The Annals of Statistics
1994, Vol. 22, No. 4, 1787–1807

A UNIFIED CONDITIONAL FREQUENTIST AND BAYESIAN TEST FOR FIXED AND SEQUENTIAL SIMPLE HYPOTHESIS TESTING[1]

BY JAMES O. BERGER, LAWRENCE D. BROWN AND ROBERT L. WOLPERT

Purdue University, University of Pennsylvania and Duke University

Preexperimental frequentist error probabilities are arguably inadequate, as summaries of evidence from data, in many hypothesis-testing settings. The *conditional frequentist* may respond to this by identifying certain subsets of the outcome space and reporting a *conditional* error probability, given the subset of the outcome space in which the observed data lie. Statistical methods consistent with the *likelihood principle*, including Bayesian methods, avoid the problem by a more extreme form of conditioning.

In this paper we prove that the conditional frequentist's method can be made *exactly* equivalent to the Bayesian's in simple versus simple hypothesis testing: specifically, we find a conditioning strategy for which the conditional frequentist's reported conditional error probabilities are the same as the Bayesian's posterior probabilities of error. A conditional frequentist who uses such a strategy can exploit other features of the Bayesian approach—for example, the validity of sequential hypothesis tests (including versions of the *sequential probability ratio test*, or SPRT) even if the stopping rule is incompletely specified.

1. Introduction.

1.1. *The basic conditional test.* We observe the random variable $X \in \mathcal{X}$ and wish to test the simple hypotheses

$$H_0\text{: } X \text{ has density } f_0(x) \quad \text{versus} \quad H_1\text{: } X \text{ has density } f_1(x).$$

Denote the *likelihood ratio of H_0 to H_1* or *Bayes factor in favor of H_0* by

$$B(x) = f_0(x)/f_1(x).$$

[Note that *small* values of $B(x)$ correspond to rejection of H_0.] Let F_0 and F_1 be the c.d.f.'s of $B(X)$ under f_0 and f_1, respectively. For ease of exposition, we will assume that F_0 and F_1 are continuous and invertible, and denote their inverses by F_0^{-1} and F_1^{-1}.

The standard equal-tailed (minimax) most powerful test is determined by the critical value c, which satisfies

$$F_0(c) = 1 - F_1(c).$$

Received June 1993; revised February 1994.

[1]Research supported by NSF Grants DMS-8923071, DMS-8903842, DMS-9303556 and SES-8921227. The project was initiated at the Cornell Workshop on Conditional Inference, sponsored by the Army Mathematical Sciences Institute and the Statistics Center, June 3–14, 1991.

AMS 1991 *subject classifications*. Primary 62A20; secondary 62A15.

Key words and phrases. Likelihood principle, conditional frequentist, Bayes factor, likelihood ratio, significance, Type I error, Bayesian statistics, stopping rule principle.

This test rejects H_0 if $B(x) \leq c$, accepts H_0 if $B(x) > c$ and has error probabilities of Type I and Type II given by $\alpha = \beta = F_0(c) = 1 - F_1(c)$.

Consider, instead, the testing procedure T_1 defined as follows:

if $B(x) \leq c$, reject H_0 and report the conditional error probability $\alpha(B) = B/(1+B)$;

if $B(x) \geq c$, accept H_0 and report the conditional error probability $\beta(B) = 1/(1+B)$.

This simple procedure has a number of attractive properties. First, $\alpha(B)$ and $\beta(B)$ are the posterior probabilities of H_0 and H_1, respectively (assuming equal prior probabilities), so the reported error probabilities are those that a Bayesian would report. Second, T_1 is a test often considered by likelihoodists. Birnbaum (1961) called $\alpha(B)$ and $\beta(B)$ the "intrinsic significance levels." Third, and quite surprisingly, T_1 is also a valid test thanuentist procedure, more precisely a valid conditional frequentist procedure. (See Section 2 for definitions and justifications.) The situation here thus differs from that in classical testing of simple hypotheses where the unconditional error probabilities α and β are typically very different from the Bayesian (or likelihood) $\alpha(B)$ and $\beta(B)$. Indeed, this is one of the common criticisms of classical testing of simple hypotheses: the reported error probabilities do not distinguish between data at the boundary of the rejection or acceptance regions and data deep within the region.

EXAMPLE 1. Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. $\mathcal{N}(\theta, 1)$ and that it is desired to test H_0: $\theta = -1$ versus H_1: $\theta = 1$. Then

$$B = \prod_{i=1}^{n} \frac{(2\pi)^{-1/2} \exp\{-\frac{1}{2}(x_i + 1)^2\}}{(2\pi)^{-1/2} \exp\{-\frac{1}{2}(x_i - 1)^2\}} = \exp\{-2n\bar{x}\},$$

and T_1 becomes:

if $B \leq 1$ (i.e., $\bar{x} \geq 0$), reject H_0 and report error probability $\alpha(B) = (1 + \exp\{2n\,\bar{x}\})^{-1}$;

if $B > 1$ (i.e., $\bar{x} < 0$), accept H_0 and report error probability $\beta(B) = (1 + \exp\{-2n\,\bar{x}\})^{-1}$.

For comparison, the classical Neyman–Pearson test with equal error probabilities utilizes the same rejection and acceptance regions, but reports the error probabilities (of Type I and Type II) $\alpha = \beta = 1 - \Phi(\sqrt{n})$, where Φ is the standard normal c.d.f. When $n = 4$, a comparison of the reports is given in Table 1. For completeness, we also report the P-value against H_0, which here is $1 - \Phi(2(\bar{x} + 1))$.

The intuitive attractiveness of $\alpha(B)$ and $\beta(B)$ is clear. If the data are $\bar{x} = 0$, intuition suggests that the evidence equally supports H_0: $\theta = -1$ and H_1: $\theta = 1$; $\alpha(B)$ and $\beta(B)$ so indicate, while α and β (and the P-value) do not. When $\bar{x} = 1$, in contrast, intuition would suggest overwhelming evidence for H_1 (note that $\bar{x} = 1$ is four standard deviations from $\theta = -1$); again $\alpha(B)$ and $\beta(B)$ reflect this. (We delay further discussion of the P-value to the end of Section 1.4.)

TABLE 1
Comparison of error probabilities when $n = 4$

	For Test T_1		For $N-P$ Test		
$\bar{x}$	$\alpha(B)$	$\beta(B)$	α	β	P-Value
0	0.5	0.5	0.025	0.025	0.025
1/4	0.12	0.88	0.025	0.025	0.0062
1	0.00034	0.99966	0.025	0.025	0.00003

The above succinctly summarizes the major point of the paper: in testing simple hypotheses the counterintuitive $N-P$ test can be replaced by the much more attractive T_1, with complete Bayesian, likelihood and frequentist justification. (Indeed, it will be argued in Section 3 that T_1 is actually a better frequentist test than the $N-P$ test.)

1.2. *A Bayesianly motivated modification.* To a Bayesian, T_1 will appear to be somewhat unnatural when $c \neq 1$. If, say, $c = 3$ and $B(x) = 2$, then T_1 would be "reject H_0 and report conditional error probability $\alpha(B) = \frac{2}{3}$." The obvious question is "why should H_0 be rejected if that action has a $\frac{2}{3}$ chance of being wrong?"

This might, of course, be a perfectly rational thing to do if the losses in accepting and rejecting are asymmetric. Indeed, in Section 2, T_1 will be seen to be a true Bayes test for certain losses. For inference without a specified loss, however, this behavior of T_1 is unappealing and may seem silly to practitioners.

A useful modification of T_1, that alleviates this difficulty, is to incorporate a "no decision" region. To describe this modification, T_1^*, define

$$r = 1 \text{ and } a = F_0^{-1}\big(1 - F_1(1)\big) \quad \text{if } F_0(1) \leq 1 - F_1(1),$$
$$r = F_1^{-1}\big(1 - F_0(1)\big) \text{ and } a = 1 \quad \text{if } F_0(1) > 1 - F_1(1).$$

Then:

if $B(x) \leq r$, reject H_0 and report the conditional error probability $\alpha(B) = B/(1+B)$;

if $r < B(x) < a$, make no decision;

if $B(x) \geq a$, accept H_0 and report the conditional error probability $\beta(B) = 1/(1+B)$.

An example will be given in Section 2.4, which will indicate that the "no decision" region is typically rather small. Indeed, T_1^* and T_1 agree (i.e., the "no decision" region disappears) whenever

$$F_0(1) = 1 - F_1(1), \tag{1.1}$$

in which case $r = a = 1$ in T_1. The main situations in which (1.1) is satisfied are situations of "symmetry"; see Sections 3.1 and 4.2 for definition and illustration.

1.3. *Use of T_1 in sequential analysis.* In sequential testing, use of T_1 (or T_1^*) can greatly simplify the analysis and has surprising foundational implications. First, consider the following example.

EXAMPLE 2. Consider the scenario of Example 1, but suppose the data are observed sequentially. Letting N denote the (random) stopping time for the experiment, it is still true that, upon stopping at $N = n$, $B_n(x_1, \ldots, x_n) = \exp\{-2n\,\bar{x}_n\}$. (Bayes factors do not depend on the stopping rule.) Condition (1.1) (which guarantees that $T_1^* = T_1$) can be written (noting that $B_n \leq 1 \Leftrightarrow \bar{x}_n \geq 0$) $P(\bar{X}_N \geq 0 \,|\, \theta = -1) = P(\bar{X}_N \leq 0 \,|\, \theta = 1)$. This will be satisfied by stopping rules that are "symmetric," such as those of the form

(1.2) "stop" at the first n for which $|\bar{x}_n| \geq g(n)$,

where $g(n)$ is any nonnegative function such that the stopping rule is proper (i.e., "stops" with probability 1). The SPRT with equal error probabilities is of this form, with $g(n) \propto 1/n$.

For this situation, the test T_1 becomes:

if $\bar{x}_n \geq g(n)$, stop experimentation, reject H_0 and report the conditional error probability $\alpha(B_n) = 1/[1 + \exp(2n\,\bar{x}_n)]$;

if $\bar{x}_n < -g(n)$, stop experimentation, accept H_0 and report the conditional error probability $\beta(B_n) = 1/[1 + \exp(-2n\,\bar{x}_n)]$.

Observe first that the test T_1 is remarkably easy to implement. There is no need for the usually difficult computation of unconditional frequentist error probabilities, as is necessary for the SPRT. Even the choice of $g(n)$ does not necessarily involve such computations; for instance, the choice

$$g(n) = \frac{1}{2n} \log\left(\frac{1}{\alpha^*} - 1\right)$$

is attractive, guaranteeing that the reported conditional error probability will not exceed α^*.

The foundational surprise here is that the reported error probabilities arising from T_1 do not depend on the specific symmetric stopping rule chosen. This is surprising in light of the common frequentist opposition to the *stopping rule principle* (*SRP*). The SRP states that final inferences should not depend on the stopping rule used to obtain the data [cf. Berger and Berry (1988) or Berger and Wolpert (1988)]. The fact that the "optimal" frequentist procedure, T_1, in the above example, seems to ignore the stopping rule in its error report is startling. This issue is discussed more fully in Section 4.

1.4. *Background on conditional testing and P-values.* There is a long history of attempts to modify frequentist theory by utilizing some form of conditioning. Earlier works are summarized in Kiefer (1977) and Berger and Wolpert (1988). Kiefer (1977), together with Kiefer (1975, 1976) and Brownie and Kiefer (1977), formally established the *conditional confidence approach*; a modification

of this approach is discussed in Section 2.1 and forms the basis of our analysis. We seek, however (for the problem of testing simple hypotheses), to overcome certain difficulties with the conditional confidence approach, difficulties discussed in the above works and by the discussants of Kiefer (1977).

Perhaps the main difficulty is that there is a plethora of conditional confidence procedures, and choosing among them is a daunting task. Brown (1978) makes a serious effort in this direction (see Section 3.2), but a general practical prescription seems remote. A second potentially serious difficulty is that conditional confidence procedures that do not have a Bayesian basis can remain incoherent and anti-intuitive. In this regard, Birnbaum (1961) and Barnard [in the discussion of Kiefer (1977)] argue that any effort to develop intuitively sensible conditional answers, for the testing problem we consider, must be based on $B(x) = f_0(x)/f_1(x)$. Even Kiefer (1977) mentioned the appeal of doing so, but felt (as did Birnbaum) that this was incompatible (in general) with a frequentist interpretation. The main thrust of this paper is to show that the two are compatible; that one can retain the coherency and ease of interpretability of $B(x)$, while achieving a conditional frequentist interpretation for its use.

Although this paper is basically a descendant of Kiefer (1977) [and Birnbaum (1961)], there has been a considerable subsequent literature on conditional frequentist testing from the "estimated confidence" perspective. This approach was also proposed in Kiefer (1977) (although it, too, has earlier roots). The focus of the approach (in testing) is to provide an "estimate" of the indicator function: 1 if H_0 is true, 0 if H_0 is false. Developments along these lines can be found in Schaafsma, Tolboom and van der Meulen (1989), Hwang, Casella, Robert, Wells and Farrell (1992) and van der Meulen (1992). Chatterjee and Chattopadhyay (1992) develop a related approach based on a betting interpretation of evidence. Exploration of these new approaches is valuable, but they frequently are susceptible to the same criticisms that we mentioned earlier in relation to the conditional confidence approach.

Finally, use of the P-value for testing H_0 versus H_1 should be mentioned. Because of the intuitive objections to the $N - P$ test that were discussed in Section 1, use of the P-value has become quite popular. There is, however, no sound justification for using it in simple versus simple hypothesis testing. It has no Bayesian or likelihood interpretation, typically differing significantly from, say, T_1. Thus, in Table 1 for $\bar{x} = 0$, the P-value is the anti-intuitive 0.025. And it has no true frequentist justification; for instance, the expected P-value under f_0, conditional on rejecting, is only one-half the actual probability of rejecting under f_0, so that "on average over the rejection region" the P-value substantially underestimates the actual Type I error rate. Since T_1 is completely justified from all foundational perspectives and is as "data-adaptive" as the P-value, T_1 is clearly to be preferred.

2. Conditional frequentist testing.

2.1. *Basic elements.* The approach we consider here is that formalized by Kiefer (1975, 1976, 1977) and Brownie and Kiefer (1977), called the *conditional*

confidence approach. [See also Brown (1978) and Berger (1985a, b).] The idea is to partition the sample space, $\mathcal{X}$; that is, let

$$\mathcal{X} = \bigcup_{s \in S} \mathcal{X}_s, \quad \text{with } \mathcal{X}_s \cap \mathcal{X}_{s'} = \emptyset,$$

and then develop frequentist measures conditional on $\mathcal{X}_s$.

For testing, the usual conditional frequentist measures considered are the conditional probabilities of Type I and Type II errors,

$$\alpha(s) = P_0(\text{Type I error} \mid \mathcal{X}_s) = P_0(\text{rejecting} \mid \mathcal{X}_s), \tag{2.1}$$

$$\beta(s) = P_1(\text{Type II error} \mid \mathcal{X}_s) = P_1(\text{accepting} \mid \mathcal{X}_s). \tag{2.2}$$

One operates by observing the partition $\mathcal{X}_s$ in which the data happen to lie and then reporting the relevant $\alpha(s)$ or $\beta(s)$ as the error probability (or, perhaps, reporting both). It is worth noting that, if one treats s as random under f_0 and/or f_1, then

$$\begin{aligned} E_0[\alpha(s)] &= E_0\left[P_0(\text{Type I error} \mid \mathcal{X}_s)\right] = P_0(\text{Type I error}) = \alpha, \\ E_1[\beta(s)] &= P_1(\text{Type II error}) = \beta, \end{aligned} \tag{2.3}$$

so the conditional tests can be viewed as simply dividing up the overall probabilities of Type I and Type II errors among the various partitions.

It should be remarked that Kiefer worked in terms of "conditional confidence," which is 1 minus the "conditional error." Also, although Kiefer stopped short of explicitly recommending a particular conditional frequentist procedure for simple versus simple hypothesis testing, he [as well as Brown (1978)] seemed to favor the procedure arising from requiring $\alpha(s) = \beta(s)$ for all s, with $\{\mathcal{X}_s: s \in S\}$ chosen to be as fine a partition as possible, subject to this constraint. See Kiefer (1977), expressions (3.12) to (3.14), for development of this procedure. Although this "equal probability continuum" partition has some attractive properties, it can be shown to yield counterintuitive results in many nonsymmetric situations; hence our preference for T_1 and T_1^*.

2.2. *The general conditional test for simple hypotheses.* The tests T_1 and T_1^*, described in Section 1, can be usefully generalized. Consider the test $T_{l,\rho}$ defined as follows for $0 < l < \infty$ and $0 < \rho < \infty$: define

$$r_{l,\rho} = l\rho \quad \text{and} \quad a_{l,\rho} = F_0^{-1}(1 - \rho F_1(l\rho)) \quad \text{if } F_0(l\rho) \le 1 - \rho F_1(l\rho), \tag{2.4}$$

$$r_{l,\rho} = F_1^{-1}\left(\frac{1}{\rho}[1 - F_0(l\rho)]\right) \quad \text{and} \quad a_{l,\rho} = l\rho \quad \text{if } F_0(l\rho) \ge 1 - \rho F_1(l\rho), \tag{2.5}$$

then the generalization of T_1^* is given by:

if $B(x) \le r_{l,\rho}$, reject H_0 and report the conditional error probability $\alpha_\rho(B) = B/(\rho + B)$;

if $r_{l,\rho} < B(x) < a_{l,\rho}$, make no decision;

if $B(x) \geq a_{l,\rho}$, accept H_0 and report the conditional error probability $\beta_\rho(B) = \rho/(\rho + B)$.

The motivation for considering this test is again Bayesian. Indeed, suppose it is desired to test H_0: f_0 versus H_1: f_1, with π_0 and π_1 being the prior probabilities of H_0 and H_1, respectively ($\pi_1 = 1 - \pi_0$), and where the loss is assumed to be 0 for a correct decision and l_i for an incorrect decision when H_i is true. Then defining

$$\rho = \pi_1/\pi_0 \quad \text{and} \quad l = l_1/l_0,$$

the optimal Bayes test $T^*_{l,\rho}$ can be written as:

if $B \leq l\rho$, reject and report posterior risk $l_0\alpha_\rho(B)$;
if $B > l\rho$, accept and report posterior risk $l_1\beta_\rho(B)$.

This is the test $T_{l,\rho}$, except for the presence in $T_{l,\rho}$ of the "no decision" region $(r_{l,\rho}, a_{l,\rho})$. [Clearly, multiplying $\alpha_\rho(B)$ by l_0 and $\beta_\rho(B)$ by l_1 converts the conditional error probabilities in $T_{l,\rho}$ to conditional risks.] The "no decision" region is the price that must be paid to obtain a conditional frequentist interpretation for the optimal Bayes test. This interpretation is developed in the following section.

A strict frequentist can also use $T_{l,\rho}$ effectively. Suppose a frequentist were planning to use the unconditional $N - P$ test with critical value c and error probabilities $\alpha = F_0(c)$ and $\beta = 1 - F_1(c)$. This can be replaced by T_{l^*,ρ^*}, with

$$\rho^* = (1 - \alpha)/(1 - \beta) \quad \text{and} \quad l^* = c/\rho^*.$$

This test is the generalization of T_1. It is straightforward to verify that T_{l^*,ρ^*} still has c as a critical value, and it does not have a "no decision" region. But instead of reporting the unconditional α and β, one now reports the conditional error probabilities $\alpha_{\rho^*}(B)$ and $\beta_{\rho^*}(B)$. Note that the choice of l^* and ρ^* is here just viewed as formalism; their interpretation in terms of losses and priors is not necessary.

While we feel that use of T_{l^*,ρ^*} is considerably better than use of the $N - P$ test, some of us hesitate to recommend its general use in practice because of the intuitive concerns raised in Section 1.2 and prefer to allow a "no decision" region. This can clearly be left to the personal taste of the practitioner, however.

2.3. *Conditional frequentist interpretation of* $T_{l,\rho}$. Define

$$\psi(s) = F_0^{-1}\big(1 - \rho F_1(s)\big) \quad \text{for } 0 < s \leq r_{l,\rho}. \tag{2.6}$$

Note that $a_{l,\rho} = \psi(r_{l,\rho})$. Define, for $0 < s \leq r_{l,\rho}$,

$$\mathcal{X}_s = \big\{x \in \mathcal{X}\colon B(x) = s \text{ or } B(x) = \psi(s)\big\},$$

and let $\mathcal{X}_0 = \{x \in \mathcal{X}\colon r_{l,\rho} < B(x) < a_{l,\rho}\}$. Then $\{\mathcal{X}_s\colon s \in [0, r_{l,\rho}]\}$ is a partition of $\mathcal{X}$.

Under a "symmetry" condition to be defined in Section 3.1, this partition turns out to be a "maximal ancillary" partition. An ancillary partition

is a partition in which, for each s, $\mathcal{X}_s$ has the same probability (or density) under f_0 as under f_1. The term "maximal" means that no finer partition, among partitions based on the sufficient statistic $B(X)$, can be ancillary. In the non-symmetric case, the above partition is not ancillary and hence has no clear intrinsic justification; rather, it is extrinsically justified by yielding error probabilities that are equal to the intuitively attractive and foundationally secure posterior probabilities.

THEOREM 1. *For the test $T_{l,\rho}$ and the above partition, if $s > 0$, then*

$$\alpha(s) = P_0\big(\textit{rejecting } H_0 \mid \mathcal{X}_s\big) = \alpha_\rho(B) = B/(\rho + B),$$
$$\beta(s) = P_1\big(\textit{accepting } H_0 \mid \mathcal{X}_s\big) = \beta_\rho(B) = \rho/(\rho + B).$$

PROOF. Let $f_i^*(\cdot)$ denote the density of the (sufficient) statistic $B(X)$ under f_i, $i = 1, 2$. We first show that

$$f_0^*(b) = bf_1^*(b). \tag{2.7}$$

To see this, observe that

$$\begin{aligned}\int_0^b f_0^*(y)\,dy &= P_0\big(B(X) \le b\big)\\ &= \int_{\{x:\, B(x) \le b\}} f_0(x)\,dx\\ &= \int_{\{x:\, B(x) \le b\}} B(x) f_1(x)\,dx\\ &= \int_0^b y f_1^*(y)\,dy,\end{aligned}$$

the last step following from the change of variables $y = B(x)$. Differentiating both sides with respect to b establishes (2.7). Calculus also yields

$$\frac{d}{ds}\psi(s) = -\rho f_1^*(s)/f_0^*\big(\psi(s)\big). \tag{2.8}$$

Applying first (2.8) and then (2.7) thus yields

$$\begin{aligned}\alpha(s) &= P_0\big(\text{rejecting} \mid \mathcal{X}_s\big)\\ &= f_0^*(s) \Big/ \left[f_0^*(s) + f_0^*\big(\psi(s)\big) \left| \frac{d}{ds}\psi(s) \right| \right]\\ &= f_0^*(s) \big/ \big[f_0^*(s) + \rho f_1^*(s) \big]\\ &= B/[B + \rho],\end{aligned}$$

concluding the proof for $\alpha(s)$. The proof for $\beta(s)$ is similar. □

2.4. *Discussion of* $T_{l,\rho}$. As promised, the test $T_{l,\rho}$ simultaneously has a Bayesian and a frequentist justification, with the *decision* and the *reported risk* (loss $\times$ error probability) being the same under either paradigm. Of course, the interpretation of the risk (or error probability) will differ for Bayesians and frequentists, but this will be irrelevant to practice. Indeed, those who see merit in both the Bayesian and frequentist philosophies might be especially pleased in the dual interpretation of the reported risk.

Strict frequentists might complain that, in operation, $T_{l,\rho}$ does not provide all needed information. Frequentist dogma asserts that one cannot, say, only look at the Type I error probability when rejecting, but must also look at Type II error. In the conditional frequentist setting, this dogma would imply that both $\alpha(s)$ and $\beta(s)$ must be reported, whether one rejects or accepts. The given procedure, $T_{l,\rho}$, only provides $\alpha(s)$ upon rejection, and $\beta(s)$ upon acceptance.

There are several possible replies to this concern. First of all, the map (2.6) is available and could be used to compute the other error probability if really desired. In practice, however, we suspect that this will not be done, primarily because it is almost impossible to see how one should, say, use $\beta(s)$ upon rejecting. Second, it can be argued that the source of the frequentist dogma concerning the reporting of both error probabilities lies in an attempt to intuitively compensate for not conditioning; such compensation is clearly not needed for $T_{l,\rho}$. Finally, in the very important "symmetric" situation of Sections 3.1 and 4.2, we will see that $\alpha(s) = \beta(s)$ for all s, so that the issue does not then arise.

The major potential thorn in the use of $T_{l,\rho}$ is the presence of the "no decision" region $(r_{l,\rho}, a_{l,\rho})$. In a sense, this is the region over which one cannot force agreement between Bayesians and frequentists. If this region is too large, then the utility of the test $T_{l,\rho}$ is reduced (although recall that a frequentist who is less concerned about agreement with Bayesians can just use T_1 or T_{l^*,ρ^*}, neither of which has a "no decision" region).

An encouraging fact is that it can be shown, in general, that, for any l or any ρ, a solution in the other variable can be found to

$$r_{l,\rho} = a_{l,\rho} = l_\rho, \tag{2.9}$$

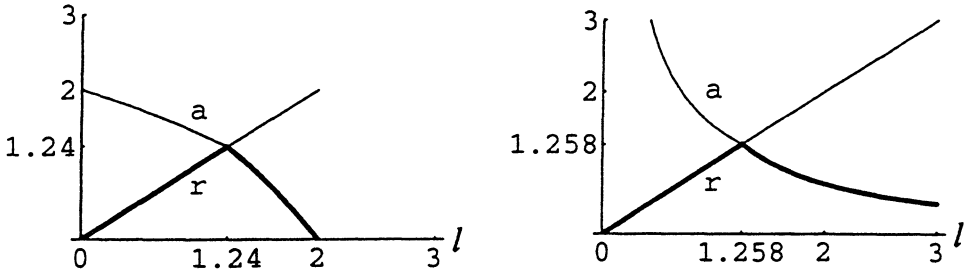
in which case the "no decision" region is empty. It can also be shown that the solutions to (2.9) in l and ρ are inversely related. To investigate how large the "no decision" region can be, consider the following example.

EXAMPLE 3. Suppose that $X_1, \ldots, X_n$ are i.i.d. exponential, with density $f(x_i \mid \theta) = \theta^{-1} \exp\{-x_i/\theta\}$ for $x_i > 0$, and that it is desired to test H_0: $\theta = \theta_0$ versus H_1: $\theta = \theta_1$, where $\theta_0 < \theta_1$. Defining $\gamma = \theta_0/\theta_1$, it is immediate that

$$B = \gamma^{-n} \exp\{-n(1-\gamma)\bar{x}/\theta_0\},$$

which is monotonically decreasing in $\bar{x}$. Note that $0 \le B \le \gamma^{-n}$. Also, computation yields, for the c.d.f. of B under H_i,

$$F_i(b) = 1 - \Gamma\left(n, \frac{-\gamma^i \log(b\gamma^n)}{(1-\gamma)}\right), \tag{2.10}$$



(a) $n = 1$ (b) $n = \infty$

FIG. 1. *The "no decision" region, as a function of l, for testing H_0: $\theta = 1$ versus H_1: $\theta = 2$ under equal prior probabilities.*

where $\Gamma(n,t) = (1/\Gamma(n)) \int_0^t y^{n-1} \exp\{-y\}\, dy$ is the incomplete Gamma function.

Consider, first, the effect of l, the loss ratio l_1/l_0. For the special case $n = 1$,

$$F_0(b) = (b\gamma)^{1/(1-\gamma)} \quad \text{and} \quad F_1(b) = (b\gamma)^{\gamma/(1-\gamma)}.$$

From these and (2.4) and (2.5), $r_{l,\rho}$ and $a_{l,\rho}$ can be explicitly obtained. When $\rho = 1$ and $\gamma = \frac{1}{2}$ for instance (say, when testing H_0: $\theta = 1$ versus H_1: $\theta = 2$ with equal prior probabilities),

$$r_{l,1} = \min\left\{l, 2(1 - l/2)^2\right\} \quad \text{and} \quad a_{l,1} = \max\left\{l, 2(1 - l/2)^{1/2}\right\}$$

if $0 \le l \le 2$; when $l > 2$, $r_{l,1} = 0$ and $a_{l,1} = 2$. (Note that the range of B here is $0 \le B \le 2$.) These are plotted in Figure 1(a).

From Figure 1(a), it is apparent that l can have a dramatic effect. First of all, unless $0 < l < 2$, the "no decision" region is the entire space. And for l small or near 2, the "no decision" region is most of the space. Note that, at $l = \sqrt{5} - 1 \cong 1.24$, the "no decision" region is empty. The resulting test $T_{1.24,1}$ is thus the test T_1 defined in Section 1.1 and could beneficially be used by a frequentist instead of the minimax test (which has critical value $B = 1.24$).

Turning to large n, three cases can be distinguished. Technical details are given in the Appendix.

CASE 1. $\rho > 1$. Then, as $n \to \infty$, $a_{l,\rho} = l_\rho$ and

$$r_{l,\rho} = \exp\left\{ (1 - \gamma^{-1} - \log\gamma)n + (1 - \gamma^{-1}) \left[z_{(1-\rho^{-1})}\sqrt{n} + \frac{1}{3}\left(z^2_{(1-\rho^{-1})} - 1\right) \right] \right\} (1 + o(1)), \tag{2.11}$$

where z_α denotes the αth quantile of the standard normal distribution. Note that $(1 - \gamma^{-1} - \log\gamma) < 0$ (recall that $0 < \gamma < 1$), so that $r_{l,\rho}$ decreases to 0 exponentially fast as $n \to \infty$.

CASE 2. $\rho < 1$. Then, as $n \to \infty$, $r_{l,\rho} = l\rho$ and

$$a_{l,\rho} = \exp\left\{(\gamma - 1 - \log\gamma)n + (\gamma - 1)\left[z_\rho\sqrt{n} + \frac{1}{3}(z_\rho^2 - 1)\right]\right\}(1 + o(1)). \tag{2.12}$$

Note that $(\gamma - 1 - \log\gamma) > 0$ (for $0 < \gamma < 1$), so that $a_{l,\rho}$ increases to ∞ exponentially fast as $n \to \infty$.

CASE 3A. $\rho = 1$, $l > -(1 + g(\gamma))$. Here

$$g(\gamma) \equiv (1 - \gamma)(\log\gamma)/(1 - \gamma + \gamma\log\gamma). \tag{2.13}$$

Then, as $n \to \infty$, $a_{l,1} = l$ and

$$r_{l,1} = l\left[-l/(1 + g(\gamma))\right]^{g(\gamma)} + o(1). \tag{2.14}$$

If $\gamma = \frac{1}{2}$, then the condition on l is $l > 1.258$, and $r_{l,1} \cong (1.679)l^{(-1.258)}$. Note that, if $l = 1.258$, then the "no decision" region is (asymptotically) empty.

CASE 3B. $\rho = 1$, $l < -(1 + g(\gamma))$. Then $r_{l,1} = l$ and, as $n \to \infty$,

$$a_{l,1} = l\left[-l/(1 + g(\gamma))\right]^{-g(\gamma)/(1+g(\gamma))} + o(1). \tag{2.15}$$

If $\gamma = \frac{1}{2}$, then $a_{l,1} \cong (1.510)l^{(-.795)}$. For the "objective" choice $l = 1$, note that this asymptotic "no decision" region of (1, 1.510) is quite close to the corresponding region $(1, \sqrt{2}\,)$ when $n = 1$, seemingly indicating a stability of the "no decision" region with respect to sample size (in this case).

DISCUSSION OF EXAMPLE 3. When $\rho \neq 1$, (2.11) and (2.12) show that the "no decision" region grows enormously as $n \to \infty$. Furthermore, it can be shown in Cases 1 and 2, respectively, that, as $n \to \infty$,

$$P_{\theta_1}(\text{"no decision"}) \to 1 - \rho^{-1}, \qquad P_{\theta_0}(\text{"no decision"}) \to 1 - \rho.$$

Hence the "no decision" region is even nonnegligible probabilistically.

The story is very different when $\rho = 1$. Then (2.14) and (2.15) show that $r_{l,1}$ and $a_{l,1}$ stay bounded. Indeed, as shown in Figure 1(b) for the case $\gamma = \frac{1}{2}$, the "no decision" region for $n = \infty$ is remarkably similar to that for $n = 1$, unless l is extreme. And the probabilities of the "no decision" region under θ_0 and θ_1 go to 0 exponentially fast as $n \to \infty$.

While generalization from a single example is hazardous, we expect the above pattern to hold for other distributions. Thus, for large n and $\rho \neq 1$, the "no

decision" region might be too large to make $T_{l,\rho}$ attractive. But for $\rho = 1$ (which, of course, is the natural "balanced" or "noninformative" assumption) we expect the "no decision" region to be small, even for large sample sizes.

From a theoretical perspective, there are various oddities here. Perhaps the most interesting is that the loss ratio, l, and prior odds ratio, ρ, have very different effects, in contrast to usual Bayesian reasoning. Selecting $\rho = 1$ seems to be important, while $l = 1$ has no special effect.

3. Symmetry and optimality.

3.1. *Likelihood ratio symmetry.* A particularly attractive situation arises when the following is satisfied.

LRS PROPERTY. The testing problem is said to possess *likelihood ratio symmetry* (*LRS*) if $f_0(X)/f_1(X)$ has the same distribution under f_0 as $f_1(X)/f_0(X)$ has under f_1.

EXAMPLE 4. If $X = (X_1, \ldots, X_n)$ arises from a coordinatewise symmetric location density $f(x \mid \theta) = g(|x_1 - \theta|, \ldots, |x_n - \theta|)$ and it is desired to test H_0: $\theta = \theta_0$ versus H_1: $\theta = \theta_1$, then it is straightforward to show that the LRS property holds.

One benefit from the LRS property is indicated in the following lemma.

LEMMA 1. *If the LRS property holds, then* $F_0(1) = 1 - F_1(1)$; *hence the "no-decision" region for the test* T_1^* *in Section* 1.2 *is empty and* $a = r = 1$. (*Clearly,* T_1^* *is then equivalent to* T_1.)

PROOF. Clearly,

$$F_0(b) = P_0(B \le b) = P_1\big(1/B \le b\big) = P_1\big(1/b \le B\big) = 1 - F_1\big(1/b\big).$$

Setting $b = 1$ yields the result. □

Thus, under the LRS property, the Bayesian (with $\rho = 1$ and $l = 1$) and the conditional frequentist using T_1 always report identical numbers. This was noticed by Kiefer (1977), who seemed quite happy with T_1 in this case, in part because it also then corresponds with his "equal probability continuum" procedure (see Section 2.1).

3.2. *Optimality.* Consider the general conditional frequentist testing scenario defined in Section 2.1. Any partition $\{\mathcal{X}_s\colon s \in \mathcal{S}\}$ of $\mathcal{X}$ corresponds to a possible conditional frequentist test T. A natural question to ask (from a frequentist perspective) is whether an optimal T (i.e., optimal partition) exists.

Brown (1978) studied this problem and proposed the following approach for simple versus simple hypothesis testing. Let $h(\cdot)$ be a nondecreasing, convex function [e.g., $h(v) = v^2$] and define the *B-utility* of a conditional frequentist test

T to be the pair $(U_T(0), U_T(1))$, where

$$U_T(i) = E_{f_i}\left[h\left(1 - \max\{\alpha_T(s), \beta_T(s)\}\right)\right],$$

with $\alpha_T(s)$ and $\beta_T(s)$ being the conditional error probabilities defined in (2.1) and (2.2). It is desirable to have the $U_T(i)$ large, because of the following two points:

(i) Since h is nondecreasing, reducing both conditional error probabilities (clearly desirable) will cause the $U_T(i)$ to increase.

(ii) Since h is convex, making the conditional error probabilities more variable in s will cause the $U_T(i)$ to increase; variability in $\alpha_T(s)$ and $\beta_T(s)$ is desirable, because it allows for reflection of varying evidentiary strength for different data.

As an illustration of this last point, the following lemma shows that, under the LRS property, the conditional frequentist test T_1 is superior to the classical Neyman–Pearson test with equal error probabilities.

LEMMA 2. *Suppose the LRS property holds and let T_0 denote the classical Neyman–Pearson test with rejection region $\{B \leq 1\}$. Then*

$$U_{T_1}(i) \geq U_{T_0}(i), \qquad i = 0, 1,$$

with strict inequality if h is strictly convex.

PROOF. The LRS property implies that $\alpha_{T_1}(B) = \beta_{T_1}(B)$ and $\alpha_{T_0} = \beta_{T_0}$. Thus Jensen's inequality yields

$$\begin{aligned} U_{T_1}(i) &= E_{f_i}^{B}\left[h\left(1 - \alpha_{T_1}(B)\right)\right] \\ &\geq h\left(1 - E_{f_i}^{B}\left[\alpha_{T_1}(B)\right]\right) \\ &= h\left(1 - \alpha_{T_0}\right) \qquad \text{[using (2.3)]} \\ &= U_{T_0}(i), \end{aligned}$$

with strict inequality if h is strictly convex. □

Furthermore, it is clear that the same reasoning applies to any other symmetric conditional frequentist test, that is, any test for which each $\mathcal{X}_s$ is defined as

$$\mathcal{X}_s = \{x \in \mathcal{X}: B(x) \in A_s \text{ or } 1/B(x) \in A_s\},$$

for some set A_s. If $\alpha_T(\cdot)$ for such a test differs from $\alpha_{T_1}(\cdot)$ with positive probability and h is strictly convex, then $U_{T_1}(i) > U_T(i)$.

Being "best" among all symmetric conditional tests is quite compelling, but Brown (1978) also establishes two optimality properties of T_1 (under the LRS property) among *all* conditional frequentist tests. First, he shows that T_1 is the *unique* test that is *U-admissible* for all h; thus, for any other test T^*, one can

find a nondecreasing convex h and another test T^{**} such that $U_{T^{**}}(i) \geq U_{T^*}(i)$, with strict inequality for $i = 0$ or 1.

The second global property that Brown establishes is that T_1 is the *unique* test that is *totally maximin*, that is, for which

$$\min\{U_{T_1}(0), U_{T_1}(1)\} = \sup_T \min\{U_T(0), U_T(1)\}$$

holds for all nondecreasing convex h. [Brown's uniqueness result applies because, here, $B(X)$ is assumed to have a nonatomic distribution.]

Brown further suggests that, among monotone procedures, T_1 is probably strictly optimal [i.e., minimizes both $U_T(0)$ and $U_T(1)$]. These optimality results are all particularly compelling because of the great generality allowed in choice of h. The bottom line is that, using reasonable frequentist criteria alone, T_1 appears to be best, in a variety of ways, among all valid (conditional) frequentist tests, at least when the LRS property holds. It does not seem to be possible to establish such strong optimality of T_1 if the LRS property does not hold, but T_1 undoubtedly remains admissible and reasonable even then.

4. Sequential tests.

4.1. *The general conditional sequential test.* Suppose $\mathbf{X} = (X_1, X_2, \ldots)$ is a sequential sample and that it is desired to test $H_0: \mathbf{X} \sim f_0$ versus $H_1: \mathbf{X} \sim f_1$. By this we mean that, for $i = 0, 1$, $f_i = \{f_{i,1}, f_{i,2}, \ldots\}$ with $\mathbf{X}^{(n)} = (X_1, \ldots, X_n)$ having density $f_{i,n}(\mathbf{x}^{(n)})$ for $n \geq 1$. Define

$$B_n = f_{0,n}(\mathbf{x}^{(n)}) / f_{1,n}(\mathbf{x}^{(n)}); \tag{4.1}$$

let N denote the stopping time of the sequential experiment (a proper stopping rule being assumed) and let $F_i(\cdot)$ be the c.d.f. of B_N under f_i, $i = 0, 1$.

The situation is slightly more complicated than that discussed previously in the paper, because F_i is typically not invertible. Usually, however, the following condition is satisfied.

CONDITION S. The stopping rule is such that $B_N \notin (R, A)$, where $R \leq 1 \leq A$; the rejection and acceptance regions are $\{B_N \leq R\}$ and $\{B_N \geq A\}$, respectively; and, for $i = 0, 1$, $F_i(b)$ is invertible for $b \notin (R, A)$, with $F_i(R) = F_i(A)$.

EXAMPLE 5 (The sequential probability ratio test). The famous Wald SPRT is defined by:

if $B_n \leq R$, stop sampling, reject H_0 and report error probability $\alpha = P_0(B_N \leq R)$;

if $B_n \geq A$, stop sampling, accept H_1 and report error probability $\beta = P_1(B_N \geq A)$.

It is immediate that $B_N \notin (R, A)$ and that the rejection and acceptance regions are as stated in Condition S. (Note that stopping and rejecting or accepting

when $\{B_n \leq R\}$ or $\{B_n \geq A\}$ is considerably stronger than the Condition S assumption $\{B_N \leq R\}$ and $\{B_N \geq A\}$; the former actually states that one stops and rejects or accepts upon crossing R or A, while the latter only states that, if one happens to stop and has crossed R or A, then one must reject or accept, respectively). It is virtually always the case that $R \leq 1 \leq A$ for the SPRT.

Under Condition S, the appropriate generalization of T_1^* in Section 1.2 is defined as is T_1^* but with $B(x)$ replaced by B_N and r and a defined by

$$r = R \text{ and } a = F_0^{-1}\big(1 - F_1(A)\big) \quad \text{if } F_0(R) \leq 1 - F_1(A),$$
$$r = F_1^{-1}\big(1 - F_0(R)\big) \text{ and } a = A \quad \text{if } F_0(R) > 1 - F_1(A).$$

It is straightforward to show that the conditional frequentist interpretation of this test is still valid. Note, however, that this generalized T_1^* does not specify the stopping rule; it merely gives the conclusion to report, upon stopping.

There are two difficulties with T_1^*. First, computing F_0^{-1} or F_1^{-1} can be quite challenging in sequential settings. Second, it is somewhat troubling if the "no decision" region (r, a) is larger than the initial (R, A), since the latter is often, in a sense, constructed to be the desired "no decision" region.

These difficulties disappear and, indeed, turn into a delightful advantage if the stopping rule is chosen so that

$$F_0(R) = 1 - F_1(A). \tag{4.2}$$

This is equivalent to the condition that the classical sequential test is constructed to have equal error probabilities, $\alpha = \beta$. Then the conditional frequentist test can be written as follows:

(4.3)
if $B_N \leq R$, reject H_0 and report the conditional error probability $\alpha(B_N) = B_N/(1 + B_N)$;

if $B_N \geq A$, accept H_0 and report the conditional error probability $\beta(B_N) = 1/(1 + B_N)$.

This is the analogue of the test T_1 defined in Section 1.1. Again note, however, that this does not specify when to stop, just what to do upon stopping.

It is of considerable interest that the conditional error probabilities are available explicitly here, while classical (unconditional) error probabilities are typically very hard to compute. Even for the SPRT, computation of α and β usually requires difficult analysis of the "overshoot," the amount by which B_N overshoots R or A [cf. Siegmund (1985)]. The conditional error probabilities are not only trivially computable, but, interestingly, incorporate the overshoot into the error statement; the more the overshoot, the less the stated error.

4.2. *Symmetric sequential tests.* In Section 4.1 it was assumed that stopping is governed by a stopping rule separate from T_1^*. In this section we explore the extent to which that assumption can be relaxed.

We will consider the situation in which $X_1, X_2, \ldots$, arise as i.i.d. observations from $f_0(x_i)$ or $f_1(x_i)$, with $Y_i = f_0(X_i)/f_1(X_i)$ having the LRS property (i.e., the distribution of Y_i under f_0 is the same as that of $1/Y_i$ under f_1). We also assume that the stopping rule $\tau = \{\tau_1, \tau_2, \ldots\}$ depends only on $Y_1, Y_2, \ldots$ and is *symmetric* in the sense that

$$\tau_i(Y_1, \ldots, Y_i) = \tau_i\big(1/Y_1, \ldots, 1/Y_i\big) \tag{4.4}$$

for all $i \geq 1$. (As usual, τ_i gives the probability, typically 0 or 1, of stopping sampling upon observing $Y_1, \ldots, Y_i$. More general stopping rules involving other chance mechanisms could be allowed, so long as they are "noninformative"—see Berger and Wolpert (1988) for definition—but the generality here suffices to make the basic point.)

EXAMPLE 6. A common class of symmetric stopping rules is given by

$$\tau_n(Y_1, \ldots, Y_n) = \begin{cases} 1 \text{ (i.e., stop)}, & \text{if } \left|\sum_{i=1}^{n} \log Y_i\right| \geq g(n), \\ 0 \text{ (i.e., continue)}, & \text{if } \left|\sum_{i=1}^{n} \log Y_i\right| < g(n), \end{cases}$$

where $g(n)$ is any arbitrary function for which τ is a proper stopping rule (i.e., is guaranteed to eventually result in "stop"). The SPRT with symmetric boundaries, $A = R^{-1}$, corresponds to $g(n) = \log A$.

For this symmetric situation, it is straightforward to verify that (4.2) holds, so that (4.3) defines T_1^*; note that it is not even necessary here to explicitly calculate the quantities in (4.2). That (4.3) defines a valid conditional frequentist test, in this situation, was already recognized by Kiefer (1977). Finally, note that if τ is restricted to be as above (i.e., depends only on the Bayes factors $B_n = \prod_{i=1}^{n} Y_i$), then Brown (1978) applies and shows that T_1^* in (4.3) defines the "optimal" conditional frequentist test. To understand the practical ramifications of this situation, consider the following example.

EXAMPLE 7. A sequential experiment is conducted involving i.i.d. $\mathcal{N}(\theta, 1)$ data for testing H_0: $\theta = 0$ versus H_1: $\theta = 1$ under a symmetric stopping rule (or at least a rule for which $\alpha = \beta$). Suppose the report states that sampling stopped after 20 observations, with $\bar{x}_{20} = 0.7$. One can then "replace" whatever sequential test was used by T_1^* in (4.3). Computing

$$B_{20} = \prod_{i=1}^{20} \big[f(x_i \mid 0)/f(x_i \mid 1)\big] = \exp\left\{-20\left(\bar{x}_{20} - \frac{1}{2}\right)\right\} = 0.018,$$

it follows that your conclusion should be to reject H_0, with associated conditional error probability $\alpha(B_{20}) = B_{20}/(1 + B_{20}) = 0.018$. This will be a "better" conclusion than that reached in the study (unless they also used T_1^*). Note that you do not need to explicitly know the stopping rule used in order to perform the optimal

analysis. This is quite attractive in practice because, all too often, the exact stopping rule is incompletely specified and, perhaps, even incompletely known to the experimenters! For instance, if the experimenters had not prespecified the stopping rule, but simply monitored the data stream and stopped when they wished, analysis with T_1^* would still be possible (under the weak assumption of symmetry).

The ramifications for sequential testing of simple hypotheses are profound. First, it appears to be possible to avoid the typically difficult classical computations involving the stopping rule. Second, and more importantly, it seems that preexperimental analysis can be avoided; one can simply start collecting data and stop whenever one desires, as long as T_1^* is then employed. This last fact strongly refutes the usual frequentist argument against the stopping rule principle, the argument which asserts that allowing one to monitor the data stream and arbitrarily stop allows a "biasing" of the result.

It is interesting to consider why the stopping rule "disappears" here. If we were to actually compute the conditioning partitions, $\mathcal{X}_s$, corresponding to T_1^*, we would indeed need to know the stopping rule. But since we will choose the partitions which guarantee $\alpha(B_n)$ and $\beta(B_n)$ as the conditional error probabilities, there is no need to actually compute the $\mathcal{X}_s$.

5. Conclusions and generalization. For the simple versus simple testing problem, we feel that the procedures proposed in this paper should become standard statistical practice. They are easy to understand, interpret and use (the sequential versions, for instance, often being much easier than, say, the SPRT); they are correct from Bayesian and likelihood perspectives; and they are valid (and often optimal) frequentist procedures.

A second conclusion from the paper is foundational: frequentist theory, itself, seems to suggest that optimal conditional frequentist procedures will ignore the stopping rule in sequential experimentation. At the very least, the classical argument against the stopping rule principle is dramatically weakened by the results here.

While the testing of simple hypotheses is often considered as a "practical" approximation in sequential settings, it is admittedly very specialized. From a Bayesian perspective, however, *any* problem of testing H_0: *X has density* $f_0(x \mid \theta_0)$ versus H_1: *X has density* $f_1(x \mid \theta_1)$, where θ_0 and/or θ_1 are unknown, can be reduced to simple versus simple testing: the Bayes factor of H_0 to H_1 is

$$B(x) = m_0(x)/m_1(x) \equiv \int f_0(x \mid \theta_0)\pi_0(d\theta_0) \Big/ \int f_1(x \mid \theta_1)\pi_1(d\theta_1),$$

where π_i is the prior distribution of θ_i, $i = 0, 1$, so that a Bayesian is implicitly testing H_0: *X has density* m_0 versus H_1: *X has density* m_1. As the latter test is simple versus simple, T_1 and/or T_1^* can be applied. The key question, however, is whether or not these tests can be given a satisfactory (conditional) frequentist interpretation in the original problem.

If H_0 is simple (i.e., θ_0 is absent or, equivalently, assumes a specific value), then the answer is—yes! The conditional Type I error probability is precisely the posterior probability of H_0 (assuming equal prior probabilities of the hypotheses), while the conditional Type II error probability (which for T_1 or T_1^* is, of course, the posterior probability of H_1) has an interpretation as a (posterior) expected frequentist Type II error. Clarification and discussion of this idea will be presented elsewhere, but the preliminary indication is that Bayesian and frequentist testing may be generally compatible; the severe conflicts that have, in the past, been observed between the two are, perhaps simply due to use of an inferior frequentist test, namely the unconditional test.

It is of interest that the "better" conditional frequentist tests will (in the composite hypothesis case) depend on the prior distributions assigned to the unknown parameters of the composite hypotheses. Robust Bayesian theory can perhaps provide conditional frequentist tests that are guaranteed to be better than the unconditional tests, but the preliminary indication for composite hypotheses is that some utilization of prior information will be necessary in defining good (conditional) frequentist tests. Note, however, that this use of prior information will probably be no more severe than is the customary use of prior information in selecting power levels for unconditional frequentist testing.

APPENDIX

Technical Details for Example 3.

LEMMA 3. *For the incomplete Gamma function* $\Gamma(n,t)$, *the following hold*:

(i) *If* $c \neq 1$, *then*, *as* $n \to \infty$,

$$\text{(A1)}\quad \Gamma\big(n, cn + d + o(1)\big) = 1_{(1,\infty)}(c) + \big(ce^{1-c}\big)^n \frac{\exp\Big\{d\big(c^{-1}-1\big)\Big\}}{\sqrt{2\pi n}(1-c)}\big(1+o(1)\big).$$

(ii) *If*, *as* $n \to \infty$,

$$\text{(A2)}\qquad \Gamma(n,t_n) = \xi + o\left(\frac{1}{\sqrt{n}}\right)$$

for $0 < \xi < 1$, *then*

$$\text{(A3)}\qquad t_n = n + z_\xi\sqrt{n} + \frac{1}{3}\big(z_\xi^2 - 1\big) + o(1).$$

PROOF. First consider $c > 1$ in part (i). A valid expansion of $\Gamma(n, \cdot)$ is

$$\Gamma(n,t) = 1 - \frac{1}{\Gamma(n)} t^{(n-1)} e^{-t}\left[1 + \frac{(n-1)}{t} + \frac{(n-1)(n-2)}{t^2} + \cdots\right].$$

Setting $t_n = cn + d + o(1)$, it is straightforward to show that

$$\left[1 + \frac{(n-1)}{t_n} + \frac{(n-1)(n-2)}{t_n^2} + \cdots\right] \to \left[1 + \frac{1}{c} + \frac{1}{c^2} + \cdots\right] = \frac{c}{c-1}.$$

Approximating $\Gamma(n)$ by Stirling's formula then yields

$$\Gamma(n, t_n) = 1 - \frac{\big(cn + d + o(1)\big)^{n-1} e^{-(cn+d+o(1))} c}{e^{-n} n^{(n-1/2)} \sqrt{2\pi}(c-1)} \big(1 + o(1)\big).$$

Clearly,

$$\frac{\big(cn + d + o(1)\big)^{n-1}}{n^{n-1}} = c^{n-1}\left(1 + \frac{d + o(1)}{cn}\right)^{n-1} = c^{n-1} e^{d/c}\big(1 + o(1)\big),$$

from which the result follows. The proof of part (i) for $c < 1$ is similar, but now uses the expansion

$$\Gamma(n, t) = e^{-t} \sum_{k=n}^{\infty} t^k / k!.$$

Details are omitted.

To prove part (ii), we use the fact that $\Gamma(n, t) = F(2t \mid 2n)$, where $F(\cdot \mid \nu)$ is the c.d.f. of the chi-squared distribution with ν degrees of freedom. Thus (A2) can be rewritten as

$$F\big(2t_n \mid 2n\big) = \xi + o\left(\frac{1}{\sqrt{n}}\right) \equiv \xi_n.$$

But, as $n \to \infty$, the ξ_nth quantile of the chi-squared $(2n)$ distribution is

$$\begin{aligned} \chi^2_{\xi_n} &= 2n + z_{\xi_n}\sqrt{4n} + \frac{2}{3}\big(z^2_{\xi_n} - 1\big) + o(1) \\ &= 2n + z_{\xi}\sqrt{4n} + \frac{2}{3}\big(z^2_{\xi} - 1\big) + o(1). \end{aligned}$$

This directly yields (A3). □

From (2.10), observe that

(A4) $$F_0(l\rho) = 1 - \Gamma(n, c_0 n + d_0), \qquad F_1(l\rho) = 1 - \Gamma(n, c_1 n + d_1),$$

where

(A5) $$c_0 = \frac{-\log \gamma}{(1-\gamma)}, \qquad d_0 = \frac{-\log(l\rho)}{(1-\gamma)}, \qquad c_1 = \gamma c_0, \qquad d_1 = \gamma d_0.$$

Since $0 < \gamma < 1$, it can be shown that $c_0 > 1$ and $0 < c_1 < 1$. It is then immediate, from (A1) and the fact that $c \exp\{1 - c\} < 1$ for $c \neq 1$, that

(A6) $$F_0(l\rho) = o\left(\frac{1}{\sqrt{n}}\right) \quad \text{and} \quad F_1(l\rho) = 1 - o\left(\frac{1}{\sqrt{n}}\right).$$

Verification of (2.11) *and* (2.12). Consider, first, the case $\rho > 1$. It is clear from (A6) that $F_0(l\rho) > 1 - \rho F_1(l\rho)$ for large n. Hence [see (2.5)], $a_{l,\rho} = l\rho$ and

$$F_1(r_{l,\rho}) = \frac{1}{\rho}\left[1 - F_0(l\rho)\right] = \frac{1}{\rho} + o\left(\frac{1}{\sqrt{n}}\right).$$

This can be rewritten as

$$\Gamma\left(n, c_1 n - \frac{\gamma \log(r_{l,\rho})}{(1-\gamma)}\right) = 1 - \frac{1}{\rho} + o\left(\frac{1}{\sqrt{n}}\right).$$

Hence (A3) yields

$$c_1 n - \frac{\gamma \log(r_{l,\rho})}{(1-\gamma)} = n + z_{(1-1/\rho)}\sqrt{n} + \frac{1}{3}\left(z^2_{(1-1/\rho)} - 1\right) + o(1).$$

Solving for $r_{l,\rho}$ yields (2.11). The derivation of (2.12) for $\rho < 1$ is similar.

Verification of (2.14) *and* (2.15). Since $\rho = 1$, the inequality defining applicability of (2.4) or (2.5) is $F_0(l) \geq 1 - F_1(l)$. Using (A1), (A4) and (A5), this condition will be satisfied as $n \to \infty$ only if $l > -(1 + g(\gamma))$ [see (2.13)]. From (2.5), we then know that $a_{l,1} = l$ and $F_1(r_{l,1}) = 1 - F_0(l)$, which can be rewritten, using (2.10) and (A4), as

$$\Gamma\left(n, c_1 n - \frac{\gamma \log(r_{l,1})}{(1-\gamma)}\right) = \Gamma(n, c_0 n + d_0). \tag{A7}$$

The solution to this equation is given in (2.14). To show this, note that then

$$\log(r_{l,1}) = \log\left(l\left[-l/\left(1 + g(\gamma)\right)\right]^{g(\gamma)}\right) + o(1),$$

so that (A1) can be applied to both sides of (A7); algebra then verifies their equality.

If $l < -(1 + g(\gamma))$, (2.4) applies. Thus $r_{l,1} = l$ and an argument similar to that above verifies (2.15).

REFERENCES

BERGER, J. (1985a). The frequentist viewpoint and conditioning. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (L. Le Cam and R. Olshen, eds.) 15–44. Wadsworth, Monterey, CA.

BERGER, J. O. (1985b). A review of J. Kiefer's work on conditional frequentist statistics. In *Jack Carl Kiefer: Collected Papers. Supplementary Volume* (L. D. Brown, I. Olkin, J. Sacks, H. P. Wynn, eds.) 48–56. Springer, New York.

BERGER, J. and BERRY, D. (1988). The relevance of stopping rules in statistical inference (with discussion). In *Statistical Decision Theory and Related Topics 4* (S. S. Gupta and J. O. Berger, eds.) **1** 29–72. Springer, New York.

BERGER, J. O. and WOLPERT, R. L. (1988). *The Likelihood Principle*, 2nd ed. IMS, Hayward, CA.

BIRNBAUM, A. (1961). On the foundations of statistical inference: binary experiments. *Ann. Math. Statist.* **32** 414–435.

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

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

CHATTERJEE, S. K. and CHATTOPADHYAY, G. (1992). Detailed statistical inference—an alternative non-Bayesian approach: two-decision problem. Technical Report, Dept. Statistics, Calcutta Univ.

HWANG, J. T., CASELLA, G., ROBERT, C., WELLS, M. T. and FARRELL, R. (1992). Estimation of accuracy in testing. *Ann. Statist.* **20** 490–509.

KIEFER, J. (1975). Conditional confidence approach in multi-decision problems. In *Multivariate Analysis* (P. R. Krishnaiah, ed.) **4** 143–158. Academic, 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.

SCHAAFSMA, W., TOLBOOM, J. and VAN DER MEULEN, E. A. (1989). Discussing truth or falsity by computing a Q value. In *Statistical Data Analysis and Inference* (Y. Dodge, ed.) 85–100. North-Holland, Amsterdam.

SIEGMUND, D. (1985). *Sequential Analysis: Tests and Confidence Intervals.* Springer, New York.

VAN DER MEULEN, E. A. (1992). Assessing weights of evidence for discussing classical statistical hypotheses. Ph.D. thesis, Univ. Groningen.

JAMES O. BERGER
DEPARTMENT OF STATISTICS
PURDUE UNIVERSITY
WEST LAFAYETTE, INDIANA 47907-1399

LAWRENCE D. BROWN
DEPARTMENT OF STATISTICS
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104-6302

ROBERT L. WOLPERT
INSTITUTE OF STATISTICS AND DECISION SCIENCES
DUKE UNIVERSITY
DURHAM, NORTH CAROLINA 27708-0251