Simultaneous Bayesian-Frequentist Sequential Testing of Nested Hypotheses

Revision #1, January 98

BY JAMES O. BERGER

Institute of Statistics and Decision Sciences, Duke University, Durham, North Carolina, 27708, USA

BENZION BOUKAI

Department of Mathematical Sciences, Indiana University-Purdue University, Indianapolis, Indiana, 46202, USA

AND YINPING WANG

Department of Mathematical Sciences, Indiana University-Purdue University, Indianapolis, Indiana, 46202, USA

SUMMARY

Conditional frequentist tests of a precise hypothesis versus a composite alternative have recently been developed, and have been shown to be equivalent to conventional Bayes tests in the very strong sense that the reported frequentist error probabilities equal the posterior probabilities of the hypotheses. These results are herein extended to sequential testing, and yield fully frequentist sequential tests that are considerably easier to use than are conventional sequential tests. Among the interesting properties of these new tests is the lack of dependence of the reported error probabilities on the stopping rule, seeming to lend frequentist support to the Stopping Rule Principle.

Some key words: Bayes factor; Composite hypothesis; Conditional test; Error probabilities; Likelihood ratio; SPRT; Sequential test.

1. INTRODUCTION

1·1. *Background*

Sequential testing of composite hypotheses from a frequentist viewpoint is typically viewed to be quite difficult. Wald (1947) considered a generalization of the Sequential Probability Ratio Test (SPRT) by utilizing a weight function (some sort of prior distribution) on composite hypotheses. Subsequent references and efforts to deal with the problem can be found in Ghosh (1970) and Siegmund (1985). The primary difficulties are in the computation of error probabilities, and in the related matter of choosing a suitable stopping rule.

There are also inherent deficiencies in unconditional frequentist testing, partic-

ularly when testing precise hypotheses. Most notable is that the unconditional frequentist tests report error probabilities that are independent of the given data; thus, for an $\alpha = 0.05$ level test, one reports the same error probability upon rejection whether the data is just at the rejection boundary or far within the rejection region. (This criticism only applies when there can be substantial "overshoot" of the stopping boundary, although related criticisms arise when truncated stopping times are used and one stops at the truncation time.) The traditional classical way of addressing this problem is to report a $p - value$ or attained significance level, which were developed for certain sequential settings in Siegmund (1985, Sections 3.4 and 4.5). Note, however, that $p - values$ are not true frequentist error measures, in the sense of being error rates in *real* repeated experimentation; hence we do not view them to be frequentist solutions to the problem. Indeed, there has been extensive discussion as to how $p - values$ can be highly misleading when interpreted as frequentist error rates in testing a precise null hypothesis against a composite alternative. (See Berger, Boukai & Wang (1997) for discussion and earlier references).

Since Kiefer (1977), it has been recognized that frequentists can attempt to employ conditional testing methodology to overcome the above inadequacies. The basic idea behind this conditional frequentist approach is to construct a statistic measuring 'strength of evidence' in the data (historically, an ancillary statistic, if available) and then to compute the appropriate error probabilities conditional on this statistic. Implementation of the approach proved troublesome, however, in part because of the difficulty in choosing the conditioning statistic and in part because of the difficulty in interpreting the resulting conditional error probabilities.

Berger, Brown & Wolpert (1994) proposed an appealing implementation of the conditional frequentist method, for the testing of simple versus simple hypotheses. They argued for choice of a conditioning statistic based on Bayesian reasoning, and

indeed showed that the resulting conditional frequentist error probabilities would exactly equal the Bayesian posterior probabilities of the hypotheses. One startling aspect of the Berger, Brown & Wolpert (1994) result for sequential testing was that they identified (common) situations in which the new Conditional Frequentist test did not depend on the stopping rule, lending frequentist support to the Stopping Rule Principle. (See Berger & Wolpert, 1988, for discussion and history of this famous, but contentious, principle.)

The new conditional testing method was generalized, for the fixed sample situation, in Berger, Boukai & Wang (1997) to testing problems involving a composite alternative. In the present paper, we show how the same testing method can be adapted to sequential testing involving precise and composite hypotheses. Although, the theory becomes rather involved, the resulting methodology is actually quite simple; indeed, is much simpler than current unconditional sequential testing methodology. To illustrate this, Section 1·2 presents an application of the new methodology to sequential t-testing. While certain theoretical details will be left until later, the practical appeal of the new methodology will, hopefully, be evident. Section 2 gives the needed notation and preliminaries. Section 3 presents the formal development of the new sequential conditional frequentist test, emphasizing the Bayesian-Frequentist duality. Section 4 gives applications to sequential testing involving a normal mean. Section 5 compares the new test with certain unconditional sequential tests, and makes some concluding philosophical remarks.

1·2. *The (truncated) conditional sequential t-test*

Suppose we can observe a sequence of responses, $X_1, X_2, \ldots$, of independent and identically distributed $\mathcal{N}(\theta, \sigma^2)$ random variables, with unknown mean θ and variance σ^2, and that we wish to sequentially test $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

A (conventional) Bayesian approach to this problem is described in Section 4, leading to a test statistic B_n, defined by (4·2) and (4·4). (In fact, B_n is the

Bayes factor of H_0 to H_1, based on the data $x_1, \ldots, x_n$, and is often described by Bayesians as the odds of H_0 to H_1 arising from the data.) The statistic B_n forms the basis for the new (truncated) conditional sequential t-test, T^*, defined as follows:

Step 1. Choose constants $0 < R < 1 < A$, a positive integer m and *stopping time*, $N \equiv N_m = \{\text{first } n < m \text{ such that } B_n \notin (R, A); \text{ else, choose } n = m\}$.

Intuitively, R could be thought of as the desired odds of H_0 to H_1 at which one would wish to reject H_0 (e.g., $R = 1/10$), while A could be thought of as the desired odds at which one would wish to accept H_0 (e.g. $A = 10/1$). The truncation time, m, is the maximum number of observations that will be taken. (Untruncated stopping times will be discussed later.)

Step 2. After stopping experimentation at time N, make *inference* as follows:

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

The constant a is computed through equation (4·5) and is typically less than A, in which case the *no decision region* is relevant only if the truncation time m is reached. (In rare circumstances a "rejection critical value," r, defines the *no decision region*; see Section 3.)

As a specific example, consider the data set in Table 6.3 of Armitage (1975). The data arose as the difference in time to recovery between paired patients who were administered different hypotensive agents. Armitage used a type of approximate sequential t-test truncated at $m = 62$. Suppose $R = 0.1$ and $A = 9$ are chosen; these, together with $m = 62$, define the *stopping boundaries*, and are shown in Figure 1 which also presents the data, graphed as B_n versus n. Computation then yields $a = 3.72$; the resulting *decision boundaries* are also shown in Figure 1.

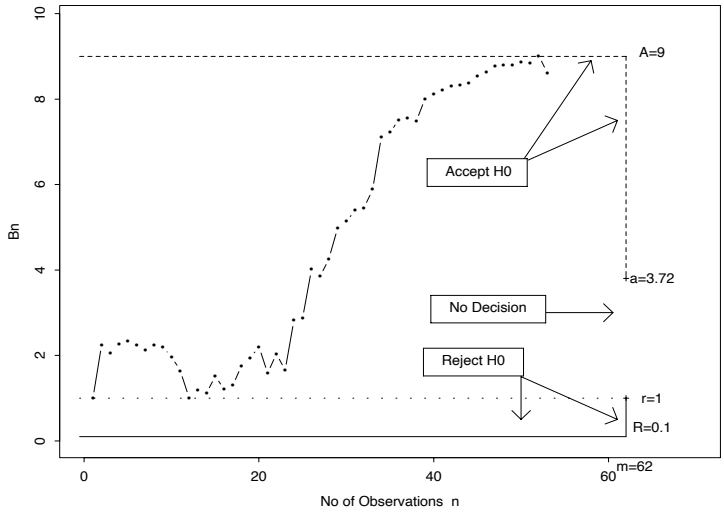

Figure 1: The truncated conditional sequential test for Armitage's (1975) data with R=0.1, A=9, and m=62. For this test, experimentation would have been stopped at N=52.

For the given data, the stopping boundary $A = 9$ would have been reached with $n = 52$ observations; indeed, $B_{52} = 9.017$. The conclusion of the test would then be to accept H_0 and report *conditional error probability* $\beta^*(B_{52}) = 1/(1 + B_{52}) = 0.100$. Note that the experiment reported by Armitage actually was stopped at the 53^{rd} observation.

For design purposes, one might be interested in some pre-experimental properties of the test, such as its unconditional error probabilities (see Section 4). For the situation depicted in Figure 1, the unconditional error probabilities of Type I and Type II are calculated to be $\alpha' = 0.048$ and $\beta' = 0.109$.

1·3. *Comments and motivation*

I) The major motivation for the new test is that the error probabilities, $\alpha^*(B_N)$ and $\beta^*(B_N)$, will be seen to be interpretable either as the posterior probability of H_0 and H_1, respectively (for a conventional Bayes test), or as frequentist probabilities of Type I and Type II error, respectively, conditional on a statistic, S, measuring "strength of evidence" in the data.

II) Note that the quantities R, A, and m each have a clear intuitive interpretation in isolation. In contrast, stopping boundaries in unconditional sequential experimentation only have explicit interpretation through the resulting unconditional error probabilities.

III) Use of the new test is typically trivial computationally. The only potentially difficult computation is the determination of a (see Section 3), but even this can often be avoided in practice. Indeed, a is clearly irrelevant if one ends up rejecting the null hypothesis. Also for a moderate truncation time, m, the constant a will typically be less than A, so that one can typically ignore a when crossing the acceptance boundary, A, at any time $N < m$. In contrast, unconditional (truncated and untruncated) sequential testing of composite hypotheses typically involves challenging stochastic process computations.

IV) The new test is an *exact* frequentist test, not involving any type of approximation such as ignoring "overshoot". The conditional error probabilities are available explicitly and incorporate the overshoot in the error statement; the greater the overshoot, the smaller the stated error.

V) One of the most contentious ideas in statistics is the Stopping Rule Principle (SRP), which states that, upon stopping experimentation, inference should not depend on the reason experimentation was stopped. Bayesian analysis automatically follows the SRP, but the SRP has been perceived to be incompatible with frequentist analysis. Interestingly, the new test is fully frequentist and yet is compatible with the SRP, insofar as the reported error probabilities never depend on the stopping rule. The constant a (which determines the "no decision region") will depend on the stopping rule used but, as argued in Comment III, this is often irrelevant in practice.

VI) For those who do not like the fact that one might end up in the "no decision region", note that it is always possible to choose R and A so that the no-decision

region disappears (i.e. $a = 1$). Indeed, for any choice of R, there is a choice of A for which this will be so. (We are not, necessarily, advocating such a choice, however.) Furthermore, the presence of the no-decision region in the vertical (truncated) stopping boundary will be appealing to many; intuitively, an experiment which stops by hitting the vertical stopping boundary near its 'center' has simply failed to be conclusive. (Note, also, that the conditional error probabilities near the center of this vertical boundary will be much larger than those at the ends, which again corresponds to intuition but is not captured by unconditional error probabilities.) Finally, the new conditional test avoids the problem of determining how to "divide" the vertical stopping boundary between the rejection and the acceptance regions; the new test prescribes this automatically.

2. Basic notation and preliminaries

Let $X_1, X_2, \ldots,$ be a sequence of observable random variables and, for each $n = 1, 2, \ldots,$ write $\underset{\sim}{X}_n = (X_1, \ldots, X_n)$ and let $\mathcal{F}_n = \sigma\{\underset{\sim}{X}_n\}$ denote the corresponding sigma-algebra. Let $\mathcal{F}_\infty$ denote the smallest sigma-algebra containing all the $\mathcal{F}_n$. Here, n represents time and $\underset{\sim}{X}_n$ is the data available at time n. Let N be a proper stopping time (adaptive to $\{\mathcal{F}_n\}$) and let $\mathcal{F}_N$ denote the collection of all events $D \in \mathcal{F}_\infty$ determined prior to N, i.e. $\mathcal{F}_N \equiv \sigma\{\, D \in \mathcal{F}_\infty;\ \ \{N = n\} \cap D \in \mathcal{F}_n, \forall n \geq 1\}$, (c.f. Woodroofe (1982)). Given $\theta \in \Theta$, a subset of some Euclidean space, let P_θ denote the unique probability measure on $\mathcal{F}_\infty$ under which $\underset{\sim}{X}_n$ has joint probability density function $f_n(\underset{\sim}{x}_n|\theta)$, $\underset{\sim}{x}_n = (x_1, \ldots, x_n)$, for each $n = 1, 2, \ldots$. In fact, $f_n(\underset{\sim}{x}_n|\theta)$ may be viewed as the density corresponding to the restriction, P_θ^n, of P_θ to $\mathcal{F}_n$.

Consider sequential testing of a simple hypothesis versus a composite alternative as given by

$$H_0 :\ \theta = \theta_0 \ \text{ versus } \ H_1 :\ \theta \in \Theta_1, \tag{2·1}$$

for some $\theta_0 \in \Theta$, $\theta_0 \notin \Theta_1 \subset \Theta$. Frequently, Θ_1 is given by $\Theta_1 = \{\theta \in \Theta :\ \theta \neq \theta_0\}$.

In the default Bayesian framework (cf. Jeffreys (1961)), one specifies equal prior probabilities of $1/2$ for H_0 and H_1 being true, and chooses a default prior density $\pi(\theta)/2$ on Θ_1, where $\pi(\cdot)$ is a proper density on Θ_1 (with respect to Lebesgue measure). Typically (under, say, "0-1" loss), one then accepts (rejects) H_0 as its posterior probability is greater than (less than) $1/2$ and the reported error probability is just the posterior probability of H_1 (H_0).

For each fixed n, the marginal density functions of $\underset{\sim}{X}_n$ under H_0 and H_1 in $(2\cdot1)$ are given, respectively, by

$$m_{0,n}(\underset{\sim}{x}_n) = f_n(\underset{\sim}{x}_n|\theta_0) \quad \text{and} \quad m_{1,n}(\underset{\sim}{x}_n) = \int_{\Theta_1} f_n(\underset{\sim}{x}_n|\theta)\pi(\theta)d\theta. \tag{2·2}$$

For a Bayesian, the sequential test of the hypotheses $(2\cdot1)$ is equivalent to the sequential test of the simple hypotheses

$$H_0' : \underset{\sim}{X}_n \sim m_{0,n}(\underset{\sim}{x}_n), \quad \forall n \geq 1, \quad \text{versus} \quad H_1' : \underset{\sim}{X}_n \sim m_{1,n}(\underset{\sim}{x}_n), \quad \forall n \geq 1, \tag{2·3}$$

which is based, for each fixed $n \geq 1$, on the corresponding likelihood ratio for $\underset{\sim}{x}_n$,

$$B_n = \frac{m_{0,n}(\underset{\sim}{x}_n)}{m_{1,n}(\underset{\sim}{x}_n)}. \tag{2·4}$$

As was mentioned before, B_n is also the Bayes factor in favor of H_0, which is often viewed as the odds of H_0 to H_1 (of H_0' to H_1') arising from the data. Moreover, it can be verified that, upon stopping at $N = n$, the posterior probability of H_0 being true, given the data $\underset{\sim}{x}_n$, is

$$\alpha^*(B_n) \equiv \text{pr}(H_0|\underset{\sim}{x}_n) = B_n/(1 + B_n), \tag{2·5}$$

whereas the posterior probability of H_1 being true is

$$\beta^*(B_n) \equiv \text{pr}(H_1|\underset{\sim}{x}_n) = 1/(1 + B_n). \tag{2·6}$$

Note that, in accord with the SRP, posterior probabilities are unaffected by the choice of the stopping rule.

The classical frequentist approach to sequentially testing the hypotheses $(2\cdot3)$ is to construct appropriate *rejection* and *acceptance* regions, reporting error probabilities of Type I and Type II given by α' (the probability of incorrect rejection of the null) and β' (the probability of incorrect acceptance of the null). When considering a composite alternative, as is H_1 in $(2\cdot1)$, the probability of Type II error is a function of θ, $\beta(\theta) = \text{pr}(\textit{Accepting}\ \ H_0|\theta) \equiv P_\theta(\textit{Accepting}\ \ H_0)$, for $\theta \in \Theta_1$. Note that (α', β') and $\beta(\theta)$ will depend on the stopping rule used.

To allow the reporting of data–dependent error probabilities, the Conditional Frequentist considers some statistic $S(\underset{\sim}{X}_N)$, where the magnitude of $S(\underset{\sim}{X}_N)$ reflects the evidentiary strength in the data (for, or against, H_0), and then reports the error probabilities conditional on $S(\underset{\sim}{X}_N) = s$, where s denotes the observed value of $S(\underset{\sim}{x}_N)$. These *conditional error probabilities* (abbreviated, here, by CEP) of Type I and Type II can be obtained, respectively, as $\alpha(s) = \text{pr}(\textit{Rejecting}\ \ H_0|\theta_0,\ S(\underset{\sim}{X}_N) = s)$ and $\beta(s|\theta) = \text{pr}(\textit{Accepting}\ \ H_0|\theta,\ S(\underset{\sim}{X}_N) = s)$ for $\theta \in \Theta_1$. The challenge in conditional frequentist inference has been to find suitable statistics upon which to condition (see Kiefer, 1977, Brown, 1978, and the discussion in Berger, Boukai & Wang, 1997). In the following, we use Bayesian intuition to guide the choice.

3. The modified Bayes-sequential test

We present the sequential version of the modified Bayesian test studied in Berger, Boukai & Wang (1997) for the composite alternative case. Thus, in the testing situation of Section 2, let N be a proper stopping time of the sequential experiment (i.e., $\text{pr}(N < \infty|\theta) = 1 \quad \forall\, \theta \in \Theta$). Note that B_N is then $\mathcal{F}_N$ measurable. Let P_i denote probability under H_i', $i = 0, 1$, in $(2\cdot3)$. That is, for any $\mathcal{D} \in \mathcal{F}_N$,

$P_0(\mathcal{D}) = \mathrm{pr}(\mathcal{D}|H_0') \equiv P_{\theta_0}(\mathcal{D})$, while

$$P_1(\mathcal{D}) = \mathrm{pr}(\mathcal{D}|H_1') \equiv \int_{\Theta_1} P_\theta(\mathcal{D})\pi(\theta)d\theta. \qquad (3\cdot 1)$$

For $i = 0, 1$, let $F_i(\cdot)$ denote the distribution function of B_N: $F_i(b) \equiv P_i(B_N \leq b)$, $b \in \mathbb{R}$. Wherever they exist, we write F_i^{-1} for the inverse function of F_i, $i = 0, 1$, and denote

$$\psi(s) = F_0^{-1}(1 - F_1(s)) \quad \text{and} \quad \psi^{-1}(s) = F_1^{-1}(1 - F_0(s)). \qquad (3\cdot 2)$$

Condition I. Assume that the range of B_N is of the form $\mathcal{B} = (R_L, R_U] \cup [A_L, A_U)$, where $R_U \leq 1 \leq A_L$, R_L could be zero, and A_U could be infinity. Furthermore, assume that ψ exists on $(R_L, R_U]$ and ψ^{-1} exists on $[A_L, A_U)$. Since we are dealing with continuous densities, this condition will be satisfied by all but very strange stopping rules. Here are few examples.

Example 1. A standard "open-ended" stopping rule, familiar from the SPRT, is

$$N = \min\{\, n \geq 1 : \quad B_n \notin (R, A)\}, \qquad (3\cdot 3)$$

where $R < 1 < A$. If the B_n can range from zero to infinity, it is easy to see that $(R_L, R_U] = (0, R]$ and $[A_L, A_U) = [A, \infty)$, and the remaining part of Condition I can be easily verified.

Example 2. With N as in $(3\cdot 3)$, consider the truncated at m stopping time $N_m = \min(N, m)$ (see also Section $(1\cdot 2)$). Clearly, this is a proper stopping time. Since the range of B_m must include 1, so must that of B_N; hence $R_U = 1 = A_L$. For the examples in this paper, the range of B_n is of the form $(0, C_n)$, where the C_n are increasing in n. Thus $R_L = 0$ and $A_U = C_m$ and the range of B_N is therefore $\mathcal{B} = (0, C_m)$. The remaining part of Condition I can be easily verified.

Example 3. Variants on the stopping time $(3\cdot 3)$, for which the same conclusion applies, are: *(i)* $N_m^1 = \min\{n \geq m : \ B_n \notin (R, A)\}$, of a "two-stage" study (with

an initial sample of $m > 1$ observations taken in a single batch and then followed by a sequential sampling), and *(ii)* $N_m^2 = m \cdot \inf\{n \geq 1 : \quad B_{nm} \notin (R, A)\}$, of a "group-sequential" study which samples (sequentially) m units at-a-time.

Corresponding to the constants $R_{_U}$ and $A_{_L}$ defined above, let

$$r = \min(R_{_U}, \psi^{-1}(A_{_L})), \qquad a = \max(A_{_L}, \psi(R_{_U})). \tag{3 · 4}$$

These constants, r and a, define the "decision boundaries" for the modified Bayes-sequential test, T^*, as follows:

$$\begin{cases} \text{if } B_N \leq r, & \text{reject } H_0, \text{ and report the CEP } \alpha^*(B_N) = B_N/(1 + B_N), \\ \text{if } r < B_N < a, & \text{make no decision (though the experiment is stopped),} \\ \text{if } B_N \geq a, & \text{accept } H_0, \text{ and report the CEP } \beta^*(B_N) = 1/(1 + B_N). \end{cases} \tag{3 · 5}$$

In Theorem 1 below, we show that the modified Bayes-sequential test, T^*, indeed defines a valid sequential test for the Conditional Frequentist. The theorem also exhibits the extent of agreement between the Bayesian and Conditional Frequentist interpretations of T^*. The test arises from conditioning on the statistic

$$S(B_N) = \min(B_N, \psi^{-1}(B_N)), \tag{3 · 6}$$

defined over the domain $\mathcal{B}_S = \{b : b \in \mathcal{B}; \quad 0 \leq S(b) \leq r\}$. (For a discussion of this specific choice of the conditioning statistic, see Berger, Boukai & Wang (1997).) Observe that the complement of $\mathcal{B}_S$ is the *no decision region*, and can be viewed as the part of the sample space over which agreement between Bayesians and Frequentists cannot be achieved. From (3·2), (3·4) and (3·5), it is clear that the no-decision region disappears if the stopping rule is chosen so that $F_1(R_{_U}) + F_0(A_{_L}) = 1$. This can virtually always be achieved, if desired. Since T^* rejects H_0 if $B_N \leq r$ and accepts H_0 if $B_N \geq a$, it follows immediately that the conditional error probabilities are

$$\begin{aligned} \alpha(s) &= \text{pr}(B_N \leq r | \theta_0, S(B_N) = s) \\ \beta(s|\theta) &= \text{pr}(B_N \geq a | \theta, S(B_N) = s), \ \theta \in \Theta_1. \end{aligned} \tag{3 · 7}$$

The proof of the following theorem is given in the Appendix.

THEOREM 1. *Consider the test T^*, of the simple versus composite hypotheses (2·1), under Condition I and with the conditioning statistic $S(B_N)$ given in (3·6). Then, in the rejection and acceptance regions, respectively,*

$$\alpha^*(B_N) = \alpha(s) \qquad \text{and} \qquad \beta^*(B_N) = \int_{\Theta_1} \beta(s|\theta)\pi(\theta|s)d\theta, \tag{3·8}$$

where $\pi(\theta|s)$ denotes the posterior density of θ conditional on H_1 being true and on the observed value s of $S(B_N)$.

The first equality in (3·8) provides the key result that the conditional Type I error probability in (3·7) and the posterior probability of H_0 in (2·5) are equal. The second equality states that $\beta^*(B_N)$ in (2·6) (the posterior probability of H_1) is the average of the conditional Frequentist Type II error probability, $\beta(s|\theta)$ given in (3·7), with respect to the "weight function" provided by the conditional posterior density of θ given $S(B_N) = s$. This "averaging" of the conditional probability of Type II error is rather appealing, especially in comparison with the common classical practice of simply picking a plausible value of θ and reporting the power at that value. Nevertheless, one could utilize T^* in this latter fashion if desired, or indeed report the entire function $\beta(s|\theta)$.

Note that application of the modified Bayesian test T^* is straightforward. The only computational challenge is determination of the constants r and a in (3·4), defining the decision boundaries for T^*, through solution of $F_0(a) = 1 - F_1(r)$. For the default conditioning statistics that we recommend, it seems to virtually always be the case that $r = 1$, so that only a need be computed. Furthermore, as was mentioned in Section 1·3, one often does not need to determine a explicitly in order to use T^*. Finally, even if it does become necessary to consider a (i.e., one has stopped experimentation and $B_N > 1$), then checking whether or not $B_N > a$ is equivalent to checking whether or not $F_0(B_N) > 1 - F_1(R_U)$, and the latter may be comparatively easy to determine.

4. Application to normal testing of a precise hypothesis

In this section we illustrate the application of the proposed conditional sequential testing procedure to the "two-sided" normal testing problem. Let $X_1, X_2, \cdots$ be a sequence of independent and identically distributed $\mathcal{N}(\theta, \sigma^2)$ random variables and consider testing

$$H_0 : \theta = \theta_0 \qquad \text{versus} \qquad H_1 : \theta \neq \theta_0. \tag{4·1}$$

Case 1: Consider first the case of known σ^2. Here, $\Theta = I\!R$ and we assume that over $\Theta_1 \equiv \{\theta \neq \theta_0\}$, θ has the (conjugate) prior $\pi(\theta) = \mathcal{N}(\theta_0, \xi\sigma^2)$, for some fixed $\xi > 0$. For each $n = 1, 2, \ldots$, let Z_n be the standard test statistic $Z_n = \sqrt{n}(\bar{X}_n - \theta_0)/\sigma$, where $\bar{X}_n$ denotes the average of the first n observations. It is convenient to consider $Z_1, Z_2, \ldots$ as the sequential observations. For each $n \geq 1$, the Bayes factor, (2·4), is $B_n = (1+\xi n)^{\frac{1}{2}} \exp\{-\xi n z_n^2/2(1+\xi n)\}$.

For some predetermined stopping boundaries R and A ($R < 1 < A$), we consider the open-ended stopping time N as in (3·3). It can be easily verified that $B_n \to \infty$ (almost surely) under P_{θ_0} and $B_n \to 0$ (almost surely) under P_θ for all $\theta \neq \theta_0$; thus the stopping time is proper. To verify Condition I, note that each B_n ranges between 0 and $(1+\xi n)^{1/2}$. Since n can be arbitrarily large under the open-ended stopping rule, it is easy to see that the range of B_N is $\mathcal{B} = (0, R] \cup [A, \infty)$, and that Condition I holds over this domain. The test T^* in (3·5) can thus be applied with "decision boundaries" r and a as determined by (3·4). Unless R is close to 1, it will be the case that $F_1(R) + F_0(A) < 1$, so that $r = R$ and $a = \psi(R)$.

In Table 1 below we compute a for the common default choice of $\xi = 2$ and various values of R and A. Also provided are other quantities of potential interest, including the unconditional probabilities of Type I and Type II errors, $\alpha' = \text{pr}(B_N < r | H_0')$ and $\beta' = \text{pr}(B_N > a | H_1')$, respectively; the expected stopping times $(E_0(N), E_1(N))$ and variances $(var_0(N), var_1(N))$ under H_0' and H_1', respectively; and the corresponding probabilities of *no-decision*, $p_0 = \text{pr}(r < B_N < a | H_0')$

and $p_1 = \text{pr}(r < B_N < a|H_1')$. The numerical evaluation of these unconditional quantities was carried out by simulation (see for example, Siegmund, 1985, Remark 2.26), based on $M = 10^5$ runs of the sequential process under H_0. It should be emphasized that α' and β' are given here for design or evaluation purposes only. For the test T^*, the reported error probabilities will still be $\alpha^*(B_N)$ and $\beta^*(B_N)$.

Perhaps the most interesting feature of Table 1 is the tradeoff between p_0, the probability under H_0 of stopping in the no-decision region, and the expected stopping times; if the expected stopping times are small, the evidence is likely to be weak and, hence, one is more likely to be in the no-decision region. The expected stopping time under H_0, $E_0(N)$, is particularly large, and may be of special concern in planning the experiment.

Table 1. Two-sided normal sequential testing with known σ^2

						$R = 0.1$			
A	a	α'	β'	p_0	p_1	$E_0(N)$	$E_1(N)$	$var_0(N)$	$var_1(N)$
3.0	3.176	0.034	0.207	0.254	0.082	8.6	7.6	66	84
4.0	4.084	0.039	0.177	0.185	0.046	14.5	10.9	212	289
5.0	5.075	0.042	0.153	0.137	0.027	22.5	14.2	611	616
6.0	6.043	0.044	0.134	0.108	0.018	31.8	18.0	1168	1154
7.0	7.040	0.046	0.118	0.084	0.012	43.0	21.6	2083	2120
8.0	8.025	0.047	0.106	0.067	0.008	55.8	25.2	3817	2947
9.0	9.023	0.048	0.096	0.054	0.006	70.4	29.6	6072	6000
10.0	10.015	0.049	0.088	0.043	0.004	86.8	33.7	10550	8548
						$R = 0.05$			
3.0	3.187	0.017	0.204	0.276	0.089	9.0	8.9	106	163
4.0	4.091	0.019	0.176	0.209	0.051	15.3	12.9	344	461
5.0	5.087	0.021	0.152	0.163	0.032	23.7	17.1	983	1327
6.0	6.052	0.022	0.133	0.133	0.022	33.5	21.3	1632	2182
7.0	7.051	0.023	0.118	0.110	0.016	45.4	26.3	2962	3605
8.0	8.034	0.023	0.105	0.095	0.012	58.8	29.9	5272	4440
9.0	9.033	0.024	0.096	0.081	0.009	74.0	35.3	8791	8182
10.0	10.023	0.025	0.088	0.070	0.007	91.2	39.9	13904	12738

Case 2: We continue with the sequential test of (4·1), but now assume that both θ and σ^2 are unknown. For the construction of the Bayesian-Frequentist sequential test for this case, we proceed along the lines of Example 5 in Berger, Boukai & Wang

(1997), using a hierarchical prior structure defined as follows. For the first-stage prior distribution of θ, take $\pi_1(\theta|\sigma^2,\xi) = \mathcal{N}(\theta_0, \xi\sigma^2)$. For the second-stage prior of (σ^2,ξ), take $\pi_2(\sigma^2,\xi) = \sigma^{-2}g(\xi)d\sigma^2 d\xi$, where $g(\cdot)$ is some proper prior density for $\xi > 0$ (to be specified later). Straightforward computation yields, for $n \geq 2$,

$$B_n = (n-1+y_n)^{-n/2} \times \left[\int_0^\infty \frac{(1+n\xi)^{(n-1)/2}}{[(n-1)(1+n\xi)+y_n]^{n/2}} g(\xi)d\xi\right]^{-1}, \qquad (4\cdot 2)$$

where $y_n = n(\bar{x}-\theta_0)^2/S_n^2$ and S_n^2 is the usual sample variance.

Instead of analyzing $x_1, x_2, \ldots$, we will assume we are analyzing the sequential t-statistics $y_2, y_3, \ldots$. For each $n \geq 2$, we write the density of y_n as $f(y_n|\mu)$, where $\mu = (\theta-\theta_0)/\sigma$. Then the test can be rewritten as a test of $H_0 : \mu = 0$, which is a simple hypothesis. Furthermore, under H_1, the hierarchical prior defined earlier becomes: $\pi_1(\mu|\xi)$ is $\mathcal{N}(0,\xi)$, while $\xi > 0$ still has proper prior $g(\xi)$. The implied prior, $\pi(\mu) = \int \pi_1(\mu|\xi)g(\xi)d\xi$, is thus proper. Note that, in this case, the marginal density of y_n under H_0 and H_1, respectively, becomes $m_{0,n}(y_n) = m(y_n|0)$ and $m_{1,n}(y_n) = \int m(y_n|\xi)g(\xi)d\xi$, where $K_n = \Gamma(\frac{n}{2})(n-1)^{(n-1)/2}/\Gamma(\frac{1}{2})\Gamma(\frac{n}{2}-\frac{1}{2})$ and

$$m(y_n|\xi) = \int f(y_n|\mu)\pi_1(\mu|\xi)d\mu = K_n \frac{{y_n}^{-1/2}(1+n\xi)^{(n-1)/2}}{[(n-1)(1+n\xi)+y_n]^{n/2}}. \qquad (4\cdot 3)$$

For this testing problem, we recommend (see Berger, Boukai, & Wang (1997)) the prior

$$g(\xi) = (2\pi)^{-\frac{1}{2}}\xi^{-\frac{3}{2}}\exp\{-\frac{1}{2\xi}\}, \qquad \xi > 0. \qquad (4\cdot 4)$$

For some predetermined stopping boundaries R and A $(R < 1 < A)$, we consider, as in Section $1\cdot 2$, the truncated at m stopping time N_m, discussed in Example 2. Along the same lines, it can be easily verified that the test T^* in $(3\cdot 5)$ applies here with $R_U = A_L = 1$, so that, by $(3\cdot 4)$, a satisfies the equation

$$F_0(a) = 1 - F_1(1). \qquad (4\cdot 5)$$

Table 2 presents the numerical evaluation of a, as was determined by $M = 10^4$ simulation runs for selected boundaries R and A and various choices of the

truncation value m. As in Table 1, other potentially interesting unconditional quantities are also presented.

Table 2. *Truncated two-sided normal sequential testing with unknown* σ^2

A	m	a	α'	β'	p_0	p_1	$E_0(N_m)$	$E_1(N_m)$	$var_0(N_m)$	$var_1(N_m)$
					$R = 0.1$					
8	50	3.670	0.048	0.116	0.129	0.059	44.4	16.6	92	393
	100	6.050	0.042	0.103	0.094	0.035	56.2	21.9	551	936
	200	8.014	0.038	0.104	0.084	0.017	62.7	25.9	1581	1845
	300	8.016	0.035	0.104	0.083	0.013	64.9	27.0	2330	2428
9	50	3.640	0.049	0.117	0.128	0.059	48.0	17.3	83	420
	100	4.540	0.045	0.097	0.091	0.036	67.3	23.6	517	1079
	200	9.008	0.040	0.094	0.073	0.019	77.7	28.5	1901	2208
	300	9.015	0.037	0.094	0.071	0.012	80.8	30.4	3005	3103
10	50	3.630	0.048	0.117	0.129	0.060	48.1	17.2	78	421
	100	4.010	0.046	0.095	0.089	0.039	77.6	24.8	453	1195
	200	7.850	0.042	0.086	0.065	0.021	92.7	31.3	2070	2644
	300	10.008	0.040	0.086	0.061	0.013	97.9	34.8	3651	3960
					$R = 0.05$					
8	50	3.775	0.039	0.115	0.139	0.063	44.8	17.9	81	416
	100	6.075	0.031	0.103	0.108	0.038	57.1	123.9	545	1025
	200	8.016	0.023	0.104	0.101	0.020	64.1	29.0	1648	2173
	300	8.018	0.020	0.104	0.100	0.016	66.8	30.4	2577	2940
9	50	3.795	0.040	0.115	0.138	0.062	48.0	18.3	69	437
	100	4.875	0.033	0.096	0.106	0.042	68.3	26.1	498	1176
	200	9.013	0.025	0.094	0.092	0.021	55.1	31.8	1933	2585
	300	9.019	0.021	0.094	0.089	0.015	55.1	35.3	3216	3913
10	50	3.775	0.039	0.115	0.141	0.064	48.5	18.4	63	435
	100	4.345	0.034	0.093	0.102	0.044	78.6	27.5	411	1263
	200	10.003	0.026	0.086	0.084	0.024	94.3	35.2	2069	3027
	300	10.011	0.023	0.086	0.078	0.016	100.6	39.5	3847	4759

One interesting feature of the choice of m is that, for smaller m (50 or 100), it happens that $a < A$. For such m, the *no decision region* can occur only at the truncation time m (see Figure 1). When m is larger, however, the *no decision region* could come into play even when one stops by crossing the A boundary.

5. Comparison with unconditional sequential testing

We have previously discussed general issues in comparison of conditional and unconditional sequential testing, such as ease of use and generality of application. Here

we consider explicit comparison of the new sequential test with two unconditional tests in the literature, Siegmund's (1977) Repeated t-test, and the Wald SPRT.

Siegmund's Repeated t-test was designed for the normal mean testing situation of Case 2 in Section 4. Define, for a fixed $\xi > 0$ and each $n \geq 2$, $L_n^*(\xi) = m(y_n|\xi)/m(y_n|0)$, where $m(y_n|\cdot)$ is as given in (4·3). (Note that $1/L_n^*$ would be the Bayes factor in (4·2) were $g(\cdot)$ chosen to be a point mass.) Siegmund's Repeated t-test is based on the stopping time $N_m^* \equiv \min(N^*, m)$, where $N^* = \inf\{n \geq 2;\ L_n^*(\xi) > (1+\xi n)^{-1/2}\exp(k);\ \text{else } n = \infty\}$, for some $k > 0$; the test rejects H_0 if $N^* \leq m$ and accepts H_0 if $N^* > m$. Table 3 gives the unconditional properties of our new testing procedure and the Type I error of the repeated t-test (in parentheses), for some situations presented in Table 2 of Siegmund (1977). For better comparability, we based the new sequential test on the stopping time N_m^* above. Note that both tests will thus have the same expected stopping time.

Table 3. *Properties of the new sequential test and the repeated t-test.*

m	k	$1/\xi$	r	a	α'	β'	p_0	p_1	$E_0(N_m^*)$	$E_1(N_m^*)$
50	3	1	1	2·590	0·1060 (0·0981)	0·1297	0·0571	0·03270	46·96	15·05
100	3	1	0·590	1	0·1199 (0·1239)	0·1173	0·0061	0·0089	91·45	22·11
50	3	2	1	3·060	0·0852 (0·0741)	0·1245	0·0818	0·0419	47·93	15·65
100	3	2	1	2·249	0·0975 (0·0951)	0·1056	0·0246	0·0154	94·00	22·59
150	3	2	0·620	1	0·1038 (0·116)	0·0989	0·0123	0·0178	138·10	28·26
100	3	4	1	2·820	0·0828 (0·0793)	0·1019	0·0415	0·0216	95·92	24·30
150	3	4	1	2·019	0·0903 (0·0890)	0·0936	0·0121	0·0079	141·66	29·31
200	3	4	0·670	1	0·0922 (0·1040)	0·0874	0·0125	0·0167	186·23	34·62
100	4	1	1	3·600	0·0577 (0·0459)	0·0972	0·0750	0·0351	96·95	25·94
200	4	1	1	3·000	0·0634 (0·0590)	0·0798	0·0313	0·0173	191·47	38·45
200	4	2	1	3·159	0·0526 (0·0477)	0·0771	0·0475	0·0221	193·55	39·39
400	4	2	1	2·719	0·0617 (0·0579)	0·0704	0·0220	0·0124	383·33	62·89
200	4	4	1	4·000	0·0454 (0·0391)	0·0754	0·0561	0·0249	195·15	41·80
400	4	4	1	3·400	0·0528 (0·0504)	0·0675	0·0301	0·0146	386·10	63·81
200	4	8	1	4·310	0·0345 (0·0270)	0·0744	0·0694	0·0253	196·88	41·34
400	4	8	1	3·910	0·0404 (0·0398)	0·0657	0·0408	0·0181	389·82	65·63

Numbers in parentheses are for the repeated t-test.

While the unconditional behavior of the two tests appears to be similar, it is worthwhile to recall the basic point that unconditional reports of error can be quite

different from conditional reports. Consider, for instance, the case $m = 100$, $\xi = 1$, $k = 3$, and suppose that $N^* = 100$ with $y_n = 6 \cdot 3$. Then the repeated t-test rejects, reporting unconditional Type I error of $0 \cdot 12$, while the conditional test rejects but reports the much larger conditional error $0 \cdot 32$. On the other hand, if $y_n = 12.3$ happened to be observed, both tests reject and the unconditional Type I error probability of the repeated t-test remains unchanged, but the conditional test reports (the much smaller) conditional error probability of $0 \cdot 026$. Note, however, that these situations are somewhat atypical, and the repeated t-test will usually have reasonable performance.

The next comparison is with the Wald (1947) SPRT (with stopping boundaries $R < 1 < A$), as applied to the testing of (2·3). Since the hypotheses in (2·3) are simple, one can approximate the probabilities of Type I and Type II error, α' and β', by the usual Wald approximations $\widetilde{\alpha} \equiv R(A-1)/(A-R)$ and $\widetilde{\beta} \equiv (1-R)/(A-R)$. Furthermore, it is straightforward to show that β' is related to the Type II error probability for the composite hypothesis H_1 in (2·1), $\beta(\theta) = \mathrm{pr}(B_N \geq A|\theta)$, $\theta \in \Theta_1$, by the expression

$$\beta' = \int_{\Theta_1} \beta(\theta)\pi(\theta)d\theta. \qquad (5 \cdot 1)$$

This Wald test has interesting similarities with T^* for the open-ended stopping time in (3·3), especially if R and A are chosen so that the no-decision region is absent. Indeed, then T^* makes exactly the same decisions as does the Wald test, and must therefore have the same unconditional properties (α', β', $E_0(N)$, $E_1(N)$, etc.). Furthermore, T^* would then inherit any of the (unconditional) optimality properties of the Wald test, and one could even use the Wald approximation for design decisions.

In spite of these similarities, there are still several fundamental differences between the procedures. The most obvious difference is the use of $\pi(\theta)$ in (5·1), instead of $\pi(\theta|s)$ as in (3·8), to compute the average reported probability of Type II error.

Pre-experimentally, use of $\pi(\theta)$ is, of course, necessary but, after seeing the data, use of the conditional $\pi(\theta|s)$ would seem more natural.

The differences in reported error probability (of both Type I and Type II) can also be considerable if there is substantial "overshoot" of the boundaries A or R. The (approximate) Wald error probabilities ignore this overshoot, while the conditional error probabilities do not. Interestingly, if the overshoot is ignorable, the Wald test will report essentially the same Type I error probability as will T^*, indicating that the Wald test then possesses reasonable conditional performance.

Our final comment concerning the comparison of T^* with standard sequential analysis is a bit more speculative, but its potential practical importance justifies the speculation. Many experimenters will acknowledge that they begin experimentation without a clear idea as to when they will stop, and indeed that they monitor the data as it becomes available to help decide when they will stop. Without a clear pre-experimental stopping rule, however, unconditional frequentist analysis is formally precluded, and what is often done is to simply ignore the issue and pretend that the final sample arose from a fixed sample size experiment. This can result in highly misleading inference. (Note that one can sometimes address such problems from an unconditional frequentist perspective, if it can be ascertained that the stopping time was bounded by a given horizon; see Siegmund (1977) for one such analysis.)

Consider, now, use of T^* in such situations. The reported conditional error probabilities, $\alpha^*(B_N)$ and $\beta^*(B_N)$, do not depend on the stopping rule, and so the only possible "mistake" in having failed to prespecify the stopping rule is that one might actually be in the *no decision region*, without being aware of it. We have seen considerable evidence, however, that this is unlikely to happen. Furthermore, $\beta^*(B_N)$ (or $\alpha^*(B_N)$), for B_N in the *no decision region* will tend to be quite large, and reporting a large error instead of "no decision" will do little harm. The new sequential test, T^*, would thus seem to be quite safe for use in situations in which

the stopping rule is unknown or ill-specified.

Acknowledgement

This work was partially supported by the United States National Science Foundation. We thank the Editor, Associate Editor and a referee for comments which led to substantial improvement in the paper.

Appendix

Proof of Theorem 1

We consider only the case $A_L < \psi(R_U)$ in $(3\cdot4)$, so that $r = R_U$ and $a = \psi(R_U)$ in T^*. The other cases follow similarly. Let f_i^* denote the probability density function of B_N under H_i', $i = 0, 1$; let f_θ^* be the conditional probability density function of B_N given $\theta \in \Theta$ (under P_θ); and define $\mathcal{D}_n(b) = \{\underset{\sim}{x}_n : N = n \text{ and } B_n \leq b\}$. From $(2\cdot2)$, $(2\cdot3)$ and $(3\cdot1)$ and the fact that $\pi(\cdot)$ is a proper probability density function over Θ_1, repeated use of Fubini's theorem shows that

$$\int_0^b f_1^*(y)dy = \int_0^b \int_{\Theta_1} f_\theta^*(y)\pi(\theta)d\theta dy.$$

Hence, for $b \in int\,\mathcal{B}$,

$$f_1^*(b) = \int_{\Theta_1} f_\theta^*(b)\pi(\theta)d\theta. \tag{$A\cdot1$}$$

By $(3\cdot1)$ and Wald's Likelihood Ratio Identity (see for example, Siegmund 1985, Proposition 2.24),

$$f_0^*(b) = bf_1^*(b). \tag{$A\cdot2$}$$

Moreover, a direct calculation (see $(3\cdot2)$) yields

$$\psi'(s) \equiv \frac{d}{ds}\psi(s) = -\frac{f_1^*(s)}{f_0^*(\psi(s))}, \tag{$A\cdot3$}$$

for each $R_L < s < R_U$. Using $(A\cdot2)$, $(A\cdot3)$, $(3\cdot2)$, $(3\cdot6)$ and $(3\cdot7)$ yields

$$\alpha(s) = P_0(B_N \leq r | B_N = s \text{ or } \psi(s)) = \frac{f_0^*(s)}{f_0^*(s) + f_0^*(\psi(s))|\psi'(s)|} = \frac{s}{1+s}.$$

But B_N equals s on the set $\{B_N \leq r \text{ and } S(B_N) = s\}$, which, in view of $(2\cdot5)$, completes the proof of the first assertion. Next, observe (see $(3\cdot6)$, $(3\cdot7)$) that

$$\beta(s|\theta) = P_\theta(B_N \geq a | B_N = s \text{ or } \psi(s)) = \frac{f^*_\theta(\psi(s))|\psi'(s)|}{f^*_\theta(s) + f^*_\theta(\psi(s))|\psi'(s)|}. \qquad (A\cdot4)$$

Moreover, utilizing $(\mathrm{A}\cdot1)$, it is straightforward to verify that the posterior density of θ, conditional on knowing only that $\{S(B_N) = s\}$, is

$$\pi(\theta|s) \equiv \pi(\theta|S(B_N) = s) = \frac{f^*_\theta(s) + f^*_\theta(\psi(s))|\psi'(s)|}{f^*_1(s) + f^*_1(\psi(s))|\psi'(s)|}. \qquad (A\cdot5)$$

Combining $(\mathrm{A}\cdot4)$ and $(\mathrm{A}\cdot5)$ and utilizing $(\mathrm{A}\cdot3)$ yields

$$\int_{\Theta_1} \beta(s|\theta)\pi(\theta|s)d\theta = \frac{f^*_1(\psi(s))|\psi'(s)|}{f^*_1(s) + f^*_1(\psi(s))|\psi'(s)|} = \frac{1}{1+\psi(s)}.$$

Since $B_N = \psi(s)$ on the set $\{B_N \geq a \text{ and } S(B_N) = s\}$, this final expression is the posterior probability of H_1 given the data (see $(2\cdot6)$), completing the proof.

References

Armitage, P. (1975). *Sequential Medical Trials*, 2nd Ed. New York: John Wiley.

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

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

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

Ghosh, B. K. (1970). *Sequential Tests of Statistical Hypotheses.* Cambridge Mass.: Addison-Wesley Publishing.

Jeffreys, H. (1961). *Theory of Probability*, 3rd ed.. London: Oxford University Press.

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

Siegmund, D. (1977). Repeated significance tests for a normal mean. *Biometrika*, **64**, 177–189.

Siegmund, D. (1985). *Sequential Analysis.* New York: Springer-Verlag.

Wald, A. (1947). *Sequential Analysis.* New York: John Wiley.

Woodroofe, M. (1982). *Nonlinear Renewal Theory In Sequential Analysis.* Philadelphia: Society for Industrial and Applied Mathematics.