Statistical Science
1997, Vol. 12, No. 3, 133–160

Unified Frequentist and Bayesian Testing of a Precise Hypothesis

J. O. Berger, B. Boukai and Y. Wang

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 in 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 for postexperimental assessment of accuracy. The new unified testing procedure is discussed and illustrated in several common testing situations.

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

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. Use of this traditional (α, β)-frequentist approach in postexperimental inference has been criticized for reporting error probabilities that do not reflect information provided by 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. However, 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 mean-

J. O. Berger is Arts and Sciences Professor, Institute of Statistics and Decision Sciences, Duke University, Durham, North Carolina 27708-0251 (e-mail: berger@stat.duke.edu). B. Boukai is Associate Professor and Y. Wang was a Ph.D. student, Department of Mathematical Sciences, Indiana University–Purdue University, Indianapolis, Indiana 46202 (e-mail: boukai@math.iupui.edu).

ing 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. Kiefer, 1977, Discussion).

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, e.g., 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 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 in postexperimental settings. These are actual 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 numerically equal, 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 $z = \sqrt{n}(\bar{x}_n - \theta_0)/\sigma$:

1. The *classical frequentist test*,

$$T_C\colon \begin{cases} \text{if } |z| \ge z_{\alpha/2}, & \text{reject } H_0 \text{ and report} \\ & \quad \text{error probability } \alpha, \\ \text{if } |z| < z_{\alpha/2}, & \text{accept } H_0 \text{ and report} \\ & \quad \text{error probability } \beta(\theta), \end{cases}$$

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

2. The *P-value test*,

$$T_P\colon \begin{cases} \text{if } |z| \ge z_{\alpha/2}, & \text{reject } H_0 \text{ and report the} \\ & \quad P\text{-value } p = 2(1 - \Phi(|z|)), \\ \text{if } |z| < z_{\alpha/2}, & \text{do not reject } H_0 \text{ and} \\ & \quad \text{report } p; \end{cases}$$

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.

3. A *new conditional test*,

$$T_1^*\colon \begin{cases} \text{if } B(z) \le 1, & \text{reject } H_0 \text{ and report} \\ & \quad \text{error probability} \\ & \quad \alpha^* = B(z)/(1 + B(z)), \\ \text{if } 1 < B(z) < a, & \text{make no decision,} \\ \text{if } B(z) \ge a, & \text{accept } H_0 \text{ and report} \\ & \quad \text{error probability} \\ & \quad \beta^* = 1/(1 + B(z)), \end{cases}$$

where $B(z) = \sqrt{1 + 2n}\exp\{-z^2/(2 + n^{-1})\}$ 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 $n = 20$, $\theta_0 = 0$, $\sigma^2 = 1$, and $\alpha = 0.05$ 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 z. Note that $z_{\alpha/2} = 1.96$ and $a = 3.26$.

The acceptance and rejection regions of all three tests are the same, except that T_1^* makes no decision when $1.18 < |z| < 1.96$. (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 $|z|$. 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 $|z| = 1.96$) to $\alpha^* = 0.0026$ (when $|z| = 4$). The variability in the reports for T_1^* is 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 (for discussion, see the articles mentioned at the beginning of the Introduction). In the rejection region, p does vary with $|z|$, but it is smaller than α^* by a factor of at least 10. Since we will argue that α^* is a sensible

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

Values of $\lvert z\rvert$		T_C	T_P	T_1^*
— — — — — —	0 —		$p = 1$	$\beta^* = 0.135$
		$(\beta(1) = 0.006)$		
Acceptance region	1 —		$p = 0.317$	$\beta^* = 0.203$
	1.18 —			No–decision region
— — — —	1.96 —		– $p = 0.05$ –	— $\alpha^* = 0.496$ —
		$(\alpha = 0.05)$		
Rejection region	3 —		$p = 0.0026$	$\alpha^* = 0.074$
	4 —		$p = 0.0000$	$\alpha^* = 0.0026$

conditional error probability, this discrepancy provides further evidence that P-values can be highly misleading (if interpreted as conditional 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 counterintuitive properties of the test. Indeed, when this is more fully discussed in Section 2.4, it will be observed that, in some settings, 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 $X \in \mathscr{X}$ and wish to test the following "*simple*" hypotheses:

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

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 $\mathscr{B}$ denote the range of $B(x)$, as x varies over $\mathscr{X}$. We will restrict attention here to the case where $\mathscr{B}$ is an interval that contains 1. 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 over the range $\mathscr{B}$ of $B(x)$. 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

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

Corresponding to the test (2.3), the frequentist reports the Type I and Type II error 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 $F_0(c) \equiv 1 - F_1(c)$.

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

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

Thus, for the conditional frequentist, the test (2.3) of these simple hypotheses becomes

$$\begin{array}{ll} \text{if } B(X) \le c, & \text{reject } H_0 \text{ and report conditional} \\ & \quad \text{error probability } \alpha(s), \\ \text{if } B(x) > c, & \text{accept } H_0 \text{ and report conditional} \\ & \quad \text{error probability } \beta(s). \end{array} \tag{2.5}$$

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

$$H_0\colon X \sim e^{-x} \quad \text{versus} \quad H_1\colon X \sim \tfrac{1}{2}e^{-x/2}.$$

Then $B(x) = 2e^{-x/2}$ and its range $\mathscr{B}$ is the interval $(0, 2)$. 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 $\{X\colon S(X) = s\}$ is just a two-point set. Calculation then yields, as the conditional frequentist test (2.5),

$$\text{(2.6)} \quad \begin{array}{ll} \text{if } B(x) \le 1, & \text{reject } H_0 \text{ and report} \\ & \quad \text{conditional error probability} \\ & \quad \alpha(s) = \dfrac{1-s}{2} = \dfrac{B(x)}{2}, \\ \text{if } B(x) > 1, & \text{accept } H_0 \text{ and report} \\ & \quad \text{conditional error probability} \\ & \quad \beta(s) = 0.5. \end{array}$$

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

It has been rare to find suitable ancillary statistics upon which to condition, as in Example 2. (For some other situations in which they have been found, 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

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

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. 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 $\pi_0 = 1/2$ 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

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

and the posterior probability that H_1 is true is

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

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

$$\mathbf{T}_1\colon \begin{cases} \text{if } B(x) \le 1, & \text{reject } H_0 \text{ and report the} \\ & \quad \text{posterior probability} \\ & \quad \alpha^*(B(x)), \\ \text{if } B(x) > 1, & \text{accept } H_0 \text{ and report the} \\ & \quad \text{posterior probability} \\ & \quad \beta^*(B(x)). \end{cases}$$

(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 a modification of $\mathbf{T}_1$ 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 $b \in \mathscr{B}$, let $\psi(b) = F_0^{-1}(1 - F_1(b))$ with $\psi^{-1}(b) \equiv F_1^{-1}(1 - F_0(b))$ and define

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

Consider the test of H_0 versus H_1 given by

$$\mathbf{T}_1^*\colon \begin{cases} \text{if } B(x) \le r, & \text{reject } H_0 \text{ and report the} \\ & \quad \text{conditional error} \\ & \quad \text{probability } \alpha^*(B(x)), \\ \text{if } r < B(x) < a, & \text{make no decision,} \\ \text{if } B(x) \ge a, & \text{accept } H_0 \text{ and report} \\ & \quad \text{the conditional error} \\ & \quad \text{probability } \beta^*(B(x)). \end{cases}$$

The "surprise" observed in BBW (see also Wolpert, 1995) is that $\mathbf{T}_1^*$ 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 $\mathscr{X}^* = \{X\colon 0 \le S(X) \le r\}$. (The complement of $\mathscr{X}^*$ is the no-decision region.) Thus, the conditional frequentist who uses the acceptance and rejection regions in $\mathbf{T}_1^*$, 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 $\beta(s) = \beta^*(B)$. [Using (2.11), it can be seen that, in terms of s, $\alpha(s) = s/(1+s)$ and $\beta(s) = 1/(1+\psi(s))$.]

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 the reported Bayesian error probabilities also be conditional frequentist error probabilities. 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 i.i.d. 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, it follows immediately that, for the test $\mathbf{T}_1^*$ and under certain conditions (cf. Chernoff, 1972, Section 9.1, pages 42–48),

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

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, which was proposed in BBW, is

$$\mathbf{T}_2^*\colon \begin{cases} \text{if } B(x) \le c, & \text{reject } H_0 \text{ and report the conditional error probability } \alpha^*(B(x)), \\ \text{if } B(x) > c, & \text{accept } H_0 \text{ and report the conditional error probability } \beta^*(B(x)); \end{cases}$$

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 between two simple hypotheses, $X \sim m_0(x)$ 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 (this situation is not that considered in the rest of the paper, in which H_1 is composite; we include this discussion here only to indicate that no-decision regions are not unnatural in related contexts):

- Statistician A chooses the hypotheses to be

$$H_0^A\colon X \sim m_0(x) \quad \text{versus} \quad H_1^A\colon X \sim m_1(x),$$

and constructs the size α most powerful test as

$$\begin{array}{ll} \text{if } B(x) \le c_0, & \text{reject } H_0^A, \\ \text{if } B(x) > c_0, & \text{accept } H_0^A, \end{array}$$

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

- Statistician B chooses the hypotheses to be

  $$H_0^B\colon X \sim m_1(x) \quad \text{versus} \quad H_1^B\colon X \sim m_0(x),$$

  and constructs the size α most powerful test as

  $$\begin{aligned} &\text{if } B(x) \ge c_1, && \text{reject } H_0^B, \\ &\text{if } B(x) < c_1, && \text{accept } H_0^B, \end{aligned}$$

  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

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

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, in 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

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

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 $X \in \mathscr{X}$ 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

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

where $\theta_0 \notin \Theta_1 \subset \Theta$. Often we will take Θ_1 to be $\Theta_1 = \{\theta \in \Theta\colon \theta \neq \theta_0\}$. As in Section 2.2, we assume the default prior probability $\pi_0 = 1/2$ for the simple hypothesis $H_0\colon \theta = \theta_0$, while assigning to Θ_1 the prior density $g(\theta)/2$, where g is a proper p.d.f. over Θ_1.

For this case, the Bayes factor in favor of H_0 is exactly as given in (2.2), that is, $B(x) = m_0(x)/m_1(x)$, 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 $H_0\colon X \sim m_0(x)$ versus $H_1\colon X \sim m_1(x)$. 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 $H_0\colon \theta = \theta_0$ against $H_1\colon \theta \in \Theta_1$, the conditional error probabilities arising from (2.4) and from use of the conditioning statistic S in (2.11) would be

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

and

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

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. To many, 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. Averaging is also typically viewed as sensible when there are nuisance parameters.

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 $S = s$. 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 $\bar{X}_n \sim \mathcal{N}(\theta, \sigma^2/n)$, σ^2 known, we wish to test

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

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

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

and

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

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

$$\begin{aligned} B(z) = \sqrt{1+kn}\exp\bigg\{-&\frac{kn}{2(1+kn)} \\ &\cdot\left(z - \frac{\Delta}{\sqrt{kn}}\right)^2 + \frac{\Delta^2}{2}\bigg\}. \end{aligned} \tag{4.4}$$

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

$$\mathbf{T}_1^*\colon \begin{cases} \text{if } B(z) \le 1, & \text{reject } H_0 \text{ and report the CEP} \\ & \quad \alpha^*(z) = \dfrac{B(z)}{B(z)+1}, \\ \text{if } 1 < B(z) < a, & \text{make no decision,} \\ \text{if } B(z) \ge a, & \text{accept } H_0 \text{ and report the CEP} \\ & \quad \beta^*(z) = \dfrac{1}{B(z)+1}. \end{cases} \tag{4.5}$$

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 $z_b^{\pm}$ be the two solutions of the equation $B(z) = b$; it follows from (4.4) that

$$(4.6)\quad z_b^{\pm} = \frac{\Delta}{\sqrt{kn}} \pm \sqrt{\frac{1+kn}{kn}\left(\log\left(\frac{1+kn}{b^2}\right) + \Delta^2\right)}.$$

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

$$(4.7)\qquad \Phi(-z_a^{+}) + \Phi(z_a^{-}) = \Phi(\Delta_k^{+}) - \Phi(\Delta_k^{-}),$$

where $z_a^{\pm}$ is given by (4.6) and

$$\Delta_k^{\pm} = \frac{\Delta\sqrt{1+kn} \pm \sqrt{\log(1+kn) + \Delta^2}}{\sqrt{kn}}.$$

It is clear that $a \equiv a(kn, \Delta)$ depends on Δ and (with a known k) on the sample size n. In Table 2 we present values of a for several choices of Δ and kn. Note also that, for the suggested default choice $k = 2$ and $\Delta = 0$, a closed form approximation to a (accurate to within 1%) was given in Example 1.

ILLUSTRATION 2. Fisher and Van Belle (1993) provide the birth weights in grams of $n = 15$ cases of SIDS (Sudden Infant Death Syndrome) born in King County in 1977:

2,013 3,827 3,090 3,260 4,309
3,374 3,544 2,835 3,487 3,289
3,714 2,240 2,041 3,629 3,345.

With the assumption of normality and a supposed known standard deviation of $\sigma = 800$ g, we consider the test of H_0: $\theta = 3{,}300$ versus H_1: $\theta \neq 3{,}300$. Here 3,300 g is the overall average birth weight in King County in 1977 (which can effectively be considered to be known), so that H_0 would correspond to the (plausible) hypothesis that SIDS is not related to birth weight. 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 $z = 0.485$ and $B(z) = 4.968$, so that $B(z) > a$. Thus, according to $\mathbf{T}_1^*$, we accept H_0 and report the CEP $\beta^* = 0.201$.

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

kn	$\|\Delta\| = 0$	**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

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

$$\mathbf{T}_1^*:\ \begin{cases} \text{if } z \le z_1^- \text{ or } z \ge z_1^+, & \text{reject } H_0 \text{ and report the CEP } \alpha^*(z), \\ \text{if } z_1^- < z < z_a^- \text{ or } z_a^+ < z < z_1^+, & \text{make no decision}, \\ \text{if } z_a^- \le z \le z_a^+, & \text{accept } H_0 \text{ and report the CEP } \beta^*(z). \end{cases}$$

Figure 1 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.

EXAMPLE 4 (One-sided normal testing). We continue with the same basic setup of Example 3, but now we wish to test the hypotheses

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

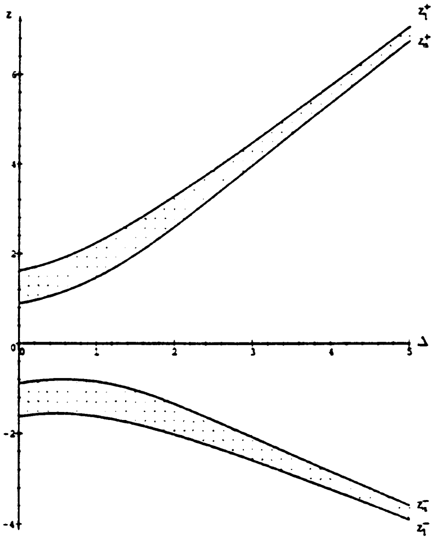

FIG. 1. *The no-decision region of $\mathbf{T}_1^*$ as a function of Δ and with $kn = 10$, for the normal two-sided test of Example 3.*

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

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

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

$$m_1(z) = \frac{2}{\sqrt{1+kn}}\phi\left(\frac{z}{\sqrt{(1+kn)}}\right)\Phi\left(\frac{knz}{\sqrt{1+kn}}\right).$$

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

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

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:

$$B(z_1) = 1, \quad B(z_a) = a,$$

$$1 - \Phi(z_a) = 2\int_{-\infty}^{z_1/\sqrt{1+kn}} \Phi(knz)\phi(z)\,dz.$$

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

$$\mathbf{T}_1^*\colon \begin{cases} \text{if } z \ge z_1, & \text{reject } H_0 \text{ and report the} \\ & \quad \text{CEP } \alpha^*(z), \\ \text{if } z_a < z < z_1, & \text{make no decision}, \\ \text{if } z \le z_a, & \text{accept } H_0 \text{ and report the} \\ & \quad \text{CEP } \beta^*(z). \end{cases}$$

Table 3 presents values of a, z_a and z_1 for selected choices of kn. Note that the no-decision region is somewhat smaller than for the two-sided test.

EXAMPLE 5 (Multisample testing). Consider p independent samples $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{in})$, $i = 1, \dots, p$, of n i.i.d. random variables from the $\mathscr{N}(\mu_i, \sigma^2)$ distribution, with unknown σ^2. We are interested in testing

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

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 $\mu_i, i = 1, \dots, p$, be i.i.d. with a first-stage $\mathscr{N}(0, \xi\sigma^2)$ 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,

TABLE 3
Values of a, z_a and z_1 for the normal one-sided test

kn	a	z_a	z_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	0.440	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
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

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

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 reformulate the test slightly. The difficulties are that (i) H_0 is, itself, composite and (ii) improper prior distributions were used. The most direct solution is initially to suppose that we will base the test on the statistic y in (4.10). We have seen a Bayesian justification for doing so, namely, that the Bayes factor in (4.9) depends only on y; and y is the standard classical test statistic for the problem at hand, arising from, say, the likelihood ratio test.

Write the density of y as $f(y|\theta_1, \dots, \theta_p)$, 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 $\mathscr{N}(0, \xi)$, independently for $i = 1, \dots, p$, while ξ still has proper prior $g(\xi)$. The implied prior

$\pi(\theta_1, \dots, \theta_p)$ is thus proper, and Theorem 1 can be applied. Note that, here,

$$m_0(y) = m(y|0)$$

and

$$m_1(y) = \int_0^\infty m(y|\xi)g(\xi)\,d\xi,$$

where

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

with

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

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

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

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

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

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

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

As an illustration, consider the case with $p = 1$; 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

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

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

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

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

This prior yields, for $p = 1$, the analysis recommended by Jeffreys (1961), since it can be shown that $\pi(\mu|\sigma^2)$ (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)].

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 $B(y)$

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	2.046
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

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 $\lambda \equiv \sum_{i=1}^{p} \mu_i^2/\sigma^2$ 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 $g(\xi) = (1/2)(1+\xi)^{-3/2}$. The resulting prior has the same tail behavior for large λ as the earlier choice, but is positive at zero.

EXAMPLE 6 (ANOVA). We continue with the same basic setup as in Example 5, but now we are interested in testing, with $p > 1$, the composite hypothesis

$$H_0\colon \mu_1 = \mu_2 = \cdots = \mu_p \quad (\text{equal to, say, } \mu) \tag{4.15}$$

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 $\mathscr{N}(\mu, \xi\sigma^2)$ distribution for the i.i.d. $\mu_1, \mu_2, \ldots, \mu_p$; choose, for the second-stage prior, the usual noninformative prior for (μ, σ^2), that is, $\pi_2(\mu, \sigma^2) = (1/\sigma^2)\, d\mu\, d\sigma^2$, which (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

$$\begin{aligned} &m(y|\xi) \\ &\quad = C\frac{y^{(p-3)/2}(1+n\xi)^{p(n-1)/2}}{[p(n-1)(1+n\xi)+(p-1)y]^{(pn-1)/2}}, \end{aligned} \tag{4.16}$$

with

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

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

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

TABLE 5
Values of a for the ANOVA test

n	$p = 2$	$p = 4$	$p = 6$	$p = 8$	$p = 10$
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
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

Now, for any specified prior $g(\xi)$, the test $\mathbf{T}_1^*$ of 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 (4.13). In Table 5 we provide the values of a for selected choices of n and p under the prior (4.14) for $g(\xi)$.

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 (in centimeters) 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 the following:

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

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

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

The value of the Bayes factor, $B(y)$ in (4.17), is $B(y) = 0.001$. 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. The decision whether or not to formulate an inference problem as one of testing a precise null hypothesis centers on assessing the plausibility of such an 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 2. 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 has been the same chemotherapy treatment, but now with (say) steroids added, then equality of treatments would have been a real possibility, since the steroids might 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 the null hypotheses in Illustrations 2 and 3 are both plausible 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. For discussion of this point, see Berger and Delampady (1987).

Choice of the Conditioning Statistic

The first point to stress is the unreasonable nature of the unconditional test (when used for postexperimental assessment of accuracy) and the even more unreasonable nature of the P-value (when incorrectly viewed as an error probability). In a postexperimental sense, the unconditional test is arguably the worst possible frequentist test; for instance, in testing of simple hypotheses, it can be formally established, under many reasonable formulations of postexperimental accuracy, that unconditional frequentist tests are worse than any conditional frequentist tests having the same rejection region. (These results will be reported elsewhere, as will partial generalizations to the type of hypotheses considered in this paper.) Furthermore, it is in some sense true that the more one can condition the better (see also Kiefer, 1977, Discussion and Rejoinder); in this regard, note that the tests we proposed have the maximal degree of conditioning that is possible. Unfortunately, among those tests with a maximal degree of conditioning, there does not appear to be any single optimal choice. (In testing simple hypotheses the situation can be different; see Brown, 1978.) Hence there will be a degree of arbitrariness to the choice of the conditioning statistic, which many may find to be unappealing. It is thus important to keep in mind that the only frequentist alternative to this arbitrariness is to use the unconditional test, which is (often) the uniquely worst test from a postexperimental perspective.

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 2 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 nonancillary statistics will "lose information" but nothing loses as much information as use of unconditional testing in postexperimental inference (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, Bayesians at least should agree that no 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 a default 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 highly attractive 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 judgement 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 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., the ANOVA example). In principle, composite null hypotheses can be treated in the same fashion as composite alternative hypotheses; that is, 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 trade-off 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 et al. (1992), Chatterjee and Chattopadhyay (1993), 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 understand properly; (ii) possessing significantly non-Bayesian or nonfrequentist 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 frequentist error probabilities are quite familiar to many statisticians; likewise the interpretation of Bayesian posterior probabilities is familiar. 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, for a default conditional test that becomes standard, need not be presented in an applied statistical report, any more than one now needs to present all the background properties of the standard 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 introduce formally a no-decision region.

Some Bayesians might be attracted by the long-run frequentist properties of the new tests, in that frequentist properties do not depend on 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 important, 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 adaptation 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:

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

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

$$f_0^*(b) = b f_1^*(b) \quad \forall\, b > 0 \tag{A.2}$$

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 follows 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

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

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

ACKNOWLEDGMENTS

This research was supported in part by NSF Grant DMS-93-03556. We are grateful to the Editor of the article for numerous suggestions that greatly improved the article.

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. *Ann. Statist.* **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** 109–122.

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. IMS, Hayward, CA.

Bernardo, J. M. (1980). A Bayesian analysis of classical hypothesis testing. In *Bayesian Statistics* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) 605–647. Valencia Univ. Press.

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

Brownie, C. and Keifer, J. (1977). The ideas of conditional confidence in the simplest setting. *Comm. Statist. Theory Methods* **6** 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. Bull.* **43** 155–180.

Chernoff, H. (1972). *Sequential Analysis and Optimal Design*. 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. *Ann. Statist.* **18** 1295–1316.

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

Evans, M. (1994). Bayesian inference procedures derived via the concept of relative surprise. Technical report, Dept. Statistics, Univ. Toronto.

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

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*. Oxford Univ. Press.

Kiefer, J. (1975). Conditional confidence approach in multi decision problems. In *Multivariate Analysis IV* (P. R. Krishnaiah, ed.) 143–158. Academic Press, New York.

Kiefer, J. (1976). Admissibility of conditional confidence procedures. *Ann. Math. Statist.* **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. Ser. 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. Decisions* **11** 201–220.

Wolpert, R. L. (1995). Testing simple hypotheses. In *Studies in Classification, Data Analysis, and Knowledge Organization* (H. H. Bock and W. Polasek, eds.) **7** 289–297. Springer, Berlin.

Comment

Dennis V. Lindley

1. INTRODUCTION

My comments embrace the present paper by Berger, Boukai and Wang (hereinafter referred to as BB′W′) and the earlier paper by Berger, Brown and Wolpert (1994, referred to as BBW). That paper dealt with two simple hypotheses but a general prior; the current paper extends the ideas to a composite alternative but restricts the prior. The essential ideas are contained in the simple case, the composite case being reduced to the simple one by a weighted average over the alternatives; but, in order to appreciate what I consider to be a defect, it is necessary to consider the role of the prior. The discussion will therefore center on BBW. When that paper appeared in the *Annals*, I prepared a paper pointing out deficiencies in it and sent copies to the Editor and to each of the authors. The Editor declined to publish it on the grounds that "It has not been customary for the *Annals* to publish commentaries on a previously published article." Even more surprisingly, not one of the three authors attempted to rebut the charges made in my paper. I am therefore most grateful to Paul Switzer for giving me the opportunity to respond to what I consider to be mischievous ideas.

2. SUMMARY

It is convenient to summarize some features of the papers in order more easily to appreciate the objections. A central problem faced by frequentist statisticians is deciding on an appropriate sample space, which usually involves restricting the "obvious" space in some way. (Bayesians restrict it completely, namely, to the single point observed, adopting the likelihood principle.) BBW's key idea is a most ingenious restriction, and we can all agree that the mathematics is beautiful. It is easy to see that the equation

$$F_0(c) = 1 - \rho F_1(c) \tag{1}$$

has a unique solution c, and that, for any $t > c$, there exists a unique $s < c$ satisfying

$$F_0(t) = 1 - \rho F_1(s) \tag{2}$$

and vice-versa. Here ρ is the prior odds in favor of H_1. In BB′W′, $\rho = 1$. Equation (1) appears as the second displayed equation in BBW. Equations (1) and (2), with $\rho = 1$, occur at the beginning of Section 2.3 in BB′W′. Equation (2) establishes a 1–1 correspondence between t and s, with the unique identity value $t = s = c$. BBW partitions X into sets

$$X_s = \{x\colon B(x) = s \text{ or } B(x) = t\}. \tag{3}$$

The authors's decision rule is to say d_1 if $B(x) < c$, to say d_0 if $B(x) \geq c$ and to quote error probabilities

$$p\left(d_1 | H_0 \text{ and } X_s\right) = \frac{B(x)}{[B(x) + \rho]}$$

and

$$p\left(d_0 | H_1 \text{ and } X_s\right) = \frac{\rho}{[B(x) + \rho]},$$

equal to the Bayesian, posterior probabilities. This procedure is subsequently modified to introduce a no-decision element. Notice that the Bayesian would, in a decision situation, choose $c = \ell\rho$, where $\ell = \ell_1/\ell_0$ and ℓ_i is the loss for an incorrect decision when H_i is true. I have six objections to the procedures in these two papers.

3. FALLACY OF THE TRANSPOSED CONDITIONAL

The fallacy consists of confusing $p(A|B)$ with $p(B|A)$, for two events A and B. It is often called the prosecutor's fallacy because of its use by the prosecution in legal cases. If $B = G'$, the event that the defendant is not guilty, and $A = E$, the evidence, for example from DNA testing; then $p(E|G')$ is often, and correctly, very small, say 10^{-6}. The prosecution then commits the fallacy and argues that $p(G'|E)$ is 10^{-6}, so that the defendant is guilty "beyond reasonable doubt." Yet if the DNA evidence is the only evidence against the defendant, reasonably the prior odds on guilt are also very small, say 10^{-6}. Since $p(E|G) = 1$ by Locard's principle, the Bayes factor is 10^6, and by Bayes's theorem, the posterior odds are 1. The posterior probability of innocence, $p(G'|E)$ is only 1/2, quite different from 10^{-6} as the prosecutor claimed.

The new test encourages the fallacy because it quotes an error $p[B(x) = s|H_0 \text{ and } X_s]$ equal to

Dennis V. Lindley is Professor of Statistics, Retired, University College London, and currently resides in Minehead, Somerset, United Kingdom.

$p[H_0|B(x) = s]$ for $s < c$. The latter is equal to the Bayesian's $p(H_0|x)$ since $B(x)$ is sufficient. Indeed, it is equal to $p[H_0|B(x) = s$ and $X_s]$ since X_s provides no further information beyond that provided by $B(x)$, and the fallacy is complete, transposing $B(x) = s$ and H_0, given X_s. Of course, frequentists commit a modified form of the fallacy every time they quote an error rate or a P-value, confusing a probability about some aspect of the data, given the null hypothesis, with the probability of the hypothesis, given the data. So perhaps they will not be concerned with this aspect of the test, although it does provide a grosser form of the fallacy, directly confusing $p(A|B)$ with $p(B|A)$. Notice that BBW arrange that these two probabilities are equal, unlike the legal case, where they can be very different. This equality can only happen if $p(A) = p(B)$. In this case, if $p(H_0|X_s) = p(B(x) = s|X_s)$, a condition that is not apparently reasonable.

4. ANCILLARITY

There is some sense in conditioning on an ancillary partition or statistic because the likelihood is unaffected, but there seems to be little reason to condition on one that is not ancillary because then the statistic contains information (here about H_0 and H_1) which is discarded by the conditioning. It is therefore of interest to investigate when the partition by X_s, equation (3), is ancillary. It is not difficult to show that the condition for ancillarity is that

$$s + \rho = 1 + \frac{\rho}{t}. \tag{4}$$

In particular, this must hold for the identity point $s = t$ which occurs at c. Hence $c = 1$, the other solution to the quadratic being negative, is the condition for ancillarity. When $\rho = 1$, this gives the symmetry condition in BBW. When $\rho \neq 1$, the corresponding condition is not a natural one. Suppose $\rho < 1$, then from (4), s has a minimum value of $1-\rho$ (as $t \to \infty$). Consequently, ancillarity demands that not all values of s, the Bayes factor, are achievable. Similarly, if $\rho > 1$, t has a maximum of $\rho/(\rho-1)$. In particular, ancillarity is not achievable for all ρ.

5. MINIMAX

The decision rule of their test has $d_1(d_0)$ if $B(x) < (\geq) c$. When $\rho = 1$, this is minimax. To see this, note that

$$\alpha(c) = p(d_1|H_0) = p(B(x) < c|H_0) = F_0(c)$$

and

$$\beta(c) = p(d_0|H_1) = p(B(x) \geq c|H_1) = 1 - F_1(c),$$

so that, from (1) with $\rho = 1$, $\alpha(c) = \beta(c)$ which, when $\ell = 1$, is the minimax solution. Yet the minimax procedure is incoherent. This was first demonstrated in 1955 by Savage and Lindley: details are in Lindley (1972). Minimax procedures also violate the principle of the irrelevant alternative. That is, if d_0 is preferred to d_1, then the introduction of a third possibility d_2 can result in d_1 being preferred to d_0, which is arguably ridiculous.

If $\rho \neq 1$, (1) may be written $[1-\alpha(c)] = \rho[1-\beta(c)]$. The same procedure can also result from taking the smaller of the two weighted powers, $\ell_0[1 - \alpha(c)]$ and $\ell_1[1-\beta(c)]$, with $\ell_1/\ell_0 = \rho$, thus making them equal; and then maximizing the common value. This procedure is often called maximin and has the same defects as minimax.

6. NO DECISION

In order to avoid what they see as an unpleasant feature of their test, namely, the conflict between it and the Bayesian approach, BBW introduce a region of values of the Bayes factor within which no decision is taken. The choice "no decision" is effectively a third decision d_2 that is available, in addition to d_0 and d_1. With a choice among three, the Bayesian would not proceed in the way that BBW suggest. The Bayesian would introduce m_i, the loss in selecting the new d_2 when H_i obtains ($i = 0, 1$), in addition to ℓ_i, the loss for an incorrect, positive decision when H_i is true. Presumably, $m_i < \ell_i$, otherwise there is no point in d_2. Having observed $X = x$, the expected losses (not risks) are

$$\ell_1 p(H_1|x), \quad \ell_0 p(H_0|x)$$

and

$$m_0 p(H_0|x) + m_1 p(H_1|x),$$

for d_0, d_1 and d_2, respectively. So d_2 will be selected iff

$$m_0 p(H_0|x) + m_1 p(H_1|x)$$

is less than both

$$\ell_1 p(H_1|x) \qquad \text{and} \qquad \ell_0 p(H_0|x).$$

Recalling that Bayes's theorem says

$$p(H_0|x)/p(H_1|x) = B(x)/\rho,$$

and doing some rearranging of these inequalities, d_2 will be selected iff

$$\frac{m_1\rho}{(\ell_0 - m_0)} < B(x) < \frac{(\ell_1 - m_1)\rho}{m_0}.$$

This interval will be nonempty if

$$\frac{m_0}{\ell_0} + \frac{m_1}{\ell_1} < 1,$$

in which case $\ell_1\rho/\ell_0 = \ell\rho$, the former critical value, will lie within it, as is easily verified by manipulating inequalities. The introduction by BBW of a "no-decision" interval of $B(x)$ does not reconcile their procedure with that of a Bayesian.

7. LOSSES AND BELIEFS

BBW point out that the choices of ℓ and ρ are "just viewed as formalism; their interpretation in terms of losses and priors is not necessary." In that case, how are they to be selected? Consider first the situation with inference, involving only ρ. For a Bayesian, ρ is the prior odds on H_1 and, for a subjectivist, would be determined by the practical meaning of the two hypotheses. In the case of the law with H_1 guilt and H_0 innocence, ρ would be about N^{-1}, where N is the population of the country within which the crime was committed. What considerations are available without such extraneous factors? Frequentists often select $\alpha = 0.05$, or some other small number, leaving β to look after itself. Are we to do something similar here and take $\rho = 1$? If so, why?

There is further difficulty once decisions are included and ℓ has to be given a value. For a Bayesian, the analysis does not depend on both ℓ_i and prior probabilities π_i; only their product matters for decision-making. This is clearly seen when only two decisions are present, since the critical value is $\ell\rho$. With the possibility of no decision, the reader can easily verify that the Bayesian procedure depends only on $\ell_i\pi_i$ and $m_i\pi_i$. In the Bayesian paradigm, losses and beliefs are inevitably intertwined. The general problem of the separation of losses and probabilities has been discussed by Rubin (1987). He concludes, I think correctly, that the separation is not possible. Yet BBW presents a procedure in which they are. ρ determines the association of one part of the partition, s, with the other, t, through (2). Yet $\ell\rho$ determines one end of the interval of no decision. Thus ℓ and ρ could be changed, keeping their product fixed, with the new decision procedure altering but the Bayesian procedure remaining unaffected.

8. INFERENCE AND DECISION

This point is general and applies beyond the immediate topic of these papers. For a Bayesian, inference is the procedure in which probabilities of the unknowns of interest, given the data, are calculated. Decision making additionally introduces a decision space, a utility, or loss, function, and chooses the decision of maximum expected utility, the expectation being calculated using the probabilities supplied by the inference. This distinction is usual in science and technology. "Pure" science makes inferential judgments. Technology applies scientific concepts to make decisions. Descriptions like that of the Bayesian test $\mathbf{T}_1$ at the end of Section 2.2 of BB′W′ are unnecessary hybrids. An inferential test provides the posterior probability of H_0 (and within H_1 if that is composite). Losses and decisions do not enter. When the latter do, some action is contemplated and "reject" and "accept" become acceptable language. The hybrid form perhaps stems from Neyman's view of inductive behavior.

9. DISCUSSION

I agree with BBW that the problem of testing one simple hypothesis against another simple hypothesis is important, despite the fact that it rarely occurs in practice. If you cannot solve the simple, how can you understand the complicated (composite)? And BBW make it clear that there are real problems, even in the simple case, that need to be resolved before we can deal with inference in situations that arise in practice. The Bayesian has an additional reason for considering the simple case because, as BBW point out, the coherent method turns composite against composite into simple against simple by introducing probabilities within each composite hypothesis and taking expectations.

The new procedure is a real advance in our appreciation of the problem but it is arguable that it is defective. It encourages the prosecutor's fallacy in a strong form. It conditions on a statistic that is typically not ancillary. Because of its minimax, or maximin, nature, it is incoherent. In an attempt to match the procedure with Bayesian, coherent methods, it ignores the change in the coherent method necessitated by the inclusion of a third decision. Unlike Bayesian procedures, it separates losses and probabilities. Finally, it fails to make the useful distinction between inference and decision. I argue that their procedures are unsatisfactory and that a unified frequentist and Bayesian method is unsound. The fact is that the frequentist and Bayesian positions are different, both philosophically and operationally. This should be recognized and attempts to reconcile them resisted.

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. *Ann. Statist.* **22** 1787–1807.

LINDLEY, D. V. (1972). *Bayesian Statistics. A Review.* SIAM, Philadelphia.

RUBIN, H. (1987). A weak system of axioms for 'rational' behavior and the non-separability of utility from prior. *Statist. Decisions* **5** 47–58.

Comment

Thomas A. Louis

1. INTRODUCTION

Berger, Boukai and Wang (BBW) engage in innovative gyrations to produce hypothesis tests that attempt to satisfy both Bayesians and frequentists. Unfortunately, as with many but by no means all Bayes–frequentist compromises, their approach hovers between the sides of a deep and wide chasm. Recent computing developments which enable application of the Bayesian formalism to a wide range of complicated and important problems have partially filled the chasm, but I encourage those planning to use the BBW procedure to wear a parachute.

2. CONDITIONING AND THE NDR

BBW create a conditioning statistic designed to line up frequentist and Bayesian statements, but doing so requires a no-decision region (NDR) to avoid "illogical" frequentist conditional statements. The conditioning statistic $S(X)$ [$= \min\{B(X), \psi^{-1}(B(X))\}$], with $B(X)$ the Bayes factor in favor of the null [formula (2.11)] ensures that outside the NDR conditional error probabilities match those produced by full conditioning. Construction of S to mimic Bayes posterior probabilities is a neat trick. Note that $B(X)$ and $S(X)$ depend on default priors and losses, tuning parameters that are best made available to the analyst.

Although $S(X)$ is a function of $B(X)$, frequentists and Bayesians will dislike that the procedure fails to condition on ancillary statistics. Specifically, the distributions of $B(X)$ under H_0 and H_1 (F_0 and F_1 in the text) do not so condition. Thus, $\psi(\cdot)$ does not and the NDR cannot. For example, consider a basic scenario in which a sample size (N) is chosen by a random, ancillary process and then N Gaussian observations are generated. The BBW procedure will not condition on the event $\{N = n\}$. All reasonable statisticians (and other scientists!) will insist on such conditioning.

The BBW procedure is "Bayes" outside of the NDR and so failure to condition on ancillaries is restricted to lack of such conditioning in defining the NDR. Practical impact depends on the probability of a sample falling in the NDR. In Section 2.4, BBW attempt to reduce our unease by showing that if data are iid, under either the null or alternative hypothesis the probability of the NDR decreases exponentially with sample size. Unfortunately, the theorem is only modestly palliative. Under a composite alternative one has two modeling choices. Either the data are exchangeable with distribution m_1 or, if one conditions on a value of θ, the data are iid but do not follow m_1. The theorem needs to be generalized for these situations.

More important, in designed studies sample size is linked to inferential goals. An increase in sample size is associated with some combination reducing error probabilities and detecting more subtle effects (e.g., in Example 3 with $\Delta = 0$, detecting alternatives with smaller k). In situations where sample size is linked to error reduction or detection of more proximal alternatives, pr(NDR) will decrease slowly at best and may remain constant. The NDR will "persist in probability." To see this for a specific case, note that (4.6) and (4.7) depend on "n" and "k" via "nk". For $\Delta = 0$, "k" is proportional to the square of the effect to be detected.

Although the BBW approach produces hypothesis tests with some preposterior properties acceptable to both Bayesians and frequentists, the NDR causes *a priori* and *a posteriori* problems for everyone. BBW defend their NDR as performing the

Thomas A. Louis is Professor, Division of Biostatistics, University of Minnesota School of Public Health, Minneapolis, Minnesota 55455 (e-mail: tom@biostat.umn.edu).

useful service of preventing the analyst from making a decision when evidence is weak. Although a designed study should be powered to achieve inferential goals, I admit that in practice this ideal is seldom attained and many studies deliver insufficient evidence. What is the analyst to do? For either simple versus simple or simple versus composite testing, Bayesians will report the posterior distribution and the posterior consequences of various decisions, possibly coupled with a sensitivity analysis for variations in the prior. Frequentists will "fail to reject" the null hypothesis, but will not "accept" it. If the alternative is composite, experienced analysts will report credible sets or confidence intervals. For Bayesians, these are conditional on observed data and on the null hypothesis being false. More generally, intelligent analysts will communicate some form of strength of evidence.

The NDR causes big trouble for frequentists who want to produce confidence intervals by inverting hypothesis tests. Should the interval be the θ's for which the test "fails to reject" or should it be the complement of the region composed of θ's for which the test "fails to accept"? With an NDR, the two approaches produce different regions.

Also in Section 2.4, in an additional attempt to mollify potential users, BBW show that under *likelihood symmetry* the NDR disappears. The temptation to eliminate the NDR by insisting on likelihood symmetry should be resisted. In the BBW procedure, the NDR is required to avoid probabilities conditional on S that are illogical in that they "go the wrong way." The illogic may signal that the Bayes factor ($B(X)$) is illogical. Neither introducing the NDR nor requiring likelihood symmetry is an effective, general solution to this problem, because there is a scientific message in such situations. For example, in simple versus simple testing, one investigator may choose m_0 as the null distribution and another may choose m_1. This choice says something about their priors and losses. Traditionally, the null hypothesis is the straw hypothesis with a relatively large loss for inappropriate rejection. If two people have different priors or losses, their inferences should not be forced to agree. These disagreements are a fact of scientific life and should not be suppressed. I would rather be a Bayesian.

3. BAYES AND FREQUENTIST CRITERIA

The foregoing leads to the central question: what do intelligent Bayesians and intelligent frequentists want from a testing procedure? Frequentists require the following:

- preposterior Type I error = α;
- complete conditioning on ancillaries;
- maximal power for some alternative or in some other sense.

In addition, frequentists may "welcome the notion of average power" and might even like "average type I error" so long as the usual preposterior properties are in place. With these in place, some conditional statements are fine, but why not have them be completely Bayes?

Bayesians will want the following:

- complete conditioning on observeds;
- good posterior properties;
- good preposterior properties.

"Good" is relative to a prior or class of priors.

4. A PROPOSAL

Many authors have promoted the Bayesian formalism as a procedure-generator irrespective of one's philosophical approach. This aphilosophic approach views the prior as providing tuning parameters, producing what Fisher (1996) calls "stylized Bayes." Carlin and Louis (1996) document the many advantages of using the Bayesian formalism in developing procedures with excellent frequentist properties. Rubin (1984) points out that preposterior evaluations are "Bayesianly justifiable and relevant frequency calculations."

Against this background, I see no need to develop a hybrid test with partial and usually inappropriate conditioning. I prefer to use the Bayesian formalism to produce a test with acceptable frequentist (preposterior) properties. The prior can be used to tune the procedure (for pure frequentists) or to reflect personal opinion (for pure, subjective Bayesians) or to capture prior empirical evidence (for objective Bayesians). Frequentists can ignore all posterior statements, communicate all of them or operate in a middle ground, depending on their degree of purity. We know what Bayesians will do.

For a specific case, consider BBW's Example 3 with $\theta_0 = 0$ and $\Delta = 0$. Allow a general prior probability π on H_0 and a general value of k. Assume we want a formal Bayes test that rejects H_0 when the posterior probability of it is less than α, that is,

$$B(x) < \frac{1-\pi}{\pi}\frac{\alpha}{1-\alpha}.$$

This Bayes test rejects H_0 when

$$Z^2 > d(u, c) = \frac{1+u}{u}[\log(1+u) - 2c],$$

where

$$c = \log\left(\frac{1-\pi}{\pi}\frac{\alpha}{1-\alpha}\right).$$

Note that $\lim_{u\to\infty} d(u, c) = \infty$, and

$$\lim_{u\to 0} d(u, c) = \begin{cases} 1, & \text{if } c = 0, \\ \infty, & \text{if } c < 0. \end{cases}$$

For this test to have frequentist size α, $d(u, c)$ must equal $z_{\alpha/2}$. To accomplish this we can adjust u (i.e., for a fixed n, adjust k) and c (by adjusting π). Although there may not be a solution for $\pi = 0.5$, in general there will be multiple combinations of k and π that produce a desired frequentist cut-point. A specific pair can be picked to maximize frequentist power for some parameter value in the alternative hypothesis space, to maximize the average power or to incorporate Bayesian considerations.

The foregoing procedure has standard frequentist, preposterior properties. The data analyst can report the posterior probability of H_0 or not, depending on how "Bayesian" he or she wants to be. The procedure conditions on ancillaries and avoids the NDR with its associated embarrassments. Is this not better than the BBW approach? Bayesians give up nothing. Frequentists who like the BBW procedure are bound to prefer this one.

5. CONCLUSION

Much of my critique relates to the role of hypothesis testing in scientific inference. A scientist requires information on strength of evidence even if hypothesis testing decisions are to be made. This information is a natural output of the Bayesian formalism; classical standard errors and confidence intervals also do a reasonable job. The BBW approach will do a good job only some of the time. Although the conditional error probabilities are attractive, failure of the BBW procedure to condition on ancillaries and the possibility of a sample falling into its NDR will block its use. To evaluate this prediction, I ask BBW to let us know if they would use their procedure in real-life settings such as providing court testimony as an expert witness or in serving on a data and safety monitoring board for a clinical trial.

Although BBW advertize their procedure as automatic, it does depend on default settings for prior parameters. If hiding these settings from the user is necessary to recruit frequentists, the same strategy can be employed for a fully Bayesian procedure. However, I do not approve of this strategy. It is deceptive and eliminates an opportunity to tune the procedure. Fortunately, Bayes procedures are now computable for complicated, relevant statistical models. These procedures can be tuned using priors and loss functions to have desired preposterior properties, whether Bayes or frequentist. The user can decide on what level of Bayesian reporting to include in a statistical summary.

As a scientist who must convince a broad group of stake holders (a broad group of priors and loss functions), I will want a procedure with good frequentist properties and I will report these. As one who is allowed to tune the procedure, I will want to report posterior summaries such as pr[H_0 | data], possibly accompanied by a sensitivity analysis. I strongly prefer this method of integrating Bayesian and frequentist approaches, but this preference in no way diminishes my respect for BBW's creativity and their ability to stimulate discussion of fundamental issues in statistical science.

ACKNOWLEDGMENT

This work was supported by Grant NIH/NIEHS: 1R01-ES-07750.

REFERENCES

CARLIN, B. P. and LOUIS, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall, London.

FISHER, L. D. (1996). Comments on Bayesian and frequentist analysis and interpretation of clinical trials. *Controlled Clinical Trials* **17** 423–434.

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172.

Comment

David Hinkley

Berger, Boukai and Wang have presented an interesting method for constructing genuinely useful additions to our statistical toolkit. Their proposal has the further merit that it should be the source of much fruitful argument among frequentist and Bayesian statisticians, both theoreticians and users of statistics.

Conventional significance tests have the shortcoming that they do not directly answer the natural question "Given these data, what is the probability that the null hypothesis is true?" Partly for this reason, and partly because hypothesis testing is often less relevant than estimation of effects, widespread use of significance tests is subject to intense criticism in some quarters. One example of this is the recent provocative article by Nester (1996). Such criticism is not limited to individuals: one newsletter that I read in 1996 reported that "...the APA [American Psychological Association] Board of Scientific Affairs at its November meeting approved in principle the creation of a Task Force to make recommendations about possibly discontinuing the use of statistical significance testing." Some of the history behind this is summarized in the interesting article by Cohen (1994). How would such a task force react to the present paper, one wonders?

What a significance test does is convert a test statistic [which measures discrepancy between data and null hypothesis (NH)] from its special scale to a p-value on the standard scale of probability. The smaller the p-value, the stronger the evidence against NH. Unfortunately, p-values are not generally comparable from one experiment to another, unless the information contents of the experiments are the same. That is, there is no universal inferential scale according to which p-values can be judged. This point has been noted by several distinguished frequentist theoreticians and underlies the advocacy of conditioning on ancillary statistics. However, no general frequentist methodology has been developed to deal with the problem.

This is of no great consequence for one-sided hypotheses about parameters, at least for likelihood-based tests, where p-values are moderately robust approximations to posterior probabilities. The difficulty can be avoided if tests are replaced by confidence interval summaries for parameters, as is usually more helpful scientifically. Moreover, if the analysis is likelihood-based, then confidence intervals will often be approximately posterior probability intervals, under quasiobjective prior distributions such as the "reference priors" developed by Bernardo, Berger and others.

So for a wide range of problems the question of interest can be answered with an appropriate assessment of uncertainty, which is approximately the same whether or not the tools used are Bayesian. This leaves the thorny problem that Berger and his coauthors address, that of the precise or sharp hypothesis. Some would argue that such hypotheses can never be exactly true (Nester, 1996), but they are often sufficiently representative to be of genuine interest. Such hypotheses also arise, at least implicitly, in the context of selection of a predictive model among a hierarchy of models—where use of conventional significance tests tends to produce poor results, compared to Bayes-like methods such as BIC (Ripley, 1996).

The authors aim to go further than the conventional significance test, and provide a sensible error rate to go with a decision for or against the null hypothesis. Given that their objective is to draw a decisive conclusion if possible, the option of "no decision" may be viewed as a sensible inclusion in the methodology. As noted earlier, the usual p-value cannot be interpreted fully without reference to the amount of relevant information available, so the authors's conditional error rate is a positive step forward. One might wish that the conditioning statistic had a more familiar feel to it, however.

One difficulty that I foresee is dealing with problems where the alternative hypothesis is quite vague, such as goodness-of-fit problems. Presumably the methodology is restricted to likelihood analysis, possibly with some appropriate extension to cover nonparametric problems? Many of the conventional goodness-of-fit tests are intermediate and informal steps in data modelling, and there is not one specific alternative hypothesis. An example is the simulation envelope test for a normal Q–Q

David Hinkley is Professor, Department of Statistics and Applied Probability, University of California at Santa Barbara, Santa Barbara, California (e-mail: hinkley@pstat.ucsb.edu).

plot; see, for example, Atkinson (1985, Section 4.2) or Davison and Hinkley (1997, Section 4.2.4). Nevertheless it would be interesting to know what the authors would do in such situations. Would they, for example, try to set up a broad range (possibly unlimited) of alternative hypotheses, and how would this affect their methodology? As another example, suppose that we have a long binary sequence and want to test the null hypothesis of homogeneity and independence: how would the authors approach this?

REFERENCES

ATKINSON, A. C. (1985). *Plots, Transformations and Regression*. Oxford Univ. Press.

COHEN, R. (1994). The Earth is round ($p < .05$). *American Psychologist* **49** 997–1003.

DAVISON, A. C. and HINKLEY, D. V. (1997). *The Bootstrap and Its Application*. Cambridge Univ. Press.

NESTER, M. R. (1996). An applied statistician's creed. *J. Roy. Statist. Soc. Ser. C* **45** 401–410.

RIPLEY, B. D. (1996). *Pattern Recognition and Classification*. Cambridge Univ. Press.

Rejoinder

J. O. Berger, B. Boukai and Y. Wang

We thank the discussants for their stimulating comments and for providing a variety of perspectives on the issues. In our Rejoinder, we will group responses by subject, rather than by discussant. Since Professor Lindley was also discussing Berger, Brown and Wolpert (1994), some specific additional comments from Lawrence Brown and Robert Wolpert are included below. Finally, as we agree with essentially everything that Professor Hinkley wrote, our comments will tend to focus on the discussions of Professors Lindley and Louis.

UNIFICATION OF STATISTICS

Professor Lindley feels that "... frequentist and Bayesian positions are different, both philosophically and operationally. This should be recognized and attempts to reconcile them resisted." While indeed they are philosophically and operationally different, we would argue that they should not be yielding fundamentally different answers in practice. Not only is it unfortunate from the perspective of the field to have one group of statisticians saying answer A is correct while the other asserts answer B, both based on the same evidence and beliefs, but this can be tragic for the applications; either the drug is effective or it is not. Furthermore, we would argue that such basic disagreement is typically the result of use of an overly limited or inadequate version of either frequentist or Bayesian methodology. Hence efforts at "unification" have the very real effect of improving statistical practice, as well as enhancing the image or our profession and its impact.

Testing of a precise hypothesis has been one of the areas in which fundamental disagreement in practical conclusions has been the rule. The point of the article is to observe that this need not be so, and that classical statisticians should, by frequentist reasoning, reach essentially the same conclusions as do Bayesians. There is no need to embrace the Bayesian philosophy for our profession to reach essential agreement concerning what to do with these testing problems.

We often treat statistical paradigms other than "our own" by attempting to limit their options and then showing that this limited version is inadequate. If one limits frequentism to unconditional (preexperimental) evaluations or to limited forms of conditioning (e.g., conditioning only on ancillary statistics), then it is easy to show that the paradigm is inadequate. Frequentists must be allowed the option of reporting general conditional (postexperimental) evidence.

Professor Louis proposes a different "unification" of Bayesian and frequentist statistics, based on allowing multiple types of reports and upon possibly matching the unconditional frequentist answer with Bayesian posterior answers. Considering this latter aspect first, let us look at actual numbers. Suppose that, in the situation of Example 3 (testing a normal mean), a frequentist wants to use $\alpha = 0.05$. Then, if $n = 20$ and the reasonable prior variance multiplier $k = 2$ is chosen, solving Louis's equation for the prior probability of H_0 yields $\pi = 0.051$. So a Bayesian can report the same number (a posterior probability of 0.05) as the unconditional frequentist's α, but only by a priori being virtually convinced that H_0 is true at the desired level! It would obviously be very questionable for the Bayesian to simply make the unconditional frequentist conclusion of significant evidence against the null hypothesis, essentially "hiding" the fact that the a priori odds were stacked 19 to 1 against H_0. Unconditional and conditional numbers are simply very different entities, and attempting to "match" them makes little sense. (In contrast, attempting to "match" conditional frequentist and conditional Bayesian numbers is, at least, plausible.) Perhaps Louis meant for the statistician in this situation to report that "if the prior probability of H_0 were 0.051, then, after observing $z = 1.96$, the posterior probability of H_0 would be 0.05, that is, equal to α." But this would be a very convoluted way of communicating the same information as that provided by the conditional error probability in Table 1. Of course, one might also want to consider alternative values of the prior variance multiplier k; indeed Bayesians would urge that this be done, with subjectively chosen ranges of k being considered. However, this additional complication is unlikely to appeal to frequentists, especially as the message does not change much. For instance, even the rather bizarre choice $k = 0.142$ yields $\pi = 0.10$, and it can be shown that this is the *largest* obtainable value of π (when $\alpha = 0.5$ and $n = 20$).

The basic fact here is that, when $z = 1.96$ in the situation of Example 1 (or Example 3), the data is roughly equally supportive of H_0 and H_1, and all statisticians must find a way to report this. Such a report occurs naturally from the Bayesian perspective and our paper shows that frequentists can also report this fact if they adopt the conditional frequentist perspective. There is, of course, nothing wrong with $\alpha = 0.05$ as a preexperimental measure of the quality of the experiment but, postexperimentally, it is simply not a scientifically tenable report from either a Bayesian or a (conditional) frequentist perspective; we thus strongly disagree with Louis's statement that "Frequentists can ignore all posterior statements, communicate them all or operate in a middle ground, depending on their degree of purity," or at least we would strongly disagree if the phrase "posterior statements" was replaced by "conditional or postexperimental statements." In this regard, Louis also makes the curious statement concerning frequentists that "... so long as the usual preposterior properties are in place ... some conditional statements are fine, but why not have them be completely Bayes?" Obviously the unconditional properties of any statistical procedure can be calculated and reported, but doing so "on the side" does not make one a frequentist. When $z = 1.96$, one can report both $\alpha = 0.05$ and that the posterior probability of H_0 is 0.496 but, if one's scientific conclusion is based on the posterior probability, the mere reporting of α does not make one a frequentist. (Of course, we have shown that 0.496 is also a long-run conditional frequentist error rate, and hence can be used by a frequentist.) We are not suggesting that a good statistician is wrong to mix frequentist and Bayesian statements; we are simply saying that one cannot arbitrarily label such mixed inferences as "frequentist," and then say that the problem is solved.

As an aside, it is worth mentioning that unconditional frequentists might argue that, for a fixed α, observations near the rejection boundary are unlikely to occur, and hence the postexperimental difficulty mentioned above is rare. This ignores two practical realities. The first is the ubiquitous use of P-values, instead of fixed α levels; P-values virtually always greatly overstate the evidence in testing of precise hypotheses. The second reality is that optional stopping is all too often used, but not reported, so that we "happen to see" data near the rejection boundary far more often than we actually should.

THE NO-DECISION REGION

Professors Lindley and Louis express concerns involving the no-decision region. Lindley points out that a Bayesian would treat "no decision" as a third possible action and would ideally introduce associated losses to deal appropriately with the three decisions. This is certainly true, and it is indeed unlikely that the ensuing procedure would exactly match the new procedure. However, inferential Bayesians (as opposed to decision-theoretic Bayesians) would probably not find our no-decision region objectionable, in that it seems to coincide in practice with data which is evidentially quite weak. In regards to Lindley's point here, Lawrence Brown adds the following comment: "I think of the no-decision region as an 'embarassing decision' region. It is a region where the conditional frequentist differs in conclusion from the Bayesian, and where it may be argued that the conditional frequentist reasoning somewhat breaks down. I am inclined to agree with Lindley that viewing this as a third decision region can, formally, lead to trouble. Fortunately, this does not occur very often."

Professor Louis's concern with the rate at which the probability, under a composite alternative, of the no-decision region approaches zero is appreciated. (Our comment in the paper was misleading in this regard since, as Louis observes, we do not really have i.i.d. observations from m_1.) Indeed, the rate will typically not be exponential for the composite alternative situation; for instance, we recently established that the rate is actually $O(n^{-1/2} \log n)$ for the situation of Example 3. (General results concerning rates for the composite alternative case are being pursued by one of us.)

Professor Louis goes on to question whether the no-decision region is really "small" in practice. Perhaps the most crucial point here is that the no-decision region virtually never intersects the rejection region (for the testing of a precise null hypothesis). Hence all we are discussing is the size of the no-decision region when considering whether to formally accept a null hypothesis or to say "no decision." Since the new testing procedure is arguably already better than classical practice in allowing for quantified acceptance of the null hypothesis, the concerns of Louis in this regard would seem to be obviated.

CONDITIONING AND ANCILLARITY

All discussants mention, with varying degrees of emphasis, that the lack of ancillarity of the conditioning statistic that we use will be a cause for concern among classical statisticians. As emphasized by Kiefer (1977), however, there is no reason, within frequentist theory, to restrict conditioning to ancillary statistics. Indeed, another interpretation of our paper is that it clarifies a situation in which frequentists apparently need to proceed beyond conditioning on ancillary statistics in order to achieve sensible answers. (This is why we so heavily stressed in the paper the unsuitability of the unconditional classical approaches to testing of precise hypotheses.) The commonly stated intuitive reasoning behind restricting conditioning to ancillary statistics, and our view as to why this reasoning is faulty, was discussed in Section 5 of the paper. We repeat only the comment that it is puzzling to see Bayesians object to the conditioning in the paper, since we show the result to be essentially equivalent to full Bayesian conditioning. Arguments aside, however, we quite agree with Professor Hinkley's "One might wish that the conditioning statistic had a more familiar feel to it." All we can say is that familiarity increases with use and, sometime down the road, it will likely feel completely natural to condition on this type of statistic.

In regards to conditioning, Professor Louis raises the interesting point that the new procedure is not guaranteed to condition on, say, an ancillary stopping time. While true, the situation is roughly that a mole hill is left behind after a mountain has been removed. The "mountain" that obstructs classical statistics in this regard is the fact that unconditional testing is highly dependent on the stopping rule used, leading, for instance, to extremely complicated procedures in sequential testing. With the new procedure, as discussed in Berger, Brown and Wolpert (1994), the only dependence on the stopping rule (and the only possible dependence on an "ancillary" stopping time), arises in determination of the no-decision region. However, the no-decision region is rarely an issue in applications, as mentioned above, so that any stopping rules (and not just those that are ancillary) become irrelevant in applications. This is of enormous practical benefit. Note that we are not objecting to the principle that one should condition on an ancillary stopping time; we are simply arguing that formal principles are often violated in minor ways to achieve major ends.

On the issue of conditioning, Robert Wolpert adds: "While the argument for conditioning is perhaps strongest when the conditioning statistic is ancillary, we should be willing to discard the modicum of information contained in a 'nearly ancillary' statistic in exchange for freedom from the dangers of misinterpretation and dependence on the stopping rule that plague the classical test."

BAYESIAN CONCERNS

Professors Lindley and Louis implicitly suggest that there is not much here for Bayesian practice. Louis even provocatively asks if we would use these procedures in actual practice. Bayesians may well prefer to continue using their existing methodology, and we have no quarrel with that. However, in practice, we frequently encounter situations in which a full Bayesian analysis is not tenable, for a variety of reasons, and we are certainly delighted then to have available a method which yields essentially the same answers but can be justified from a frequentist perspective. Second, at least some Bayesians do take comfort in knowing that their procedures have a frequentist interpretation. Finally, echoing Professor Hinkley, those who seek to understand the debates on foundations of statistics (and this includes many Bayesians) will need to adapt to the possibilities inherent in conditional frequentist analysis.

Professor Lindley takes us to task for encouraging the confusion between $P(A \mid B)$ and $P(B \mid A)$. Indeed, he has a point. Previously, one of us taught elementary testing by discussing both the classical α-level and the posterior probability of the null hypothesis, and was successful in communicating the difference between $P(A \mid B)$ and $P(B \mid A)$ because these numbers (and the resulting conclusions) were so different. However, with the new procedure, these two probabilities will be equal, so that testing of a precise null hypothesis will no longer serve as a good pedagogical example of the "prosecutor's fallacy." In a related vein, a Bayesian might dislike the new procedure because it eliminates one of the biggest contrasts between Bayesian and classical methods, and hence eliminates one of the most powerful rationales for the Bayesian position. Our guess, however, is that non-Bayesians who take the time to truly understand the issues here will end up with considerably increased sympathy for the Bayesian position.

Robert Wolpert's view of this issue is: "The intention behind the development was precisely to find a statistical testing procedure which yields the same error probabilities for frequentists who condition on the hypothesis or Bayesians who condition on the data. We do not confuse the two types of error probabilities, but sought a test which is safe to use even for those who *do* confuse them, especially the large proportion of nonstatisticians who routinely misinterpret P-values as posterior probabilities."

Professor Lindley's comments concerning minimaxity are rather curious. First of all, we did not actually recommend using the minimax rule here, since that would involve making unreasonable conclusions for certain data (corresponding to the no-decision region). Also, general criticisms about minimax procedures should not be used to indict specific minimax procedures. After all, there are numerous Bayes rules with respect to proper priors which also happen to be minimax, and we doubt if Lindley would insist that any such proper priors be barred from consideration by a subjectivist! That said, we agree with Lindley's underlying point, which is that the procedure we recommend can probably be shown to be formally incoherent. Of course, most Bayesians (as well as non-Bayesians) typically operate in practice in ways that are formally incoherent; the key question is whether the incoherence is significant or minor, and our judgement is that any incoherence found here would be of the minor variety.

LOSSES AND PRIORS

Professor Lindley, in discussion more related to Berger, Brown and Wolpert (1994), makes several observations concerning the fact that the new testing procedure can be modified to allow for varying prior probabilities of the hypotheses and varying losses for incorrect decisions. He first asks how prior probabilities and losses are to be chosen, if not in subjectivist fashion? We would agree that subjectivism is needed for their choice, but note that we are primarily advocating the new testing method for use in "default" or "inferential" fashion; hence our restriction in this paper to (essentially) the assumption of equal prior probabilities of hypotheses and equal losses in incorrect decisions. Lindley also notes that the new testing method does not depend only on the product of prior and loss, as Bayesian procedures should. Again, however, this "slight incoherency" only manifests itself in the no-decision region, not in reported expected losses. Lindley later argues for keeping inference distinct from decision, which is what we are trying to do in the present paper. We would not rule out, however, the possibility of successful development of conditional frequentist decision theory along the lines suggested by Berger, Brown and Wolpert (1994).

Professor Louis notes that, in classical testing, the choice of hypotheses is usually based on priors and losses of the experimenter, and hence contains a message. This is certainly true, but is it desirable? We do not feel that "hiding" losses and priors through such choices is a desirable feature of classical statistics. We suspect that Louis would agree with this; indeed he lauds the Bayesian approach as allowing for explicit study of sensitivity to priors and losses, and we would agree that this is a big advantage. Likewise, Louis suggests that the default priors and losses used in the new procedures be made available to the analyst; we would certainly not disagree.

GENERALIZATIONS

Professor Hinkley concludes by asking about extensions to situations where the alternative is vague or nonparametric. To see that one must remain very cautious about unconditional methods in such situations, see Delampady and Berger (1990). However, admittedly, extending the conditional frequentist approach to such problems may be quite challenging. Indeed, exploratory analysis, when alternatives are vague, may well remain mostly an art.

Professor Louis raises the interesting question of how a frequentist should produce a confidence set for θ after rejecting H_0. We first note that this is a problem common to all frequentist analysis: the conclusion from one part of the analysis can formally affect the conclusion from another. In practice, this issue is usually ignored, with the "standard" confidence set being reported upon rejection of H_0. When the "standard" confidence set is satisfactory conditionally (in the sense of approximately corresponding to posterior probability intervals, as Professor Hinkley notes), we do not view the situation as one of great concern. Incidentally, we much prefer constructing frequentist confidence sets by using "probability matching" posterior probability intervals, rather than by inverting tests. The optimality properties inherited through the "inverting" process are not very compelling; indeed, the resulting confidence sets can have very poor conditional behavior.

CONCLUDING REMARK

There is a certain irony to this discussion: although the disagreements expressed herein might seem rather severe, we suspect that the testing methods the discussants and ourselves would actually *prefer* to use in practice are similar, with heavy emphasis on Bayesian analysis with sensitivity studies. Indeed, our view of the discussions from this perspective is that they were quite wonderful, providing very good advice as to how (philosophy aside) statistical testing should be done. However, especially for non-Bayesians or Bayesians operating in non-Bayesian environments, we agree with Professor Hinkley that the new conditional testing methods are "genuinely useful additions to our statistical toolkit."