The Annals of Statistics
1980, Vol. 8, No. 6, 1219–1228

A SEQUENTIAL CLINICAL TRIAL FOR TESTING $p_1 = p_2$

BY D. SIEGMUND[1] AND P. GREGORY[2]

Stanford University

Sequential designs are proposed for clinical trials to compare two binomial success probabilities, p_1 and p_2. Approximations to the operating characteristics and expected sample size are obtained and compared with simulations. Special reference is made to the problem of comparing vasopressin and placebo for stopping upper gastrointestinal hemorrhage.

1. Introduction. This paper presents a class of sequential tests for comparing two binomial success probabilities p_1 and p_2. Although the method is a general one, it was motivated by a clinical trial for testing the efficacy of vasopressin (a hormone which constricts blood vessels) in stopping upper gastrointestinal hemorrhage; and although this report is primarily theoretical, reference will be made to the trial of vasopressin because it seems to illustrate clearly certain advantages and disadvantages of a sequential design in clinical trials.

Two conditions indicating a sequential design are (i) a serious disease, so that ethical considerations mandate the early termination of a trial in which one treatment appears especially effective, and (ii) a response time which is short compared to the time between patient arrivals, so that it is feasible to evaluate the current state of affairs before admitting new patients to the study. Massive upper gastrointestinal hemorrhage satisfies these requirements, since failure to control it within hours may lead to death or to surgical intervention.

One other circumstance which seems to indicate a sequential trial in this particular case is the existence of earlier, favorable reports on the use of intra-arterial vasopressin for stopping upper gastrointestinal hemorrhage. (See especially Conn et al., 1975.) Although certain reservations concerning these earlier trials and the desire to investigate a much simpler intravenous mode of administration of the drug suggest a new trial, a sequential design provides protection against the lengthy continuation of this trial, should the previous, favorable results be repeated.

For purposes of sequential analysis this trial is one of comparing the probability p_1 of success using vasopressin to the probability p_2 of success with placebo. Very tentative figures from previous studies indicate that both the spontaneous remission rate and the success rate with vasopressin vary with cause of bleeding. While this suggests consideration of a more elaborate model involving some kind of stratification, for the relatively small experiment envisaged here the advantage of stratification appears to be negligible. For a more thorough discussion of this point, see Siegmund and Gregory (1979).

The definition of success is somewhat arbitrary. Here it is defined as a cessation of bleeding within five hours and no recurrence within six. Other endpoints of interest are the time until bleeding initially ceases, recurrence of bleeding, severity of bleeding measured by transfusion requirements, the need for surgical intervention, and death. Choice of a sequetial design for testing $p_1 = p_2$ makes the implicit assumption that if p_1 appears to be considerably larger than p_2, that by itself is sufficient to terminate the trial and indicate the use of vasopressin. In practice one would probably be reluctant to terminate early unless the other factors also consistently favor vasopressin, although it would defeat the purpose of a sequential trial to

Received February 1979; revised July 1979.

[1] Research supported by ONR Contract N00014-77-C-0306 (NR-042-373) and by NSF Grant MCS77-16974

[2] Research supported by NIH Grant AM 19976

AMS 1970 *subject classification.* Primary 62L10.

Key words and phrases. Sequential test, clinical trial, stopping rule.

insist that these factors show "statistically significant" differences between treatment and control. Conversely, in the absence of a strong indication for vasopressin based on success rate alone, it may be desirable to analyze other factors rather carefully. For example, if vasopressin reduces bleeding sufficiently that surgery need not be performed on an emergency basis, it would be a useful treatment, although its "success" rate may be no higher than for placebo. The extent to which a sequential design introduces a bias which makes these other analyses difficult is perhaps its most serious disadvantage.

In addition to studying the specific problem of testing $p_1 = p_2$, a primary goal of this paper is to indicate how arguments developed by Siegmund (1977, 1978) to deal with normally distributed data can be adapted and supplemented by simulations to obtain a reasonably clear understanding of a similar but more difficult problem. Section 2 reviews pertinent material for an analogous problem with normal data. A modification of the repeated significance tests advocated by Armitage (1975) is suggested and their properties studied. Section 3 returns to the problem of testing $p_1 = p_2$. Mathematical results are collected in three appendices.

2. Normal Data With Known Variance. The normal distribution with known variance is relatively simple both conceptually and technically and suggests useful approximations for more complex situations. In this section known results for the repeated significance tests advocated by Armitage (e.g., Armitage, 1975) are reviewed and a modification of these tests suggested and studied.

The simplest situation occurs in a paired comparison design, in which for each $n = 1, 2, \cdots$ the observation x_n represents the difference in response of the nth pair of subjects, one of whom receives treatment A and the other treatment B. It is assumed that the x_n are independent and normally distributed with expectation μ and known variance σ^2. Let $s_n = x_1 + \cdots + x_n$, and given $b_1 > 0$ and $m_0 = 1, 2, \cdots$ define

$$T_1 = \text{first} \quad n \geq m_0 \quad \text{such that} \quad |s_n| > b_1 \sigma n^{1/2}. \tag{1}$$

Let $m_1 > m_0$ be a positive integer. The sequential test of $H_0: \mu = 0$ against $H_1: \mu \neq 0$ which terminates sampling at $\min(T_1, m_1)$ and rejects H_0 if and only if $T_1 \leq m_1$ is the repeated significance test of Armitage (1975).

Let $\theta = \mu/\sigma$. The distribution of T_1 and hence the power function of this test depend on μ and σ only through the value of θ. By repeated numerical integration McPherson and Armitage 1971)—see also Armitage (1975)—have provided tables which allow one to choose the design parameters m_1 and b_1 to attain a specified significance level $\alpha = P_0\{T_1 \leq m_1\}$ and power $1 - \beta = P_{\theta_1}\{T_1 \leq m_1\}$ at a given value $\theta_1 \neq 0$. Accurate analytic approximations to α and β were given by Siegmund (1977, 1978)—see Appendix A for a summary of the pertinent results adapted to the present requirements.

A class of modified repeated significance tests which interpolate the fixed sample size and repeated significance tests have been suggested independently by Peto et al. (1976) and Siegmund (1978), but their properties have not been studied. Let $0 < c \leq b$ and $m_0 \leq m$ be given, and let T be defined by (1) with b in place of b_1. Stop sampling at $\min(T, m)$ and reject H_0 if either $T \leq m$ or $T > m$ and $|s_m| > cm^{1/2}$. For fixed m_0 there are three parameters m, b, and c defining such a modified repeated significance test and hence there are many tests having a specified significance level and power at a given $\theta_1 \neq 0$. Relative to a given repeated significance test defined by m_1 and b_1, the corresponding modified tests have $m \leq m_1$ and $b \geq b_1$. The extreme case $b = \infty$ corresponds to a fixed sample size test with rejection region $|s_m| > cm^{1/2}$, whereas $c = b = b_1$ and $m = m_1$ give a repeated significance test. Figure 1 illustrates these relations.

Table 1 gives numerical examples illustrating various relations among fixed sample size, repeated significance tests, and the modified tests suggested here. The approximations given in Appendices A and B were used to perform the required calculations, except for those entries which could be obtained from Armitage (1975), page 104. (Simulations indicate that these approximations are quite accurate.) For all tests the over-all significance level is .05 and $m_0 = 1$. The entry m_f denotes that fixed sample size which would yield the same power at the

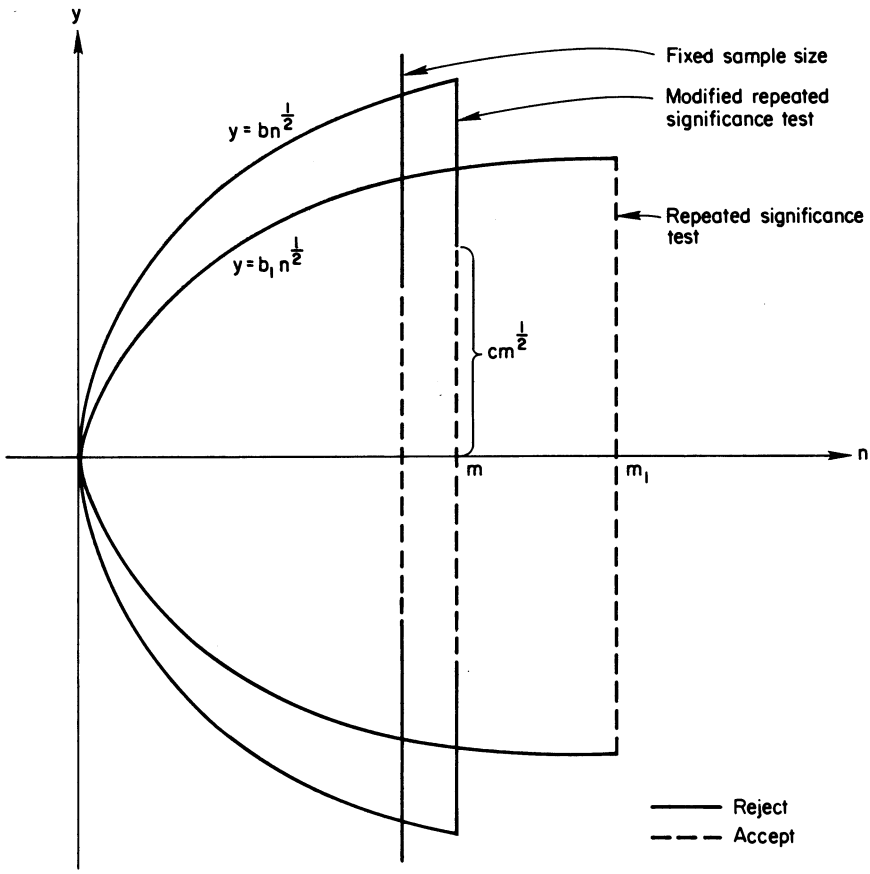

FIG. 1.

TABLE 1
Numerical comparisons

	Fixed Sample Size	Repeated Significance Test			Modified Test			
	$m = 49$	$m_1 = 49$, $b_1 = 2.8$			$m = 49$, $b = 3.15$, $c = 2.13$			
θ	$1 - \beta$	$1 - \beta$	$E_\theta(T_1 \wedge m_1)$	m_f	$1 - \beta$	$E_\theta(T \wedge m)$	m_f	$P_\theta(T \leq m)$
.6	.99	.95	21	37	.98	25	45	.90
.4	.80	.61	34	32	.75	39	44	.47
	$m = 111$	$m_1 = 111$, $b_1 = 2.89$			$m = 111$, $b = 3.25$, $c = 2.13$			
θ	$1 - \beta$	$1 - \beta$	$E_\theta(T_1 \wedge m_1)$	m_f	$1 - \beta$	$E_\theta(T \wedge m)$	m_f	$P_\theta(T \leq m)$
.4	.99	.95	47	82	.98	59	101	.88
.3	.88	.71	72	71	.85	87	100	.57

indicated θ as the given sequential test. For the modified tests b was chosen fairly large, so that $P_0\{T \leq m\}$ is slightly less than .02.

The most obvious appeal of these modified repeated significance tests is that they provide insurance against a long trial should one treatment seem considerably superior without as

large a maximum sample size or as great a loss of power to detect smaller differences as the usual repeated significance tests.

Although difficult to quantify, the following additional arguments in favor of the modified tests seem to warrant some discussion.

(i) One disadvantage of the possible early termination of a sequential test is that it may prevent the accumulation of sufficient evidence against H_0 to be thoroughly convincing. For a repeated significance test, if $T_1 = n \leq m_1$, the observed significance level or P-value of the test may be defined as $P_0\{T_1 \leq n\}$. This is a common index of how convincing the data against H_0 are; but since $P_0\{T_1 \leq n\}$ is approximately proportional to $\log n$ (Siegmund, 1977, or Appendix A), even for n much smaller than m_1 it may not be appreciably smaller than the over-all significance level, $P_0\{T_1 \leq m_1\}$. For example, for the first repeated significance test in Table 1, which has $m_1 = 49$ and $\alpha = .05$, if $T_1 = 16$, the observed significance level is $P_0\{T_1 \leq 16\} \cong .032$. By way of contrast, for a modified test the over-all significance level is

$$\alpha = P_0\{T \leq m\} + P_0\{T > m,\ cm^{1/2} \leq |s_m| \leq bm^{1/2}\}. \tag{2}$$

If $T = n \leq m$, the observed significance $P_0\{T \leq n\}$ is no greater than $P_0\{T \leq m\}$, which may be made small by taking b large. For the first modified test in Table 1, $P_0\{T \leq 49\} \cong 0.18$ and $P_0\{T \leq 16\} \cong .011$. A similar argument applies to other indices of "convincingness," e.g., a lower confidence bound on $|\theta|$ (cf. Siegmund, 1978).

(ii) As was mentioned in the introduction, focusing on a single endpoint—here represented by the single parameter θ—is a convenient but potentially misleading simplification of a complex situation. By using a modified test with b large enough that $P_\theta\{T \leq m\}$ is small for small θ, one reduces the biasing effect of the stopping rule (Siegmund, 1978) and helps to insure that sufficient data will be available for more subtle comparisons when these seem advisable.

3. Comparing Two Binomials. Now suppose that the response of a patient assigned to treatment i ($i = 1, 2$) is success with probability p_i and failure with probability $q_i = 1 - p_i$ and is independent of other responses. To simplify the discussion it is assumed that observations are taken in pairs with one member of each pair assigned to treatment and the other to control. The biased coin design of Efron (1971) provides a reasonable scheme for approximating this situation while maintaining a high level of unpredictability as to exactly which treatment will be assigned to the next patient. With suitable modifications patients may easily be assigned to treatment or control in a 2 to 1 or other ratio.

More specifically, assume that the data consists of pairs $(x_1, y_1), (x_2, y_2), \cdots$, where the x's and y's are independent random variables assuming the values 1 and 0. Let $P\{x_n = 1\} = p_1$, $P\{y_n = 1\} = p_2$, $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$. It is desired to obtain a sequential test of $H_0: p_1 = p_2$ against $H_1: p_1 \neq p_2$ which on the average requires a small number of observations to reach a decision whenever p_1 and p_2 differ substantially.

Two obvious candidates are sequential versions of the generalized likelihood ratio test and of the χ^2 test for independence in 2×2 tables. As one might expect, these tests perform similarly; but there are slight differences.

Let $H(x) = x \log x + (1 - x)\log(1 - x)$ and $I(x, y) = H(x) + H(y) - 2H[\frac{1}{2}(x + y)]$. The log generalized likelihood ratio for testing $H_0: p_1 = p_2$ against $H_1: p_1 \neq p_2$ based on n pairs of observations is $l_n = nI(\bar{x}_n, \bar{y}_n)$, where $\bar{x}_n = n^{-1}\sum_1^n x_k$ and $\bar{y}_n = n^{-1}\sum_1^n y_k$. In analogy with (1), given integers $m_0 \leq m$ and real numbers $0 < c \leq b$, define

$$T = \text{first} \quad n \geq m_0 \quad \text{such that} \quad (2l_n)^{1/2} > b. \tag{3}$$

Stop at $\min(T, m)$ and reject H_0 if either $T \leq m$ or $T > m$ and $(2l_m)^{1/2} > c$.

A Taylor series expansion about (p_1, p_2) shows that $(2nl_n)^{1/2}$ behaves approximately like the absolute value of a sum of n independent identically distributed random variables with mean

$$\mu = [2I(p_1, p_2)]^{1/2} \tag{4}$$

and variance

$$\sigma^2 = \{p_1q_1\log^2(p_1\bar{q}/q_1\bar{p}) + p_2q_2\log^2(p_2\bar{q}/q_1\bar{p})\}/2I(p_1, p_2), \tag{5}$$

where $\bar{p} = (p_1 + p_2)/2$ and $\bar{q} = 1 - \bar{p}$. This suggests that under $H_1: p_1 \neq p_2$, one may approximate the behavior of $(2nl_n)^{1/2}$ by Brownian motion with drift μ and variance σ^2 for the purpose of computing the power and expected sample size of the test. The relevant formulas are summarized in Appendices A and B, and their accuracy is supported by Table 2 below. This approximation is consistent with the customary weak convergence under local alternatives theory, but appears to be much more accurate when $|p_1 - p_2|$ is not small.

For $p_1 = p_2$ the Brownian motion theory is not accurate because extreme tail probabilities are involved. Let $P_{(p_1,p_2)}$ denote the probability measure on sequences $(x_1, y_1), (x_2, y_2), \cdots$, determined by p_1 *and* p_2. The significance level of the modified repeated significance test defined above is for $p_1 = p_2 = p$

$$\alpha(p) = P_{(p,p)}\{T \leq m\} + P_{(p,p)}\{T > m, (2l_m)^{1/2} > c\}. \tag{6}$$

The second term on the right-hand side of (6) may be approximated by the upper bound $P_{(p,p)}\{c < (2l_m)^{1/2} \leq b\} \cong 2[\Phi(b) - \Phi(a)]$, since $2l_m$ is asymptotically χ^2 with one degree of freedom under H_0. In principle the methods of Lai and Siegmund (1977) may be adapted to give an asymptotic approximation to $P_{(p,p)}\{T \leq m\}$ as $b \to \infty$, $m \to \infty$, and $m_0 \to \infty$ in such a way that $bm^{-1/2} \to \theta_1$ and $bm_0^{-1/2} \to \theta_0$. A heuristic sketch of the rather elaborate computations is given in Appendix C. The resulting formula involves a numerical integration, which was easy in the case of normal random variables, but in this case remains a difficult unsolved problem. A further "no overshoot" approximation to this integral yields a crude but simple approximate upper bound: for $p \leq \frac{1}{2}$,

$$P_{(p,p)}\{T \leq m\} \lesssim (\pi^{-1}pqa)^{1/2}e^{-a} \cdot \int_{(0,2p) \cap \{\xi:\theta_0 < [2I(\xi,2p-\xi)]^{1/2} \leq \theta_1\}} [I(\xi, 2p - \xi)\xi(1 - \xi)(2p - \xi)(2q - 1 + \xi)]^{-1/2}\, d\xi. \tag{7}$$

TABLE 2
Sequential generalized likelihood ratio test

(p_1, p_2)*	$P\{T \leq m\}$	$P\{\text{Reject } H_0\}$	$E(T \wedge m)$
	Case I: $m_0 = 7$, $m = 49$, $b = 3.15$, $c = 2.15$		
.5, .5	.017 ± .001 (.023)	.045 ± .003 (.053)	48.5 ± .1
.7, .5	.238 (.231)	.474 (.451)	44.1 ± .4
.8, .5	.629 (.623)	.851 (.849)	35.7 ± .5 (34.7)
.4, .4	.019 ± .001 (.024)	.041 ± .002 (.054)	48.3 ± .1
.6, .4	.208 (.225)	.448 (.435)	44.3 ± .4
.7, .4	.578 (.571)	.827 (.816)	36.5 ± .5 (37.0)
.8, .4	.902 (.900)	.983 (.981)	25.8 ± .4 (24.9)
.3, .3	.018 ± .001 (.024)	.046 ± .003 (.054)	48.3 ± .1
.7, .3	.885 (.880)	.979 (.976)	25.9 ± .4 (25.8)
.2, .2	.016 ± .001 (.017)	.046 ± .003 (.047)	48.4 ± .1
	Case II: $m_0 = 10$, $m = 100$, $b = 3.2$, $c = 2.15$		
.5, .5	.018 ± .001 (.021)	.045 ± .004 (.051)	98.5 ± .3
.7, .5	.506 (.510)	.802 (.775)	79.0 ± .9 (74.5)
.8, .5	.948 (.945)	.995 (.992)	45.7 ± .8 (45.4)
.4, .4	.017 ± .001 (.021)	.044 ± .004 (.051)	98.5 ± .3
.6, .4	.479 (.496)	.761 (.757)	79.1 ± .9 (74.7)
.7, .4	.917 (.918)	.988 (.986)	51.4 ± .9 (48.4)
.8, .4	.998 (.998)	1.00 (1.00)	28.8 ± .5 (27.0)
.3, .3	.019 ± .001 (.024)	.046 ± .004 (.054)	99.1 ± .3
.7, .3	.998 (.998)	1.00 (1.00)	30.2 ± .6 (28.3)
.2, .2	.017 ± .001 (.021)	.035 ± .004 (.051)	98.9 ± .3

* The cases $(p_1, p_2) = (.5, .3), (.6, .3), (.5, .2)$, and $(.6, .2)$ are by symmetry the same as $(.7, .5), (.7, .4), (.8, .5)$, and $(.8, .4)$ respectively.

Table 2 gives two numerical examples, which are roughly comparable to the normal examples in Table 1. The first entry in each cell is a Monte Carlo estimate; the parenthetical entries are analytic approximations obtained from the suggestions of the preceding paragraphs together with (A.4), (A.5), and (B.4) of the appendices. The ± figures are one estimated standard error. Where no ± figure is given, the Monte Carlo estimate is a relative frequency r, the standard error of which may be estimated by the usual $r(1-r)/N$, where N is the number of repetitions of the experiment. Except for the probabilities $P_{(p,p)}\{T \le m\}$ and $\alpha(p)$, $N = 900$. For these probabilities $N = 5000$ and the method of importance sampling mentioned in Appendix C was used. Generally speaking, the analytic approximations are reasonably good except for the null hypothesis probabilities $P_{(p,p)}\{T \le m\}$ and $\alpha(p)$, for which they are too large as expected. The authors have performed other simulations and found the approximations to hold up over a wider range of parameter values than those reported here.

Because of the discreteness of the underlying data, the choice of m_0 can have a substantial effect on $P_{(p,p)}\{T \le m\}$. Taking m_0 about equal to $m^{1/2}$ seems reasonable and has the additional desirable property of making $P_{(p,p)}\{T \le m\}$ fairly constant as a function of p, at least for p not too near 0 or 1.

The customary χ^2 statistic for testing $p_1 = p_2$ is $\chi_n^2 = n(\bar{x}_n - \bar{y}_n)^2/2p_nq_n$, where $p_n = \frac{1}{2}(\bar{x}_n + \bar{y}_n)$ and $q_n = 1 - p_n$. A sequential test analogous to that discussed above may be defined by the stopping rule (3) with χ_n in place of $(2l_n)^{1/2}$. Again sampling terminates at $\min(T, m)$ and H_0 is rejected if either $T \le m$ or $T > m$ and $\chi_m > c$. Brownian motion approximations similar to those suggested above can be developed under H_1, but the authors have not obtained an analogous null hypothesis theory.

APPENDIX A

Probability approximations for normal data

Let $x_1, x_2, \cdots$, be independent normally distributed random variables with mean θ and variance 1. Let $s_n = x_1 + \cdots + x_n$ and

$$T = \text{first} \quad n \ge m_0 \quad \text{such that} \quad |s_n| > bn^{1/2}. \tag{A.1}$$

Let $m > m_0$ and $0 < c \le b$. The over-all significance level of the modified repeated significance test studied in Section 2 is

$$\alpha = P_0\{T \le m\} + P_0\{T > m, |s_m| > cm^{1/2}\}. \tag{A.2}$$

An upper bound for the second term which is fairly accurate when c is small compared to b is

$$P_0\{cm^{1/2} < |s_m| \le bm^{1/2}\}. \tag{A.3}$$

The first term may be approximated using results of Siegmund (1977).

Let T_+ be defined by (A.1) with s_n in place of $|s_n|$. According to Siegmund (1978), if $b \to \infty$, $m \to \infty$, and $b = m^{1/2}\theta_1$, for each fixed $\theta > 0$, $x > 0$

$$P_\theta\{T_+ < m, s_m \le bm^{1/2} - x\} \cong \nu(\theta_1)\phi[m^{1/2}(\theta_1 - \theta)]e^{-\theta x}/m^{1/2}\theta. \tag{A.4}$$

For Brownian motion the corresponding approximation has 1 in place of $\nu(\theta_1)$. It is easy to see that

$$\begin{aligned} &P_\theta\{T_+ \le m\} + P_\theta\{T_+ > m, s_m > cm^{1/2}\} \\ &\qquad = P_\theta\{s_m > cm^{1/2}\} + P_\theta\{T_+ < m, s_m < cm^{1/2}\}. \end{aligned} \tag{A.5}$$

For $c = b$, (A.4) and (A.5) yield approximations to $P_\theta\{T \le m\}$ for $\theta \ne 0$. For c somewhat smaller than b one can often ignore the second term on the right hand side of (A.5) in approximating the power of a modified test.

APPENDIX B

Approximate expected sample size for normal data

Let $\{X(t), 0 \le t < \infty\}$ be a Brownian motion process with drift θ and variance 1 per unit time. Let $T = \inf\{t: t \ge m_0, |X(t)| = bt^{1/2}\}$. It may be shown (e.g., Siegmund, 1977) that for each $\theta \neq 0$, as $b \to \infty$

$$E_\theta T = (b^2 - 1)/\theta^2 + o(1). \tag{B.1}$$

For discrete normal random walk the corresponding expansion contains a term to account for excess over the stopping boundary, which can be computed numerically and for small θ is about

$$1.166/\theta \tag{B.2}$$

(cf. Lai and Siegmund, 1979).

For sequential tests of the kind discussed in this paper the expected sample size is

$$E_\theta \min(T, m) = E_\theta T - \int_{\{T>m\}} E_\theta(T - m \mid X(m))\, dP_\theta. \tag{B.3}$$

Suppose that $b \to \infty$ and $m \to \infty$ in such a way that $b = \theta_1 m^{1/2}$. For θ in a neighborhood of θ_1, say $\theta = \theta_1 + \xi m^{-1/2}$, it is possible to estimate the second term on the right-hand side of (B.3) to provide reasonable approximations to $E_\theta \min(T, m)$.

THEOREM. *Suppose $b \to \infty$ and $m \to \infty$ so that for some $\theta_1 \neq 0$, $b = \theta_1 m^{1/2}$. For $\theta = \theta_1 + \xi m^{-1/2}$*

$$\begin{aligned} E_\theta \min(T, m) = (b^2 - 1)/\theta^2 - \{m^{1/2}[\theta - \tfrac{1}{2}\theta_1]^{-1} \\ \cdot[\phi(\xi) - \xi\Phi(-\xi)]\} + \theta_1^{-2}[\Phi(-\xi)(1 + \xi^2) - \xi\phi(\xi)] + o(1). \end{aligned} \tag{B.4}$$

A sketch of a proof goes as follows. By (B.1) it suffices to consider the second term on the right-hand side of (B.3), which may be rewritten as

$$\int_0^\infty P_\theta\{X(m) \in \theta_1 m - dx\}(1 - P_0\{T < m \mid X(m) = \theta_1 m - x\})E_\theta \tau_m(x), \tag{B.5}$$

where

$$\tau_m(x) = \inf\{t: t > 0,\ X(t) = \theta_1 m^{1/2}[(m + t)^{1/2} - m^{1/2}] + x\}.$$

Since $\theta_1 m^{1/2}[(m + t)^{1/2} - m^{1/2}] \le \tfrac{1}{2}\theta_1 t$, a standard argument using Wald's identity yields

$$E_\theta \tau_m(x) \le x/(\theta - \tfrac{1}{2}\theta_1). \tag{B.6}$$

Writing the integral in (B.5) as the sum of integrals over $(0, m^{1/8})$, $(m^{1/8}, m^{1/2}\log m)$ and $(m^{1/2}\log m, \infty)$, one sees from (B.6) that the first and third integrals converge to 0 as $m \to \infty$. It may also be shown as in Siegmund (1977) that uniformly in $x \ge m^{1/8}$, $P_0\{T < m \mid X(m) = \theta_1 m - x\} = o(m^{-1})$, and hence by (B.5) and (B.6) it suffices to find an approximation for

$$m^{-1/2} \int_{m^{1/8}}^{m^{1/2}\log m} \phi(m^{1/2}(\theta_1 - \theta) - xm^{-1/2}) E_\theta \tau_m(x)\, dx. \tag{B.7}$$

A Taylor series expansion and some calculation with Wald's identity shows that uniformly for $x < m^{1/2}\log m$

$$E_\theta \tau_m(x) = x/(\theta - \tfrac{1}{2}\theta_1) - x^2/m\theta_1^2 + o(x^2/m),$$

which when substituted into (B.7) yields the theorem.

For the entries in Table 1, the quantity (B.2) was added to (B.4) to obtain a slightly better approximation to $E_\theta T$. It seems doubtful that estimating the excess over the boundary in the correction term $\int_{(T>m)} E_\theta(T - m \mid s_m)\, dP_{\theta_1}$ is worth the effort, since this term is already relatively small in those cases where the over-all approximation can be expected to be accurate. The entries in Table 2 were obtained from the Brownian motion approximation with mean and variance given by (5) and (6).

APPENDIX C

Approximate significance level for Bernoulli data

Let $\bar{x}_n$, $\bar{y}_n$, H, I, and l_n be as in Section 3. The stopping rule T defined by (4) may be rewritten

$$T = \text{first} \quad n \geq m_0 \quad \text{such that} \quad l_n > a, \tag{C.1}$$

where $a = b^2/2$. The significance level of the sequential test studied in Section 3 is given by (7). In this appendix an asymptotic expression similar to (7) is obtained for $P_{(p,p)}\{T \leq m\}$ as $m \to \infty$, $m_0 \to \infty$, $b \to \infty$ in such a way that $b = m^{1/2}\theta_1 = m_0^{1/2}\theta_0$. The method utilizes the nonlinear renewal theorem and an interesting adaptation of the methods of Lai and Siegmund (1977). Since the computations are rather elaborate, they are only given heuristically. The following likelihood ratio identity is also very helpful in simulating α.

Let $F_1 \subset F_2 \subset \cdots$ be an increasing sequence of sub-σ-algebras of a basic σ-algebra F. Let P and Q be two probabilities on F such that the restriction $P^{(n)}$ of P to F_n is absolutely continuous relative to the corresponding restriction $Q^{(n)}$ of Q. Let $L_n = dP^{(n)}/dQ^{(n)}$ be the likelihood ratio of these restrictions. One version of the fundamental identity of sequential analysis says that for any stopping time σ and any event A such that $A \cap \{\sigma = n\} \in F_n$ for all n,

$$P(A \cap \{\sigma < \infty\}) = \int_{A \cap \{\sigma < \infty\}} L_\sigma \, dQ. \tag{C.2}$$

(The proof follows at once by writing $\{\sigma < \infty\} = \cup_{n=1}^{\infty} \{\sigma = n\}$ and using the additivity of the integral.)

In what follows $P_{(p_1,p_2)}$ will be as in Section 3 and

$$Q = \int_0^1 \int_0^1 P_{(p_1,p_2)}\, dp_1\, dp_2.$$

Taking $P = P_{(p,p)}$ gives

$$L_n = dP^{(n)}_{(p,p)}/dQ^{(n)} = \binom{n}{s_n}\binom{n}{s_n^*} p^{s_n+s_n^*} q^{2n-s_n-s_n^*}(n+1)^2, \tag{C.3}$$

where $s_n = \sum_1^n x_k$ and $s_n^* = \sum_1^n y_k$. The identity (C.2) gives representations for $P_{(p,p)}\{T \leq m\}$ and $P_{(p,p)}\{T > m,\ (2l_m)^{1/2} > c\}$ which are useful in estimating these probabilities by Monte Carlo methods. One samples (x_1, y_1), (x_2, y_2), $\cdots$ according to Q and estimates $P_{(p,p)}\{T \leq m\}$, for example, by averages of $I_{\{T\leq m\}}L_T$. (See Siegmund, 1975, for a general discussion of such importance sampling in sequential analysis and Lai and Siegmund, 1977, for an application in a context similar to the present one.) This estimator has three advantages over direct simulation: (i) its variance is smaller; (ii) the expectation under Q of $\min(T, m)$ is smaller than under $P_{(p,p)}$ where it essentially equals the maximum sample size m; and (iii) $P_{(p,p)}\{T \leq m\}$ may be estimated simultaneously for several values of p using the same random numbers. For estimating $\alpha(p)$ for a test with c small compared to b this technique is not variance reducing, but advantages (ii) and (iii) hold in this case as well.

In contrast to the case of normal variables, where a direct representation of the probability that $T \leq m$ by means of (C.2) provided the starting point of a fruitful asymptotic analysis, in this case an indirect approach seems advisable. Let $u > 0$, $v > 0$, and let

$$P = \int_0^1 P_{(\xi,\xi)}\xi^u(1-\xi)^v\,d\xi/B(u+1, v+1), \tag{C.4}$$

so

$$\begin{aligned} L_n &= dP^{(n)}/dQ^{(n)} \\ &= (n+1)^2\binom{n}{s_n}\binom{n}{s_n^*}\Big/B(u+1, v+1)(u+v+2n+1)\binom{u+v+2n}{u+s_n+s_n^*}. \end{aligned} \tag{C.5}$$

Of course, P defined by (C.4) depends on u and v. If $u \to \infty$ and $v \to \infty$ in such a way that $u/(u+v) \to p$, then the distribution with density $\xi^u(1-\xi)^v/B(u+1, v+1)$ converges weakly to a point mass at p, so for each fixed m

$$P\{T \le m\} \to P_{(p,p)}\{T \le m\}. \tag{C.6}$$

Hence for large u and v with $u/(u+v) = p$ an approximation for $P\{T \le m\}$ "should be" an approximation for $P_{(p,p)}\{T \le m\}$.

Fix $0 < p_1, p_2 < 1$. Stirling's formula and the strong law of large numbers applied to (C.3) show that with $P_{(p_1,p_2)}$ probability one, as $n \to \infty$

$$\begin{aligned} \log L_n = -l_n + \tfrac{1}{2}\log n + \tfrac{1}{2}\log \bar{p}\bar{q}/p_1q_1p_2q_2 + u\log\bar{p} + v\log\bar{q} \\ - \log 2\pi^{1/2} - \log B(u+1, v+1) + o(1), \end{aligned} \tag{C.7}$$

where $\bar{p} = \tfrac{1}{2}(p_1 + p_2)$ and $\bar{q} = 1 - \bar{p}$. Suppose now that $a = b^2/2 \to \infty$, $m_0 \to \infty$, and $m \to \infty$ in such a way that $b = \theta_1 m^{1/2} = \theta_0 m_0^{1/2}$. Substitution of (C.7) into (C.2) and an argument similar to that of Lai and Siegmund (1977) yields

$$\begin{aligned} P\{T \le m\} &\sim \frac{1}{2}[B(u+1, v+1)]^{-1}(\pi^{-1}a)^{1/2}e^{-a} \\ &\quad \times \int_0^1\int_0^1\int_{\{T\le m\}} e^{-(l_T-a)}(T/a)^{1/2}(\bar{p}\bar{q}/p_1q_1p_2q_2)^{1/2}\,dP_{(p_1,p_2)}\bar{p}^u\bar{q}^v\,dp_1\,dp_2 \\ &\sim \frac{1}{2}[B(u+1, v+1)]^{-1}(\pi^{-1}a)^{1/2}e^{-a} \\ &\quad \times \iint_{\{(p_1,p_2):\theta_1<[2I(p_1,p_2)]^{1/2}<\theta_0\}} \tilde{u}(p_1, p_2)[I(p_1, p_2)]^{1/2}\,(\bar{p}\bar{q}/p_1q_1p_2q_2)^{1/2}\bar{p}^u\bar{q}^v\,dp_1\,dp_2, \end{aligned} \tag{C.8}$$

where

$$\tilde{v}(p_1, p_2) = \lim_{a\to\infty} E_{(p_1,p_2)}\exp[-(l_T - a)]$$

exists by an application of Theorem 1 of Lai and Siegmund (1977). (Actually, in order that this theorem be applicable it is necessary that a certain random walk associated with the process l_n be nonarithmetic, which is the case for all p_1, p_2 with at most a denumerable number of exceptions. This suffices in view of the subsequent integration over p_1 and p_2.)

Now suppose that $u \to \infty$ and $v \to \infty$ with $u/(u+v) = p \le \tfrac{1}{2}$. Some calculation shows that the measure $K_{u,v}(dp_1, dp_2) = [B(u+1, v+1)]^{-1}\bar{p}^u\bar{q}^v\,dp_1\,dp_2$ has total mass converging to $4p$ and converges weakly to the uniform distribution concentrated on $\tfrac{1}{2}(p_1 + p_2) = p$. Hence, $a^{1/2}e^a$ times the right-hand side of (C.8) converges to

$$\begin{aligned} \pi^{-1/2}\int_{(0,2p)\cap\{\xi:\theta_1<[2I(\xi,2p-\xi)]^{1/2}<\theta_0\}} \tilde{v}(\xi, 2p-\xi)[I(\xi, 2p-\xi)]^{-1/2} \\ \cdot[pq/\xi(1-\xi)(2p-\xi)(2q-1+\xi)]^{1/2}\,d\xi. \end{aligned} \tag{C.9}$$

Together with (C.6) this suggests that

$$P_{(p,p)}\{T \le m\} \sim C(p; \theta_0, \theta_1)a^{1/2}e^{-a} \qquad 0 < p \le \tfrac{1}{2}, \tag{C.10}$$

where $C(p; \theta_0, \theta_1)$ denotes the expression in (C.9). For $\frac{1}{2} \le p < 1$, a similar result holds with $C(q; \theta_0, \theta_1)$ in place of $C(p; \theta_0, \theta_1)$.

It should be emphasized that the preceding argument is only heuristic, although it seems to be possible to make it rigorous by taking u and v as functions of m which tend to ∞ slowly with m. The final result appears to agree formally with a similar very general result of Woodroofe (1978a), which is not directly applicable in this case because Woodroofe's condition L is not satisfied.

REFERENCES

ARMITAGE, P. (1975). *Sequential Medical Trials*, 2nd ed. Blackwell, Oxford.

CONN, H. O., RAMSBY, G. R., STORER, E. H. MUTCHNICK, M. G., JOSHI, P. H., PHILLIPS, M. M., COHEN, G. A., FIELDS, G. N., and PETROSKI, D. (1975). Intraarterial vasopressin in the treatment of upper gastrointestinal hemorrhage: a prospective, controlled clinical trial. *Gastroenterology* **68** 211–221.

EFRON, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika* **58** 403–417.

LAI, T. L. and SIEGMUND, D. (1977). A nonlinear renewal theory with applications to sequential analysis I. *Ann. Statist.* **5** 946–954.

LAI, T. L. and SIEGMUND, D. (1979). A nonlinear renewal theory with applications to sequential analysis II. *Ann. Statist.* **7** 60–76.

MCPHERSON, C. K. and ARMITAGE, P. (1971). Repeated significance tests on accumulating data when the null hypothesis is not true. *J. Roy. Statist. Soc. Ser. A* **134** 15–26.

PETO, R, PIKE, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J., and SMITH, P. G. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. *Br. J. Cancer* **34** 585–612.

SIEGMUND, D. (1975). Importance sampling in the Monte Carlo study of sequential tests. *Ann. Statist.* **4** 673–684.

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

SIEGMUND, D. (1978). Estimation following sequential test. *Biometrika* **65** 341–349.

SIEGMUND, D. and GREGORY, P. (1979). A sequential clinical trial for testing $p_1 = p_2$. Stanford Univ. Technical Report.

WALD, A. (1947). *Sequential Analysis*, Wiley, New York.

WOODROOFE, M. (1978a). Large deviations of likelihood ratio statistics with applications to sequential testing. *Ann. Statist.* **6** 72–84.

WOODROOFE, M. (1978b). Repeated likelihood ratio tests. Univ. of Michigan Technical Report.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305