xi\sigma^2)$ prior distribution, to be denoted by $\pi_1(\mu_i|\sigma^2, \xi)$. Let the second-stage prior be $\pi_2(\sigma^2, \xi) = \sigma^{-2} g(\xi) d\sigma^2 d\xi$; thus σ^2 is given the usual noninformative prior and $\xi > 0$ is given the proper prior p.d.f. g (to be defined later). Straightforward computation yields, as the Bayes factor of H_0 to H_1,

$$B(y) = (n - 1 + y)^{-pn/2} \times \left[\int_0^\infty \frac{(1 + n\xi)^{p(n-1)/2}}{[(n-1)(1+n\xi) + y]^{pn/2}} g(\xi) d\xi \right]^{-1}, \tag{4.9}$$

where

$$y = \frac{(n-1)n \sum_{i=1}^p (\bar{x}_i)^2}{\sum_{i=1}^p \sum_{j=1}^n (x_{ij} - \bar{x}_i)^2}. \tag{4.10}$$

To proceed with a Conditional Frequentist interpretation of the Bayes test, we need to slightly reformulate the test. The difficulties are that (i) H_0 is, itself, composite; and (ii) improper prior distributions were used. The most direct solution is to initially suppose that we will base the test on the statistic, y, in (4.10). We have seen a Bayesian justification for doing so; and y is the standard classical F test statistic for the testing problem at hand.

Write the density of y as $f(y|\theta_1, \ldots, \theta_p)$, where $\theta_i = \mu_i/\sigma$. Then the test can be rewritten as a test of $H_0 : \theta_1 = \theta_2 = \cdots = \theta_p = 0$, which is a simple hypothesis. Furthermore, under H_1, the hierarchical prior defined earlier becomes: the $\pi_1(\theta_i|\xi)$ are $\mathcal{N}(0, \xi)$, independently for $i = 1, \ldots, p$, while ξ still has proper prior $g(\xi)$. The implied prior, $\pi(\theta_1, \ldots, \theta_p)$, is thus proper, and Theorem 1 can be applied. Note that, here,

$$m_0(y) = m(y|0) \qquad \text{and} \qquad m_1(y) = \int_0^\infty m(y|\xi) g(\xi) d\xi,$$

where

$$\begin{aligned} m(y|\xi) &= \int f(y|\theta_1, \ldots, \theta_p) \pi(\theta_1, \ldots, \theta_p) d\theta_1, \ldots, d\theta_p \\ &= C \, \frac{y^{p/2-1}(1 + n\xi)^{p(n-1)/2}}{[(n-1)(1+n\xi) + y]^{pn/2}}, \end{aligned} \tag{4.11}$$

with

$$C = \frac{\Gamma(\frac{np}{2})(n-1)^{p(n-1)/2}}{\Gamma(\frac{p}{2})\Gamma(\frac{p(n-1)}{2})}.$$

The test $\mathbf{T}_1^*$, from Section 3, can thus be written as

$$\text{(4.12)} \qquad \mathbf{T}_1^* : \begin{cases} \textit{if } B(y) \le 1, \textit{ reject } H_0 \textit{ and report the CEP } \alpha^*(y); \\ \textit{if } 1 < B(y) < a, \textit{ make no decision}; \\ \textit{if } B(y) \ge a, \textit{ accept } H_0 \textit{ and report the CEP } \beta^*(y). \end{cases}$$

Here, using (4.9) and (4.11), a (as well as y_1 and y_a) can be solved numerically from the following system of equations:

$$\text{(4.13)} \qquad \begin{cases} B(y_1) = 1; \qquad B(y_a) = a; \\ \int_{y_a}^{\infty} m(y|0)dy = \int_0^{y_1} \int_0^{\infty} m(y|\xi)g(\xi)d\xi dy. \end{cases}$$

In terms of the statistic y in (4.10), this test has the form

$$\mathbf{T}_1^* : \begin{cases} \textit{if } y \ge y_1, \textit{ reject } H_0 \textit{ and report the CEP } \alpha^*(y); \\ \textit{if } y_a < y < y_1, \textit{ make no decision}; \\ \textit{if } y \le y_a, \textit{ accept } H_0 \textit{ and report the CEP } \beta^*(y). \end{cases}$$

As an illustration, consider the case with $p = 1$; clearly this is equivalent to the normal two-sided test with unknown σ^2. Note that, in this case, $y \equiv t^2$, where t denotes the standard t-test statistic. In comparison, the classical α-level two-sided test of (4.8) (with $p = 1$) can be given in terms of the statistic (4.10) as

$$\begin{cases} \textit{if } y > t^2_{\alpha/2} \qquad \textit{reject } H_0 \textit{ and report error probability } \alpha \text{ ;} \\ \textit{if } y \le t^2_{\alpha/2}\text{ ,} \quad \textit{accept } H_0 \textit{ and report the probability of Type II error} \text{ ;} \end{cases}$$

here $t_{\alpha/2}$ is the $\frac{\alpha}{2}$-level critical value from the $t_{(n-1)}-$ distribution.

The default prior, $g(\xi)$, that we recommend for this testing problem is

$$\text{(4.14)} \qquad g(\xi) = \frac{1}{\sqrt{2\pi}}\xi^{-\frac{3}{2}} \exp\{-\frac{1}{2\xi}\}.$$

This prior yields, for $p = 1$, the analysis recommended by Jeffreys (1961), since it can be shown that $\pi(\mu \mid \sigma^2)$ (formed by integrating over ξ) is then Cauchy$(0,\ \sigma^2)$. In Table 4, we present the value of $t_{0.025}$ along with the values of a, $\sqrt{y_1}$ and $\sqrt{y_a}$ as were determined numerically for selected choices of n under the prior (4.14).

Illustration 2 (continued): Now assume that σ is unknown. This corresponds to the case of $p = 1$ in the null hypothesis (4.8) above. The calculated value of the test statistic (4.10) is $y = 0.343$. For the default prior (4.14), we find from Table 4 that $\sqrt{y_a} = 1.123$. Thus again, we accept H_0 and report CEP $\beta^* = 0.186$ (computed from (4.9)).

Table 4. Values of a and critical points for the normal two-sided test with unknown σ^2.

n	a	$\sqrt{y_a}$	$\sqrt{y_1}$	$\lvert t_{0.025}\rvert$
2	1.302	1.342	1.983	12.706
3	1.732	1.035	1.881	4.303
4	1.962	0.993	1.863	3.182
5	2.123	0.991	1.864	2.776
6	2.250	1.001	1.872	2.571
7	2.356	1.015	1.883	2.447
8	2.447	1.030	1.894	2.365
9	2.528	1.045	1.905	2.306
10	2.600	1.060	1.917	2.262
11	2.665	1.074	1.928	2.228
12	2.725	1.087	1.939	2.201
13	2.781	1.100	1.949	2.179
14	2.832	1.112	1.959	2.160
15	2.880	1.123	1.968	2.145
20	3.083	1.174	2.011	2.093
25	3.242	1.215	2.046	2.064
30	3.374	1.250	2.076	2.045
35	3.486	1.280	2.102	2.032
40	3.583	1.306	2.126	2.023
45	3.669	1.329	2.147	2.015
50	3.746	1.351	2.165	2.010
55	3.815	1.370	2.183	2.005
60	3.879	1.387	2.199	2.001
65	3.937	1.404	2.213	1.998
70	3.991	1.419	2.227	1.995
80	4.087	1.447	2.252	1.990
90	4.172	1.471	2.273	1.987
100	4.247	1.493	2.293	1.984

For general p, the choice of $g(\xi)$ in (4.14) results in $\pi(\mu|\sigma^2)$ being the p-variate t-distribution with location $\mathbf{0}$ and scale matrix $\sigma^2\mathbf{I}$ and one degree of freedom. Note that the introduction of ξ allows $B(y)$ in (4.9) to be computed by one-dimensional integration, regardless of p.

The choice of $g(\xi)$ in (4.14) is not the only "default" choice that is reasonable. In particular, this choice of g implies that $\lambda \equiv \sum_{i=1}^{p} \mu_i/\sigma^2$ has a prior density which is roughly proportional to $\lambda^{(p-1)/2}$ for small λ. Sometimes, however, (4.8) is more naturally thought of as testing $H_0 : \lambda = 0$ versus $H_a : \lambda > 0$, in which case a prior density for λ which is positive at zero may be more intuitively appealing. A choice of g that achieves this goal is $g(\xi) = \frac{1}{2}(1+\xi)^{-3/2}$. The resulting prior has the same tail behavior for large λ as the earlier choice, but is positive at zero.

Example 6: (ANOVA II)

We continue with the same basic setup as in Example 5, but now, we are interested in testing, with $p > 1$, the composite hypothesis

$$(4.15) \qquad H_0 : \mu_1 = \mu_2 = \cdots = \mu_p \text{ (equal to, say, } \mu)$$

against the alternative H_1 : *not all μ_i are equal* . We assume a similar hierarchical prior structure for this testing problem: choose as the first-stage prior,

$\pi_1(\mu_i|\sigma^2,\ \xi)$, the $\mathcal{N}(\mu, \xi\sigma^2)$ distribution for the *i.i.d.* $\mu_1,\ \mu_2,\ \ldots,\ \mu_p$; choose, for the second-stage prior, the usual noninformative prior for $(\mu,\ \sigma^2)$, i.e. $\pi_2(\mu,\ \sigma^2) = (1/\sigma^2)d\mu\, d\sigma^2$, while (independently) ξ is given the proper p.d.f. $g(\xi)$.

It can be shown that the Bayes test and the classical test are based on the usual F statistic

$$y = \frac{p(n-1)n\sum_{i=1}^{p}(\bar{x}_i - \bar{\bar{x}})^2}{(p-1)\sum_{i=1}^{p}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2},$$

and that the test can be reformulated, as in Example 5, with $\theta_i = (\mu_i - \mu)/\sigma$ and $m(y|\xi)$ given by

$$m(y|\xi) = C\,\frac{y^{(p-3)/2}(1+n\xi)^{p(n-1)/2}}{[p(n-1)(1+n\xi)+(p-1)y]^{(pn-1)/2}}, \qquad (4.16)$$

with

$$C = \frac{\Gamma(\frac{np-1}{2})[p(n-1)]^{p(n-1)/2}(p-1)^{(p-1)/2}}{\Gamma((p-1)/2)\Gamma(p(n-1)/2)}.$$

The corresponding Bayes factor has a similar form to that of Example 5, namely

$$\begin{aligned} B(y) &= (p(n-1)+(p-1)y)^{-(pn-1)/2} \\ &\times \left[\int_0^\infty \frac{(1+n\xi)^{p(n-1)/2}}{(p(n-1)(1+n\xi)+(p-1)y)^{(pn-1)/2)}}g(\xi)d\xi\right]^{-1}. \end{aligned} \qquad (4.17)$$

Now, for any specified prior $g(\xi)$, the test $\mathbf{T}_1^*$ of the hypotheses (4.15) follows exactly as in Example 5. The values of a, y_1 and y_a are determined numerically, using (4.16) and (4.17) in the equations (4.13). In Table 5, we provide the values of a for selected choices of n and p under the prior (4.14) for $g(\xi)$.

Table 5. Values of a for the ANOVA II test.

	p=2	4	6	8	10
n=2	1.654	1.742	1.847	1.934	2.007
3	1.995	2.135	2.237	2.320	2.388
4	2.133	2.372	2.474	2.552	2.616
5	2.267	2.545	2.645	2.719	2.778
6	2.377	2.683	2.779	2.848	2.903
7	2.471	2.797	2.889	2.953	3.004
8	2.553	2.895	2.983	3.043	3.090
9	2.626	2.981	3.065	3.120	3.163
10	2.692	3.058	3.137	3.188	3.227
n=20	3.155	3.568	3.607	3.622	3.634
30	3.439	3.874	3.885	3.876	3.868
40	3.648	4.098	4.088	4.061	4.038
50	3.814	4.276	4.250	4.208	4.174
60	3.952	4.425	4.386	4.332	4.288
70	4.070	4.552	4.503	4.439	4.387
80	4.173	4.665	4.606	4.534	4.475
90	4.265	4.765	4.699	4.620	4.554
100	4.347	4.856	4.784	4.698	4.627

Illustration 3: (Pappas and Mitchell, 1985). An experiment was conducted to determine whether mechanical stress can retard the growth of soybean plants. Young

plants were randomly allocated to two groups of 13 plants each. Plants in one group were mechanically agitated by shaking for 20 minutes twice daily. At the end of the experiment, the total stem length (cm) of each plant was measured. The raw observations, in increasing order, are as follows:

Control :	25.2 29.5 30.1 30.1 30.2 30.2 30.3
	30.6 31.1 31.2 31.4 33.5 34.3
Stress :	24.7 25.7 26.5 27.0 27.1 27.2 27.3
	27.7 28.7 28.9 29.7 30.0 30.6

For these data ($n = 13$ and $p = 2$) we obtain,

$$\bar{x}_1 = 30.59, \quad \bar{x}_2 = 27.78, \quad \text{and} \quad \bar{\bar{x}} = 29.19,$$

$$\sum_{j=1}^{n}(x_{1j} - \bar{x}_1)^2 = 26.65 \quad \text{and} \quad \sum_{j=1}^{n}(x_{2j} - \bar{x}_2)^2 = 21.56,$$

$$y = \frac{p(n-1)n\sum_{i=1}^{p}(\bar{x}_i - \bar{\bar{x}})^2}{(p-1)\sum_{i=1}^{p}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2} = 25.37.$$

The value of the Bayes factor, $B(y)$ in (4.17), is $B(y) = 0.001$. Clearly, using $\mathbf{T}_1^*$, we should reject H_0 and report CEP $\alpha^* = 0.001$.

5. Concluding Remarks

Testing a Precise Hypothesis: In this paper, discussion was restricted to testing of simple hypotheses or testing of a composite alternative hypothesis and a precise (i.e., lower dimensional) null hypothesis. Deciding whether or not to formulate the test as one of testing a precise hypothesis centers on the issue of deciding if there is a believable precise hypothesis. Sometimes this is easy, as in testing for the presence of extrasensory perception, or testing that a proposed law of physics holds. Often it is less clear. In medical testing scenarios, for instance, it is often argued that any treatment will have some effect, even if only a very small effect, and so exact equality of effects (between, say, a treatment and a placebo) will never occur. While perhaps true, it will still often be reasonable to formulate the test as testing the precise hypothesis of, say, zero treatment difference, since such a test can be shown to be a very good approximation to the optimal test unless the sample size is very large (cf., Berger and Delampady, 1987). This is an important issue, because whether one formulates a test as a test of a precise hypothesis or as, say, a one-sided test can make a huge difference in the Bayesian posterior probabilities (or conditional frequentist error probabilities), in contrast to classical unconditional testing, where the error probabilities only vary by a factor of two. Since this issue is so important in Bayesian or conditional testing, we will belabor the point with an additional illustration.

Illustration 4. Suppose one is comparing a standard chemotherapy treatment for cancer with a new radiation treatment. There is little reason to suspect that the two treatments could have the same effect, so that the correct test would be a one-sided test comparing the two treatments. If, instead, the second treatment had been the same chemotherapy treatment, but now with (say) steroids added, then equality of treatments would have been a real possibility, since the steroids could have no substantial additional effect on the cancer. Hence one should now test the precise hypothesis of no treatment difference, using the Bayesian or conditional frequentist test. (We do not mean to imply that one need only carry out the relevant test here; rather we are saying that the relevant test is important to do as part of the overall analysis.) Note that both null hypotheses in Illustrations 2 and 3 are believable hypotheses.

A final comment on this issue is that precise hypothesis testing should not be done by forming a traditional confidence interval (frequentist or Bayesian) and simply checking whether or not the precise hypothesis is compatible with the confidence interval. A confidence interval is usually of considerable importance in determining where the unknown parameter (say) is likely to be, given that the alternative hypothesis is true, but it is not useful in determining whether or not a precise null hypothesis is true.

Choice of the Conditioning Statistic: The first point to stress is the unreasonable nature of the unconditional test, and the even more unreasonable nature of common procedures such as the P-value; in some sense, these are the worst possible testing procedures, and any reasonable conditional frequentist tests are better. (We hope to be able to indicate this more formally in subsequent work.) Furthermore, the conditional tests we propose have the feature of conditioning as much as is possible; there is typically no natural reason to stop short of the maximal possible degree of conditioning (see, e.g., the Discussion and Rejoinder in Kiefer, 1977). Unfortunately, among these (perhaps optimal) conditional tests, there is apparently no single optimal choice. In particular, choice of the conditioning statistic may seem rather uncertain and arbitrary.

Conditioning on ancillary statistics is familiar but, as mentioned earlier, suitable ancillary statistics rarely exist for testing. Furthermore, it is far from clear that conditioning on ancillary statistics is always best. Consider Example 2, for instance. Conditioning on the ancillary statistic led to a conditional Type II error probability that was actually constant over the acceptance region, even though the likelihood ratio (or Bayes factor) varied by a factor of two over that region! In contrast, our recommended conditioning statistic led to conditional Type II error probabilities that varied quite sensibly over the acceptance region.

It is sometimes argued that conditioning on non-ancillary statistics will "lose information" but nothing loses as much information as unconditional testing (effectively replacing the data by the indicator on its being in the acceptance or rejection region); and since our conditioning leads to Bayesian posterior probabilities as the conclusion, it is hard to see what information is "being lost." Finally, it is crucial to remember all of the advantages (mentioned in the introduction) that accrue from using a conditioning statistic that results in error probabilities with a Bayesian interpretation.

Choice of the Prior on the Alternative Hypothesis: This is the stickiest issue: each choice of prior distribution on the parameter space of the alternative hypothesis will lead to a different conditioning statistic, and hence to a different conditional frequentist test. In one sense this is wonderful, in that it says that both Bayesians and frequentists have the same problem; whether one chooses to phrase the problem in terms of choice of the prior distribution or choice of the conditioning statistic is simply a matter of taste. (Of course it can be argued that choice of the prior is much more intuitively accessible than is choice of the conditioning statistic.) But that does not settle the question of what to do.

A subjective Bayesian has a ready answer: "Elicit your subjective prior distribution on the parameter space of the alternative hypothesis, and use the Bayes test; if you wish to use a conditional frequentist test, use that with the corresponding conditioning statistic." (Actually, of course, the subjective Bayesian would also insist that the prior probabilities of the hypotheses be elicited and utilized. That would require the modifications discussed in BBW.)

We have no disagreement with this answer, except that we also want to provide an automatic test, for those who are unable or unwilling to elicit a prior distribution. What we have done in section 4, therefore, is to define what we consider to be attractive "default" Bayesian tests (following Jeffreys, 1961), and provide their conditional frequentist analogues. This, in fact, defines a new joint Bayesian - frequentist research agenda for testing: develop attractive default Bayesian tests for all situations, and then translate them into their conditional frequentist analogues. (For the development of general default Bayesian procedures, two interesting recent approaches are described in Berger and Pericchi, 1996, and O'Hagan, 1995.)

We have frequently heard the comment that non-Bayesians will not accept these conditional frequentist procedures because their development utilizes a prior distribution. It seems absurd, however, to reject a procedure that is arguably greatly superior from a pure frequentist perspective, simply because a Bayesian tool was used in its derivation. We suspect, therefore, that what is really intended by such comments is to suggest that the appearance of statistical objectivity is often considered to be important, and that there is concern that a procedure that uses a prior distribution will not be perceived to be objective. While not passing judgment here on the possibility or desirability of "objectivity," we would argue that the proposed default conditional tests have every bit as much claim to objectivity as any other frequentist procedure. They are specific procedures that can be used without subjective input, and have highly desirable frequentist properties that can be evaluated on their own merits.

Generalizations: We have not considered situations involving composite null hypotheses, except those that can be reduced to simple hypotheses by some type of invariance reduction (e.g., ANOVA II). In principle, composite null hypotheses can be treated in the same fashion as composite alternative hypotheses; i.e., be reduced to simple hypotheses by Bayesian averaging. This will be a far more controversial step for frequentists, however, since classically the treatment of null hypotheses and alternatives has been very asymmetric. For instance, many frequentists will welcome the notion of "average" power that arises from the conditional frequentist tests that we consider, but will perhaps be wary of any notion of "average" Type I error.

As discussed in BBW, the general framework applies equally well to sequential experiments. One can develop conditional frequentist tests that essentially agree with Bayesian tests, and hence which essentially ignore the stopping rule. This is potentially revolutionary for, say, clinical trials. It appears necessary, however, to "fine tune" the new sequential tests, so as to obtain a satisfactory tradeoff between the size of the no-decision region and the expected sample size of the experiment. This work will be reported elsewhere.

Other Approaches and Comparison: A number of other approaches to data-dependent inference for testing have been recently proposed. These include the developments in Bernardo (1980), Hwang, Casella, Robert, Wells, and Farrell (1992), Schaafsma and van der Meulen (1993), Evans (1994), and Robert and Caron (1995). While being interesting and worthy of study, these alternative approaches all have one or more of the following disadvantages: (i) requiring new evidential concepts that would require extensive study and experience to properly understand; (ii) possessing significantly non-Bayesian or non-frequentist properties, which would prevent members of either paradigm from accepting the approach; and (iii) being difficult to implement in all but relatively simple situations.

In contrast, the approach we advocate possesses none of these disadvantages. It does not really involve new concepts, since conditional error probabilities are quite familiar to most statisticians (and can, in any case, be understood with the usual frequentist logic); likewise the interpretation of Bayesian posterior probabilities is familiar (and very easy besides). One might argue that it is difficult to develop and understand the recommended conditioning statistic, but this understanding is really only necessary for those developing the methodology. Most practitioners would need only to know the actual test procedure, and that the reported error probabilities can either be interpreted as posterior probabilities (with, say, default priors), or as frequentist error probabilities conditioned on a reasonable statistic reflecting the strength of evidence in the data. Note, in particular, that the actual conditioning statistic need not be presented in an applied statistical report, any more than one now needs to present all the background properties of the particular unconditional test that is chosen. This is assuming, of course, that a default conditioning statistic is being used, rather than one tailored to subjective prior beliefs; in the latter case, reporting the conditioning statistic (or better, the prior) would seem only fair.

Likewise, the testing paradigm we propose should be acceptable to both frequentists and Bayesians. Although the proposed tests are mainly traditional Bayesian tests, it is perhaps the Bayesians who will most object to this paradigm; while there are compelling reasons for frequentists to shift to the conditional frequentist paradigm, there are no compelling reasons for Bayesians to alter their approach. For instance, many Bayesians would see little reason to formally introduce a "no-decision" region.

Some Bayesians might be attracted by the long-run frequentist guarantee that is carried by the new tests, in that the guarantee is independent of the prior distribution. This would seem to imply some type of robustness of the methodology with respect to the prior. The situation is unclear, however, because it could be claimed that it is "robustness for the wrong question." We would, at least, expect Bayesians to agree that these new tests are considerably better than the classical unconditional tests. And, most importantly, the answers obtained in practice by "pure" Bayesians and by non-Bayesians who adopt this new paradigm will now typically be quite similar.

Finally, implementation of the new paradigm is relatively easy, in many cases easier than implementation of classical unconditional testing. This is because Bayesian testing is often much easier to implement than unconditional frequentist testing, and the new tests are essentially based on Bayesian tests. The only significant adaption that is needed is computation of the no-decision region, which is usually a computation of only modest numerical difficulty.

Appendix: Proof of Theorem 1

We will only prove the second assertion since the proof of the first assertion is provided in BBW. We assume that $\psi(1) \geq 1$ in (2.10). The case $\psi(1) < 1$ follows similarly and therefore is omitted.

Let f_i^* denote the p.d.f. of $B(X)$ under m_i, $i = 0, 1$ and let F_θ and f_θ^* be the conditional c.d.f. and p.d.f. (respectively) of $B(X)$ given $\theta \in \Theta_1$ (under $P_\theta(\cdot)$). Notice that, since g is a proper p.d.f. over Θ_1, the following relation holds:

$$\begin{aligned} F_1(b) &= \int_0^b f_1^*(y)dy = \int_{\{B(x)\leq b\}} m_1(x)dx \\ &= \int_{\{B(x)\leq b\}} \int_{\Theta_1} f(x|\theta)g(\theta)d\theta dx = \int_{\Theta_1} \int_{\{B(x)\leq b\}} f(x|\theta)g(\theta)dxd\theta \\ &= \int_{\Theta_1} \int_0^b f_\theta^*(y)g(\theta)dyd\theta = \int_{\Theta_1} F_\theta(b)g(\theta)d\theta. \end{aligned}$$

Hence, for all $b > 0$, we have

$$f_1^*(b) = \int_{\Theta_1} f_\theta^*(b)g(\theta)d\theta. \tag{A.1}$$

Moreover, it is easy to verify (see BBW), that

$$f_0^*(b) = bf_1^*(b) \quad \forall b > 0 \tag{A.2}$$

and that

$$\psi'(b) \equiv \frac{d}{db}\psi(b) = \frac{-f_1^*(b)}{f_0^*(\psi(b))}. \tag{A.3}$$

Now, it follows from (2.10), and (2.11) that for all $\theta \in \Theta_1$, the expression for conditional Type II error in (3.4) is

$$\begin{aligned}\beta(\theta|s) =& P_\theta(B(X) > \psi(1)|S(X) = s) \\ =& \frac{f_\theta^*(\psi(s))\,|\psi'(s)|}{[f_\theta^*(s) + f_\theta^*(\psi(s))\,|\psi'(s)|]}.\end{aligned} \tag{A.4}$$

It is also straightforward to verify that, given H_1 is true, the posterior p.d.f of θ conditional on $S(X) = s$ is

$$g(\theta|s) = \frac{[f_\theta^*(s) + f_\theta^*(\psi(s))\,|\psi'(s)|]g(\theta)}{m_1^*(s)}, \tag{A.5}$$

with

$$\begin{aligned}m_1^*(s) &= \int_{\Theta_1} [f_\theta^*(s) + f_\theta^*(\psi(s))\,|\psi'(s)|]g(\theta)d\theta \\ &= [f_1^*(s) + f_1^*(\psi(s))\,|\psi'(s)|],\end{aligned}$$

where the last equality followed from relation (A.1). By combining (A.4) and (A.5) in (3.4) we obtain that

$$\begin{aligned}E^{g(\theta|s)}[\beta(\theta|s)] &\equiv \int_{\Theta_1} \beta(\theta|s)g(\theta|s)d\theta \\ &= \frac{f_1^*(\psi(s))\,|\psi'(s)|}{[f_1^*(s) + f_1^*(\psi(s))\,|\psi'(s)|]}.\end{aligned} \tag{A.6}$$

Finally, using relations (A.2) and (A.3) in (A.6), it follows that

$$E^{g(\theta|s)}[\beta(\theta|s)] = \frac{1}{[1+\psi(s)]} = \frac{1}{[1+B(x)]} \equiv \beta^*(B),$$

using the fact that $B(x) = \psi(s)$ on the set $\{B(x) > \psi(1)$ and $S(x) = s\}$. □

References

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**, 1787–1807.

Berger, J. O. and Delampady, M. (1987). Testing precise hypotheses. *Statist. Sci.*, **3**, 317-352.

Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.* **91**.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of P-values and evidence. *J. Amer. Statist. Assoc.*, **82**, 112–122.

Berger, J. O. and Wolpert, R. L. (1988). *The Likelihood Principle*, 2nd Ed. Institute of Mathematical Statistics, Hayward, CA.

Bernardo, J. M. (1980). A Bayesian analysis of classical hypothesis testing. In *Bayesian Statistics* (J. Bernardo, et. al., Eds.), 605–647. Valencia University Press.

Brown, L. D. (1978). A contribution to Kiefer's theory of conditional confidence procedures. *The Annals of Statistics*, **6** , 59–71.

Brownie, C. and Keifer, J. (1977). The ideas of conditional confidence in the simplest setting. *Commun. Statist.- Theory Meth.* **A 6(8)**, 691–751.

Casella, G. and Berger, R. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *J. Amer. Statist. Assoc.* **82**, 106–111.

Chatterjee, S. K. and Chattopadhyay, G. (1993). Detailed statistical inference-multiple decision problems. *Calcutta Statist. Assoc. Bulletin* **43**, 155–180.

Chernoff, H. (1972). *Sequential Analysis and Optimal Design.* CBMS Regional Conference Series in Applied Mathematics 8, SIAM, Philadelphia.

Delampady, M. and Berger, J. O. (1990). Lower bounds on Bayes factors for the multinomial distribution, with application to chi-squared tests of fit. *The Annals of Statistics*, **18**, 1295–1316.

Edwards, W., Lindman, H. and Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psych. Rev.*, **70**, 193–242.

Evans, M. (1994). Bayesian inference procedures derived via the concept of relative surprise. Technical Report, Department of Statistics, University of Toronto.

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

Good, I. J. (1992). The Bayesian/Non-Bayesian compromise: a brief review. *J. Amer. Statist. Assoc.* **87**, 597–606.

Hwang, J. T., Casella, G., Robert,C., Wells, M. T., and Farrell, R. (1992). Estimation of accuracy in testing. *Ann. Statist.* **20**, 490–509.

Jeffreys, H. (1961). *Theory of Probability.* London: Oxford University Press.

Kiefer, J. (1975). Conditional confidence approach in multi decision problems. In: P. R. Krishnaiah, Ed. *Multivariate Analysis IV.* New-York: Academic Press.

Kiefer, J. (1976). Admissibility of conditional confidence procedures. *Ann. Math. Statis.* **4**, 836–865.

Kiefer, J. (1977). Conditional confidence statements and confidence estimators. (with discussion), *J. Amer. Statist. Assoc.* **72**, 789–827.

O'Hagan, A. (1995). Fractional Bayes factors for model comparisons. *J. Roy. Statist. Soc. B* **57**, 99–138.

Pappas, T. and Mitchell, C. A. (1985). Effects of seismic stress on the vegetative growth of *Glycine max* (L.) Merr. cv. Wells II. *Plant, Cell and Environment* **8**, 143-148.

Robert, C. P. and Caron, N. (1995). Noninformative Bayesian testing and neutral Bayes factors. Technical Report, CREST, INSEE, Paris.

Schaafsma, W. and van der Meulen, A. E. (1993). Assessing weights of evidence for discussing classical statistical hypotheses. *Statist. and Decision* **11**, 201–220.

Wolpert, R. L. (1995). Testing simply hypotheses. In *Studies in Classification, Data Analysis, and Knowledge Organization*, Vol. 7. (H. H. Bock and W. Polasek, Eds.). Springer-Verlag, Heidelberg, 289–297.