AD-A065 733 STANFORD UNIV CALIF DEPT OF STATISTICS F/G 6/15
A SEQUENTIAL CLINICAL TRIAL FOR TESTING P1 = P2.(U)
FEB 79 D SIEGMUND, P GREGORY N00014-77-C-0306
UNCLASSIFIED TR5 NL

| OF |
AD
AD 65733

END
DATE FILMED
5-79
DDC

A SEQUENTIAL CLINICAL TRIAL FOR TESTING $p_1 = p_2$

BY

D. SIEGMUND and P. GREGORY

TECHNICAL REPORT NO. 5
FEBRUARY 2, 1979

PREPARED UNDER CONTRACT
N00014-77-C-0306 (NR-042-373)
FOR THE OFFICE OF NAVAL RESEARCH

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

A SEQUENTIAL CLINICAL TRIAL FOR TESTING $p_1 = p_2$

by

D. Siegmund and P. Gregory

Technical Report No. 5

February 2, 1979

Prepared under Contract

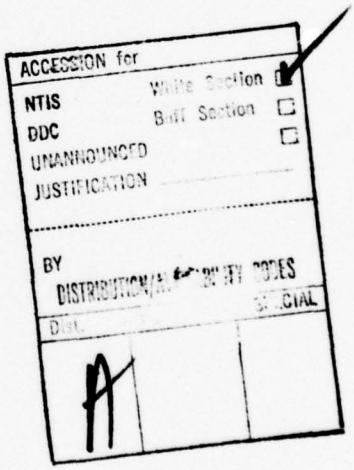
N00014-77-C-0306 (NR-042-373)

for the Office of Naval Research

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

ABSTRACT

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.



AMS 1970 Subject Classification. Primary 62L10.

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

A SEQUENTIAL CLINICAL TRIAL FOR TESTING $p_1 = p_2$

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 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 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. Success 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 sequential 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 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. The extent to which a sequential design introduces a bias which makes these other analyses difficult is perhaps its most serious disadvantage. (This point will be discussed again in Section 2.)

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 $n = 1,2,\ldots$ 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 $s_n = x_1 + \ldots + x_n$, and given $b_1 > 0$ and $m_0 = 1,2,\ldots$ define

$$(1) \qquad T_1 = \text{first } n \geq m_0 \text{ such that } |s_n| > b_1 \sigma n^{\frac{1}{2}} .$$

Let $m_1 \geq m_0$ 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.

Consider a repeated significance test and a fixed sample size test of the same significance level and power at some given $\theta_1 \neq 0$. The advantage of the sequential test is that if $|\theta|$ is large, indicating that one treatment is considerably superior, the expected sample size of the sequential test is much smaller than the fixed sample size. Concomitant disadvantages are that the sequential test has a considerably larger maximum sample size and less power for detecting values θ closer to zero than θ_1.

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 $|s_m| > cm^{\frac{1}{2}}$. 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 $|s_m| > cm^{\frac{1}{2}}$, whereas $c = b = b_1$ and $m = m_1$ give a repeated significance test. Figure 1 illustrates these relations.

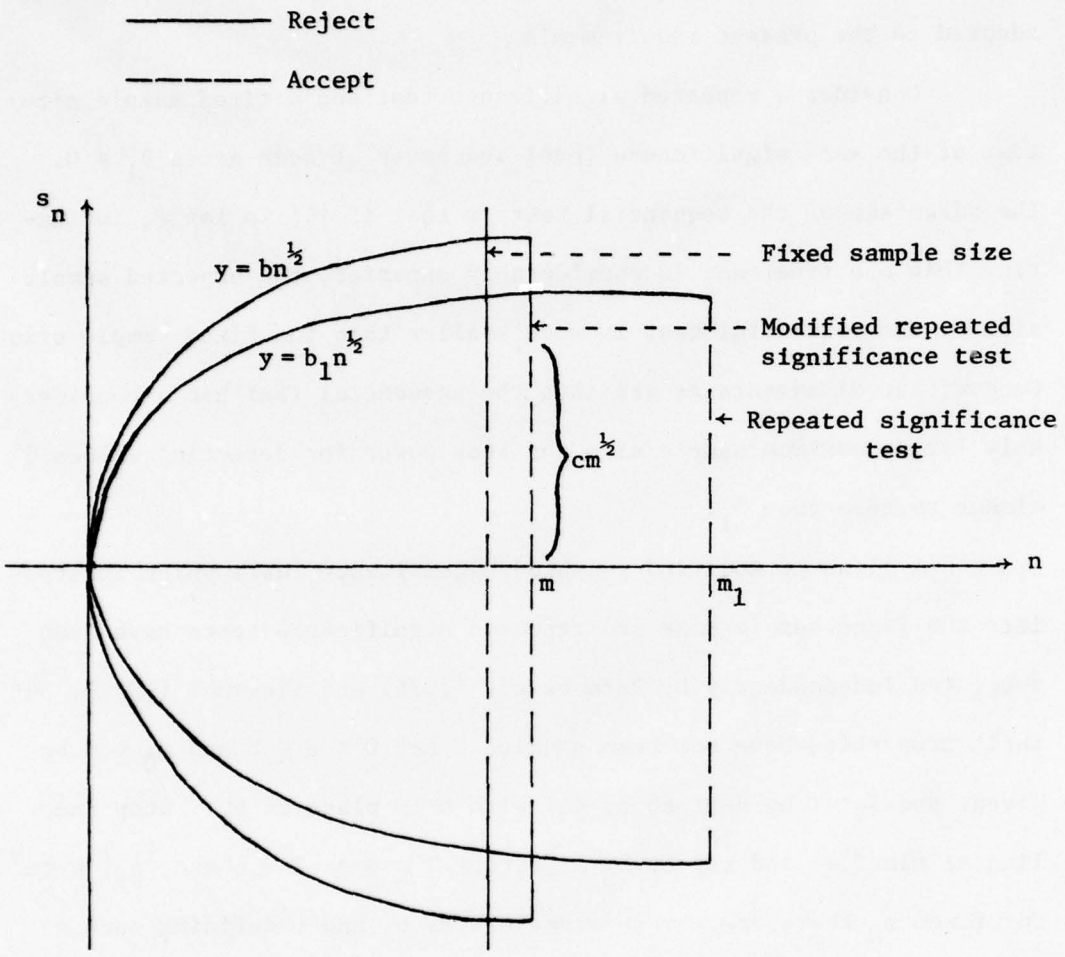

Reject
Accept
s_n
$y = bn^{\frac{1}{2}}$
$y = b_1 n^{\frac{1}{2}}$
Fixed sample size
Modified repeated significance test
Repeated significance test
$cm^{\frac{1}{2}}$
n
m
m_1

Figure 1.

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), p. 104. (Simulations indicate that these approximations are quite accurate.) For all tests the over-all significance level is .05 and $m_0 = 1$. The maximum sample sizes were taken to be equal and large enough to include the apparent range of possibilities for the proposed vasopressin study. The entry m_f denotes that fixed sample size which would yield the same power at the indicated θ as the given sequential test. One measure of the savings in sample size for the sequential tests compared to fixed sample size tests is $R = (m_f - \text{expected sample size})/m_f$. For the modified tests b was chosen fairly large, so that $P_0\{T \leq m\}$ is slightly less than .02. The entry $1 - \beta^*$ denotes the probability at θ of early termination, $P_\theta\{T \leq m\}$.

A reasonable over-all conclusion seems to be that for about a 10% increase in the maximum sample size of a fixed sample test, a modified sequential test can yield about the same power and a 40% reduction in the expected sample size for $|\theta|$ large enough that the power approaches one. Comparatively, a repeated significance test obtains a greater reduction in expected sample size for large $|\theta|$ at the cost of a considerable increase in the maximum sample size and some loss of power. The relative advantages and disadvantages of the repeated significance tests compared to the modified tests become more pronounced with larger patient horizons.

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	$1-\beta^*$
.6	.99	.95	21	37	.98	25	45	.90
.4	.80	.61	35	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	$1-\beta^*$
.4	.99	.95	47	82	.98	59	101	.88
.3	.88	.71	72	71	.85	87	100	.57

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 simple 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

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

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, $P_0\{T \leq 49\} \cong .018$ 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) In many cases the parameter θ represents one direction in a multidimensional space. Implicit in the use of a sequential test is the assumption that if $|\theta|$ appears to be large, that by itself will play a dominant role in the choice of treatment (although one would presumably look at other factors to see that they are reasonably consistent with this conclusion). However, if $|\theta|$ is small, a more detailed analysis of these other factors may be important. For example, if vasopressin reduces bleeding sufficiently that surgery need not be performed on an emergency basis, it would be a useful treatment, even though its "success" rate may be no higher than for placebo.

Use of a data dependent stopping rule introduces a bias which tends to make $|\theta|$ seem larger than it actually is. (See Siegmund, 1978, for a discussion of the magnitude of this bias.) By using a modified test with b large enough that $P_\theta\{T \leq m\}$ is relatively small for small $|\theta|$, one reduces the biasing effect of the stopping rule, particularly for small $|\theta|$ when analysis of other variables may play an important role.

3. Comparing Two Binomials

The clinical trial of vasopressin may be regarded as one of comparing two binomial success probabilities p_1 and p_2, where p_1 is the probability of success with vasopressin and p_2 the probability of success with placebo. There are possibilities for stratifying the population, e.g., according to the cause of bleeding, but initially it is easier to suppose that all data are pooled.

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. For the first two tests discussed below patients may easily be assigned to treatment or control in a 2 to 1 or other ratio. The matched pairs test described later relies heavily on approximately balanced allocation to treatment and control.

For theoretical discussions of clinical trials it is customary to consider testing $p_1 = p_2$ against the two-sided alternative $p_1 \neq p_2$, and that custom is followed here. For a comparison of treatment and placebo it is sometimes appropriate to consider only the one-sided alternative $p_1 > p_2$. The appropriate modifications are obvious, and for the numerical examples presented later give tests having a significance level about one-half that of the two-sided tests.

Assume then that the data consists of pairs (x_1,y_1), $(x_2,y_2),\ldots$, 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 which may be important in some applications.

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 $\ell_n = n\, I(\bar{x}_n, \bar{y}_n)$, where $\bar{x}_n = n^{-1} \Sigma_1^n x_k$ and $\bar{y}_n = n^{-1} \Sigma_1^n y_k$. In analogy with (1), given integers $m_0 \leq m$ and real numbers $0 < c \leq b$, define

$$T = \text{first } n \geq m_0 \text{ such that } (2\ell_n)^{\frac{1}{2}} > b \quad . \tag{4}$$

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

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

$$\mu = [2I(p_1,p_2)]^{\frac{1}{2}} \tag{5}$$

and variance

$$(6) \quad \sigma^2 = \{p_1q_1 \log^2(p_1\bar{q}/q_1\bar{p}) + p_2q_2 \log^2(p_2\bar{q}/q_1\bar{p})\}/2\, I(p_1,p_2) \quad ,$$

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 $(2n\ell_n)^{\frac{1}{2}}$ 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. Table 2 below indicates that these approximations are reasonably accurate.

Approximating the significance level of this test poses a more difficult problem. Let $P_{(p_1,p_2)}$ denote the probability measure on sequences (x_1,y_1), $(x_2,y_2),\ldots$, determined by p_1 and p_2. The significance level of the modified repeated significance test defined above is for $p_1 = p_2 = p$

$$(7) \qquad \alpha(p) = P_{(p,p)}\{T \leq m\} + P_{(p,p)}\{T > m,\ (2\ell_m)^{\frac{1}{2}} > c\} \quad .$$

The difficulties arise with the first term on the right hand side of (7). The second term may be approximated by the upper bound $P_{(p,p)}\{c < (2\ell_m)^{\frac{1}{2}} \leq b\}$, which may in turn be approximated by $2[\Phi(b) - \Phi(c)]$, since $2\ell_m$ is asymptotically χ^2 with one degree of freedome 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 $bm^{-\frac{1}{2}} \to \theta_1$ and $bm_0^{-\frac{1}{2}} \to \theta_0$. 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 $p \leq \frac{1}{2}$,

$$(8)\quad P_{(p,p)}\{T \leq m\} \lesssim (\pi^{-1}pqa)^{\frac{1}{2}}\, e^{-a} \int_{(0,2p)\cap\{\xi:\theta_0<[2I(\xi,2p-\xi)]^{\frac{1}{2}}<\theta_1\}} [I(\xi,2p-\xi)\xi(1-\xi)(2p-\xi)(2q-1+\xi)]^{-\frac{1}{2}}\, d\xi.$$

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. 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 $P_{(p,p)}\{T \leq m\}$ 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 $P_{(p,p)}\{T \leq m\}$ and $\alpha(p)$, for which they are too large as expected. The author has 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 $P_{(p,p)}\{T \leq m\}$. Taking m_0 about equal to $m^{\frac{1}{2}}$ seems reasonable and has the additional desirable property

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 (82.7)
.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 (85.7)
.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.

of making $P_{(p,p)}\{T \leq m\}$ fairly constant as a function of p, at least for p not too near 0 or 1. For this choice of m_0, $P_{(p,p)}\{T \leq m\}$ also seems to be quite well approximated by the analogous probability for normal random variables with $m_0 = 1$. Although there is no apparent reason why this should be so, it seems to provide reasonable quidelines for tentatively selecting b and m subject to verification by a Monte Carlo experiment.

The customary χ^2 statistic for testing $p_1 = p_2$ is $\chi_n^2 = n(\bar{x}_n - \bar{y}_n)^2/2\,p_n\,q_n$, 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 (4) with χ_n in place of $(2\ell_n)^{\frac{1}{2}}$. Again sampling terminates at min(T,m) and H_0 is rejected if either $T \leq m$ or $T > m$ and $\chi_m > c$.

For given b, c, m_0, and m, simulations show that $P_{(p_1,p_2)}\{T \leq m\}$ is generally smaller for this test and the expected sample size larger than for the likelihood ratio tests. If b and c are decreased to make $P_{(p,p)}\{T \leq m\}$ and the significance levels comparable, there seems to be essentially no difference in the power and expected sample size of the two tests, at least over the range studied in Table 2. Brownian motion approximations similar to those of Table 2 seem adequate when $p_1 \neq p_2$, although in this case the approximating random walk has the slightly different expectation

$$\mu_1 = (p_1 - p_2)/(2\,\bar{p}\,\bar{q})^{\frac{1}{2}}$$

and variance

$$\sigma_1^2 = p_1q_1f_+(p_1,p_2) + p_2q_2f_-(p_1,p_2) \quad ,$$

where

$$f_\pm(p_1,p_2) = [\pm\bar{p}\bar{q} - (p_1 - p_2)(q_1 - p_2)/4]^2/2(\bar{p}\bar{q})^3 \quad .$$

One slight advantage that the likelihood ratio test seems to have over the χ^2 test is that its value of $P_{(p,p)}\{T \leq m\}$ may be more nearly constant as p gets close to 0 or 1. Previous experience suggests that extreme values of p are not likely in the case of upper gastrointestinal hemorrhage, and hence they were not systematically investigated. Of course, the χ^2 statistic is easier to compute. Otherwise, the authors' slight preference for the likelihood ratio test is based on the existence of the primitive but interesting null hypothesis theory which may be developed for it and which makes selection of the design parameters m_0, m, b, and c slightly easier.

Yet a third test of $p_1 = p_2$ (one which differs from the other two in several important respects) is obtained by discarding those pairs for which $x_i = y_i$ and basing a test on the number of times $x_i = 1$ among those pairs for which $x_i \neq y_i$. This device was suggested by Wald (1947, p. 107) and has been advocated by Armitage (1975) for general use in making sequential binomial comparisons. It reduces the problem of testing the equality of two binomial parameters p_1 and p_2 to that of testing whether a single binomial parameter λ equals $\frac{1}{2}$, where $\lambda = p_1q_2/(p_1q_2 + p_2q_1) = P_{(p_1,p_2)}\{x_i = 1 | x_i \neq y_i\}$.

Let $n^{\#}$ be the number of pairs $i \leq n$ with $x_i \neq y_i$ and let $\bar{z}_{n^{\#}}$ be the proportion of pairs with $x_i = 1$ among those with $i \leq n$ and $x_i \neq y_i$. The log generalized likelihood ratio for testing $H_0 : \lambda = \frac{1}{2}$ against $H_1 : \lambda \neq \frac{1}{2}$ is

$$\ell^{\#}_{n^{\#}} = n^{\#}[H(\bar{z}_{n^{\#}}) - H(\tfrac{1}{2})] \quad .$$

In analogy with (4), given b, c, m_0, and m let

(9) $$T^{\#} = \text{first } n^{\#} \geq m_0 \text{ such that } (2\ell^{\#}_{n^{\#}})^{\frac{1}{2}} > b \quad .$$

Stop sampling at $T^{\#}$ if $T^{\#} \leq m$ and otherwise when $n^{\#} = m$. Reject H_0 if $T^{\#} \leq m$ or $T^{\#} > m$ and $(2\ell^{\#}_{n^{\#}})^{\frac{1}{2}} > c$.

Although a direct numerical approach is easy, for present purposes it is simpler to adapt the methods introduced already to compute approximately various properties of these tests. For Brownian motion approximations to the distribution of $T^{\#}$ under H_1, the appropriate mean and variance are

(10) $$\mu_2 = \{2[H(\lambda) - H(\tfrac{1}{2})]\}^{\frac{1}{2}}$$

and

(11) $$\sigma_2^2 = \lambda(1-\lambda)\log^2[\lambda/(1-\lambda)]/2[H(\lambda) - H(\tfrac{1}{2})] \quad ,$$

where $\lambda = p_1q_2/(p_1q_2 + p_2q_1)$. A "no overshoot" approximation analogous to (8) is

(12) $$P_{(p,p)}\{T^{\#} \leq m\} \lesssim (2a/\pi)^{\frac{1}{2}} e^{-a} \int_{(0,\frac{1}{2}) \cap \{\xi : \theta_0 < \{2[H(\xi) - H(\frac{1}{2})]\}^{\frac{1}{2}} < \theta_1\}} \{\xi(1-\xi)[H(\xi) - H(\tfrac{1}{2})]\}^{-\frac{1}{2}} \, d\xi \quad .$$

For $b = 3.15$, $c = 2.15$, $m = 49$, and $m_0 = 8$ (it is impossible to stop before $n^{\#} = 8$ for this value of b), (12) gives $P_{(p,p)}\{T^{\#} \leq m\} \underset{\sim}{<} .020$. Monte Carlo estimates based on 5000 repetitions for this probability and the over-all significance level are respectively $.018 \pm .000$ and $.056 \pm .002$. For $p_1 = p_2$ about equal to .5, it takes on average about two pairs (x_n, y_n) to obtain one pair with $x_n \neq y_n$. Thus, a maximum value $m = 49$ corresponds roughly to a maximum value of $n = 98$. Hence, for $p_1 = p_2 \tilde{=} .5$ this test is comparable to the second test in Table 2, although its over-all significance level is somewhat larger. Table 3 gives 900 repetition Monte Carlo and analytic approximations for some additional comparisons with the second test in Table 2. The analytic approximations to $P_{(p_1,p_2)}\{\text{Reject } H_0\}$ are obtained by a normal approximation with continuity correction; the other analytic results use Brownian motion with mean and variance given by (10) and (11). These figures suggest that in terms of operating characteristic and expected sample size there is no strong reason for preferring one test to the other. If the values of c were altered slightly to make the over-all significance levels of the matched pairs test and the two population test more nearly equal, the power functions would tend to converge, and the two tests would appear to be even more similar.

There are several important differences between the matched pairs test and the likelihood ratio test defined by (4). As was noted earlier, the matched pairs test cannot be used unless approximately equal numbers of patients are allocated to treatment and control, whereas the likelihood ratio test has no such restriction. Even in the

TABLE 3

MATCHED PAIRS TEST: $m_0 = 8$, $m = 49$, $b = 3.15$, $c = 2.15$

(p_1, p_2)	λ	$P_{(p_1,p_2)}\{T^{\#} \leq m\}$	$P_{(p_1,p_2)}\{\text{Reject } H_0\}$	$E_{(p_1,p_2)}(T^{\#} \wedge m)/(p_1q_2 + p_2q_1)$
.7, .5	.7	.506 (.501)	.822 (.809)	76.8 (73.6)
.8, .5	.8	.933 (.942)	.996 (.997)	48.6 (40.4)
.6, .4	.69	.462 (.442)	.767 (.762)	75.2 (70.8)
.7, .4	.78	.860 (.885)	.979 (.990)	51.5 (48.9)
.8, .4	.86	1.00 (.995)	1.00 (1.00)	30.2 (28.6)

case of equal allocations to treatment and control the matched pairs test has the disadvantage that its maximum sample size is a random variable. If N denotes the number of pairs (x_n, y_n) which must be observed until $n^{\#} = m$, then N is the waiting time until m "successes" in Bernoulli trials with success probability $\delta = p_1q_2 + p_2q_1$. Hence, $E_{(p_1,p_2)}N = m/\delta$ and $\text{var}_{(p_1,p_2)}(N) = m(1-\delta)/\delta^2$. If in the preceding example $p_1 = p_2 = .5$, then $E_{(p_1,p_2)}(N) = \text{var}_{(p_1,p_2)}(N) = 98$. However, if $p_1 = p_2 = .3$, then $E_{(p_1,p_2)}(N) = 116.7$ and $\text{var}_{(p_1,p_2)}(N) = (12.7)^2$, so one may with reasonably large probability attain 49 pairs having $x_n \neq y_n$ in fewer than $116.7 - 12.7 = 104$ or more than $116.7 + 12.7 \cong 130$ total pairs. The situation is worse for larger departures from $\delta = 1/2$. Without some compensating feature it seems doubtful that one would prefer this test to that determined by (4), except perhaps in those cases when p_1 and p_2 are expected to be near 0 or 1.

There are, however, compensating features, the most important of which is the use of the matched pairs test to deal with stratified populations. By pairing observations within strata it is often possible in effect to increase the value of λ and hence the sensitivity of the test. (See Armitage, 1975, especially p. 86.)

Upper gastrointestinal hemorrhage has several common causes. Very tentative figures from previous studies suggest that both the spontaneous remission rate and the success rate with vasopressin vary with cause. With some grouping of causes according to these apparent remission rates, perhaps 90-95% of all patients can be put into three strata: (a) esophagitis, gastritis, Mallory-Weiss tear; (b) various

ulcers; and (c) esophageal varices. Highly speculative success rates with and without vasopressin may be approximately .7 and .5 for (a), .5 and .3 for (b), and .7 and .3 for (c). It is also difficult to guess the size of the various strata. (For example, the percentage of variceal patients is larger in populations with a large proportion of alcoholics--hence, typically larger at V. A. hospitals than at others.) But the ratio 25:25:50 may be a reasonable approximation. If these figures are correct, then in the entire population $p_1 = .65$ and $p_2 = .35$, so random pairing gives $\lambda = .775$. Matching pairs within strata gives $\lambda = .778$. Modest variations in these assumed figures consistently yield differences in λ of less than .01. In general, stratification increases slightly the average number of pairs required to obtain a pair with $x_n \neq y_n$, which cancels to some extent the small advantage gained by increasing λ.

The failure of stratification to produce a more dramatic effect is presumably due to the fact that with the maximum sample sizes envisaged here, a treatment effect with an appreciable probability of being detected must be as large or larger than differences among strata. For larger experiments designed to detect smaller treatment effects, the value of stratification should be re-examined.

The preceding analysis ignores the possibility of a time trend, against which pairing of patients according to their arrival time provides protection. The history of the treatment of upper gastrointestinal bleeding and the simple mode of administration of vasopressin in this trial suggest that time trends do not pose serious problems.

In summary, the modest increase in sensitivity of the matched pairs test in the range of parameters and sample sizes considered for the vasopressin study does not seem sufficient compensation for the random variation in its maximum sample size and its failure to deal adequately with unbalanced allocations to treatment and control.

APPENDIX A

PROBABILITY APPROXIMATIONS FOR NORMAL DATA

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

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

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

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

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

$$P_0\{cm^{\frac{1}{2}} < |s_m| \leq bm^{\frac{1}{2}}\} \quad . \tag{A.3}$$

The first term may be approximated using results of Siegmund (1977). If $b \to \infty$ and $m \to \infty$ in such a way that $b = m^{\frac{1}{2}}\theta_1$, then

$$P_0\{T \leq m\} \sim (2/\pi)^{\frac{1}{2}} b\, e^{-b^2/2} \int_{\theta_1}^{\infty} [\nu(x)/x]dx \quad , \tag{A.5}$$

where

$$\nu(x) = 2 \exp[-2 \Sigma_{n=1}^{\infty} n^{-1} \Phi(-\tfrac{1}{2} x n^{\frac{1}{2}})]/x^2 \quad . \tag{A.6}$$

The approximation (A.5) is very accurate when $m_0 = 1$. For larger values of m_0 one seems to obtain a better approximation by assuming in addition that $m_0 \to \infty$ and $b = m_0^{\frac{1}{2}}\theta_0$. Then

$$(A.7)\quad P_0\{T \le m\} = P_0\{|s_m| > bm^{\frac{1}{2}}\} + P_0\{m_0 < T \le m\}$$

$$\sim 2[1 - \Phi(b)] + (2/\pi)^{\frac{1}{2}} b\, e^{-b^2/2} \int_{\theta_1}^{\theta_0} [\nu(x)/x]dx \quad .$$

Although the first term on the right hand side of (A.7) is of smaller order of magnitude than the second, and hence there is no good mathematical reason for including it, its inclusion seems to give more accurate numerical results when compared with simulations (cf. equation (50) of Siegmund, 1977).

It is easy to compute the function $\nu(x)$ defined in (A.6) numerically. Woodroofe (1978b) has tabled an integral equivalent to that in (A.5). Siegmund (1979) has shown that for x less than about 2, $\nu(x)$ is well approximated by $\exp(-.583x)$, and hence for many values of θ_0 and θ_1 the integrals in (A.5) and (A.7) may be obtained from tables of the exponential integral and the value

$$\int_2^\infty [\nu(x)/x]dx \cong .224 \quad .$$

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 $b = m^{\frac{1}{2}}\theta_1$, for each fixed $\theta > 0$, $x > 0$

$$(A.8)\quad P_\theta\{T_+ < m,\ s_m < bm^{\frac{1}{2}} - x\} \cong \nu(\theta_1)\phi[m^{\frac{1}{2}}(\theta_1 - \theta)]e^{-\theta x}/m^{\frac{1}{2}}\theta \quad .$$

For Brownian motion the corresponding approximation has 1 in place of $\nu(\theta_1)$. Together with the obvious decomposition

$$P_\theta\{T_+ \le m\} = P_\theta\{s_m > b\, m^{\frac{1}{2}}\} + P_\theta\{T_+ < m,\ s_m < b\, m^{\frac{1}{2}}\} \quad ,$$

(A.8) allows one to approximate $P_\theta\{T \le m\}$ for $\theta \neq 0$.

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 $T = \inf\{t : t \ge m_0,\ |X(t)| = bt^{\frac{1}{2}}\}$. It may be shown (e.g., Siegmund, 1977) that for each $\theta \neq 0$, as $b \to \infty$

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

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

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

Suppose that $b \to \infty$ and $m \to \infty$ in such a way that $b = \theta_1 m^{\frac{1}{2}}$. For θ in a neighborhood of θ_1, say $\theta = \theta_1 + \xi m^{-\frac{1}{2}}$, 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 $\theta_1 \neq 0$, $b = \theta_1 m^{\frac{1}{2}}$. For $\theta = \theta_1 + \xi m^{-\frac{1}{2}}$

(B.4) $$E_\theta \min(T,m) = (b^2-1)/\theta^2 - \{m^{\frac{1}{2}}[\theta - \tfrac{1}{2}\theta_1]^{-1}[\phi(\xi) - \xi\Phi(-\xi)]\}$$

$$+ \theta_1^{-2}[\Phi(-\xi)(1+\xi^2) - \xi\phi(\xi)] + o(1) \quad .$$

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

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

where

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

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

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

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

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

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

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

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 $\int_{(T>m)} E_\theta(T - m|s_m)dP_{\theta_1}$ 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

(C.1) $$T = \text{first } n \geq m_0 \text{ such that } \ell_n > a \ ,$$

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 (A.5) 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 $b = m^{\frac{1}{2}}\theta_1 = m_0^{\frac{1}{2}}\theta_0$. The method utilizes the non-linear 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 $F_1 \subset F_2 \subset \ldots$ 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,

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

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

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

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

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

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

where $s_n = \Sigma_1^n x_k$ and $s_n^* = \Sigma_1^n y_k$. The identity (C.2) gives representations for $P_{(p,p)}\{T \leq m\}$ and $P_{(p,p)}\{T > m, \ (2\ell_m)^{\frac{1}{2}} > c\}$ which are useful in estimating these probabilities by Monte Carlo methods. One samples (x_1,y_1), (x_2,y_2),... 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

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

so

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

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

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

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 $n \to \infty$

$$\log L_n = -\ell_n + \frac{1}{2}\log n + \frac{1}{2}\log \bar{p}\,\bar{q}/p_1 q_1 p_2 q_2 + u \log \bar{p} + v \log \bar{q}$$

$$- \log 2\pi^{\frac{1}{2}} - \log B(u+1, v+1) + o(1) \quad , \tag{C.7}$$

where $\bar{p} = \frac{1}{2}(p_1 + p_2)$ 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 $b = \theta_1 m^{\frac{1}{2}} = \theta_0 m_0^{\frac{1}{2}}$. Substitution of (C.7) into (C.2) and an argument similar to that of Lai and Siegmund (1977) yields

(C.8) $P\{T \le m\}$

$$\sim \tfrac{1}{2}[B(u+1,\ v+1)]^{-1}(\pi^{-1}a)^{\frac{1}{2}} e^{-a}$$

$$\times \int_0^1 \int_0^1 \int_{\{T \le m\}} e^{-(\ell_T - a)} (T/a)^{\frac{1}{2}} (\bar{p}\,\bar{q}/p_1 q_1 p_2 q_2)^{\frac{1}{2}} dP_{(p_1,p_2)} \bar{p}^u \bar{q}^v dp_1 dp_2$$

$$\sim \tfrac{1}{2}[B(u+1,\ v+1)]^{-1}(\pi^{-1}a)^{\frac{1}{2}} e^{-a}$$

$$\times \int\int_{\{(p_1,p_2):\theta_1 < [2I(p_1,p_2)]^{\frac{1}{2}} < \theta_0\}} \tilde{\nu}(p_1,p_2)\,[I(p_1,p_2)]^{-\frac{1}{2}} (\bar{p}\,\bar{q}/p_1 q_1 p_2 q_2)^{\frac{1}{2}} \bar{p}^u \bar{q}^v\, dp_1 dp_2,$$

where

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

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 non-arithmetic, 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 $u/(u+v) = p \le \frac{1}{2}$. 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 a measure uniform on $\frac{1}{2}(p_1 + p_2) = p$. Hence, $a^{\frac{1}{2}} e^a$ times the right hand side of (C.8) converges to

$$(C.9)\quad \pi^{-\frac{1}{2}} \int_{(0,2p)\cap\{\xi:\theta_1<[2I(\xi,2p-\xi)]^{\frac{1}{2}}<\theta_0\}} \tilde{\nu}(\xi,2p-\xi)\,[I(\xi,2p-\xi)]^{-\frac{1}{2}}\,[pq/\xi(1-\xi)(2p-\xi)(2q-1+\xi)]^{\frac{1}{2}}\,d\xi\ .$$

Together with (C.6) this suggests that

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

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 however, does not apply in this case because certain smoothness conditions important to Woodroofe's method are not satisfied.

REFERENCES

Armitage, P. (1975). Sequential Medical Trials, 2nd ed., Oxford: Blackwell.

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 non-linear renewal theory with applications to sequential analysis I, Ann. Statist. 5, 946-954.

Lai, T.L. and Siegmund, D. (1979). A non-linear renewal theory with applications to sequential analysis II, Ann. Statist. 7,

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 tests, Biometrika 65, 341-349.

Siegmund, D. (1979). Corrected diffusion approximations in certain random walk problems, Adv. Appl. Prob. 11,

Wald, A. (1947). Sequential Analysis, John Wiley and Sons, 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.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: TR-#5, TR-128	2. GOVT ACCESSION NO. / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): SEQUENTIAL CLINICAL TRIAL FOR TESTING $p_1 = p_2$.	5. TYPE OF REPORT & PERIOD COVERED: Technical report,
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): D. SIEGMUND and P. GREGORY	8. CONTRACT OR GRANT NUMBER(s): N00014-77-C-0306, NSF-MCS77-16974
9. PERFORMING ORGANIZATION NAME AND ADDRESS: DEPARTMENT OF STATISTICS, STANFORD UNIVERSITY, STANFORD, CALIF.	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: (NR-042-373)
11. CONTROLLING OFFICE NAME AND ADDRESS: STATISTICS & PROBABILITY PROGRAM (Code 436), OFFICE OF NAVAL RESEARCH, ARLINGTON, CA. 22217	12. REPORT DATE: February 2, 1979
	13. NUMBER OF PAGES: 34
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): 38p.	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE AND SALE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Also issued as Technical Report No. 128 under National Science Foundation Grant MCS77-16974 - Dept. of Statistics, Stanford University

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

SEQUENTIAL TEST

CLINICAL TRIAL

STOPPING RULE

20. ABSTRACT

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.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-014-6601

UNCLASSIFIED.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

332580