Exact Statistical Inference for Group Sequential Trials

D. Y. Lin
Department of Biostatistics, SC-32, University of Washington,
Seattle, Washington 98195, U.S.A.

and

L. J. Wei and D. L. DeMets
Department of Statistics, University of Wisconsin,
Madison, Wisconsin 53706, U.S.A.

SUMMARY
This paper considers clinical trials comparing two treatments with dichotomous responses where the data are examined periodically for early evidence of treatment difference. The existing group sequential methods for such trials are based on the large-sample normal approximation to the joint distribution of the estimators of treatment difference over interim analyses. We demonstrate through extensive numerical studies that, for small and even moderate-sized trials, these approximate procedures may lead to tests with supranominal size (mainly when unpooleed estimators of variance are used) and confidence intervals with under-nominal coverage probability. We then study exact methods for group sequential testing, repeated interval estimation, and interval estimation following sequential testing. The new procedures can accommodate any treatment allocation rules. An example using real data is provided.

1. Introduction
In clinical trials with sequential patient entry, fixed-sample-size designs are sometimes unjustified on ethical grounds and fully sequential procedures are often impractical. One compromise is the group sequential design under which “multiple looks” are performed during the trial. It is widely recognized that the use of the standard single-stage, fixed-sample-size test (or confidence interval) at each look would lead to an overall Type I error (or noncoverage rate) significantly higher than the nominal level. A number of methods have been proposed to determine proper critical values for maintaining a prespecified overall significance level in sequential testing, to construct repeated confidence intervals with a given overall coverage probability, and to obtain a valid confidence interval following a sequential trial. An excellent review on the existing methodology was given by Jennison and Turnbull (1989).

In this paper, we are mainly concerned with the commonly used group sequential trials comparing two treatments (A and B) with relatively quick dichotomous responses (success and failure). The data are reviewed periodically with cumulative $m_k$ subjects receiving treatment A and $n_k$ subjects receiving treatment B by the $k$th look. Let random variables $X_k$ and $Y_k$ denote the numbers of successes out of the $m_k$ and $n_k$ subjects, respectively. We assume that the maximum number of looks, $K$, is fixed in advance, though some of the procedures developed here can be extended to the case where $K$ is not fixed in advance.

Key words: Clinical trial; Estimation after testing; Exact unconditional test; Interim analyses; Repeated confidence intervals; Repeated significance tests.
The existing group sequential methods for testing and estimating the treatment difference $\Delta = p_A - p_B$ in the above problem rely on the large-sample normal approximation to the joint distribution of $(\hat{\Delta}_1, \ldots, \hat{\Delta}_K)$, where $\hat{\Delta}_k = \hat{p}_{Ak} - \hat{p}_{Bk}$, $\hat{p}_{Ak} = X_k/m_k$, and $\hat{p}_{Bk} = Y_k/n_k$. When the pooled method of variance estimation is adopted, these approximate sequential tests have reasonable control over their size. However, when unpooled estimators of variance are used, as are necessary for testing nonzero $\Delta$ and constructing confidence intervals for $\Delta$, the normal approximation can lead to tests with supranominal size and confidence intervals with coverage probability well below the nominal confidence level in small and even moderate-sized group sequential trials. To avoid this problem, exact statistical methods for group sequential tests, repeated confidence intervals, and interval estimation after sequential testing are proposed and evaluated in this article. An example using real data is provided in illustration of the approximate and exact procedures.

2. Exact Group Sequential Methods

We will study exact methods for constructing group sequential tests, repeated confidence intervals, and confidence intervals following group sequential trials in Sections 2.1, 2.2, and 2.3, respectively. We start each discussion with a review of the corresponding large-sample methods. All procedures are illustrated with a trial conducted at Mayo Clinic during the years 1958–1973 (O’Brien and Fleming, 1979). The data from this trial are shown in Table 1.

<table>
<thead>
<tr>
<th>Stage ($k$)</th>
<th>Outcome on Prednisone + VCR</th>
<th>Outcome on Prednisone</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Remission</td>
<td>No remission</td>
</tr>
<tr>
<td>1</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>1</td>
</tr>
</tbody>
</table>

2.1 Exact Group Sequential Tests

The $Z$ statistic for $\Delta$ with unpooled variance estimators are given by

$$Z_u(k, \Delta) = \frac{\hat{\Delta}_k - \Delta}{\hat{p}_{Ak}(1 - \hat{p}_{Ak})/m_k + \hat{p}_{Bk}(1 - \hat{p}_{Bk})/n_k}^{1/2}, \quad k = 1, \ldots, K.$$  

A maximum-likelihood group sequential procedure for testing the null hypothesis $H_0: \Delta = \Delta_0$ against two-sided alternatives with an overall significance level $\alpha$ is to reject $H_0$ at stage $k$ if $|Z_u(k, \Delta_0)| \geq c_k$, where $Z_u(k, \Delta_0)$ is the observed value of $Z_u(k, \Delta_0)$, and $c_k; \ k = 1, \ldots, K$ are constants satisfying $\text{Pr}_{\Delta_0}(\cdot) \geq c_k$, for some $1 \leq k \leq K = \alpha$. The evaluation of the probability $\text{Pr}_{\Delta_0}(\cdot)$ is based on the normal approximation to the joint distribution of $(\hat{\Delta}_k; k = 1, \ldots, K)$.

The two best-known forms of group sequential tests are due to Pocock (1977) and O’Brien and Fleming (1979). Pocock (1977) set $c_1 = \cdots = c_K$, whereas O’Brien and Fleming (1979) chose $c_k = (K/k)^{1/2}c_k \ (k = 1, \ldots, K)$. Clearly, the O’Brien–Fleming procedure spends much less Type I error in early stages than the Pocock procedure. Two additional forms of group sequential tests due to Slud and Wei (1982) and Lan and DeMets (1983) are also useful. For the Slud–Wei procedure, one prespecifies the exit probabilities $\{\pi_1, \ldots, \pi_K\}$, which are defined by $\text{Pr}_{\Delta_0}(\cdot) < c_1, \ldots, |Z_u(k - 1, \Delta_0)| < c_{k-1}, |Z_u(k, \Delta_0)| \geq c_k = \pi_k$ and $\sum_{k=1}^K \pi_k = \alpha$, and computes the $c_k$’s sequentially. An increasing
sequence of \( \pi_k \)'s is generally recommended so that \( c_K \) is not much different from the customary fixed-sample-size boundary value. The Slud–Wei approach is especially attractive for unequal and unpredictable group sizes. The method of Lan and DeMets (1983) specifies “Type I error spending functions” that generate a broad class of boundaries based on possibly unequal but independent increments in information without specifying \( K \) in advance.

When the null difference \( \Delta_0 \) is 0, the \( Z \) statistics with pooled variance estimators,

\[
Z_p(k) = \frac{\hat{\Delta}_k}{\sqrt{\hat{\sigma}_k(1 - \hat{\rho}_k)(1/m_k + 1/n_k)}}^{1/2}, \quad k = 1, \ldots, K,
\]

where \( \hat{\rho}_k = (X_k + Y_k)/(m_k + n_k) \), may be used in replacement of the \( Z_u(k, 0) \). In fact, \( Z_p(k) \) \( (k = 1, \ldots, K) \) were the original statistics used by Pocock (1977) and O'Brien and Fleming (1979) in defining their group sequential tests for binary outcomes.

The likelihood-ratio test for testing \( \Delta = \Delta_0 \) is usually preferred to the maximum-likelihood test in small samples mainly because the former does not involve variance estimation. In the group sequential setting, the log-likelihood at the \( k \)th look is given by

\[
l_k(\Delta, p_B) = X_k \log(\Delta + p_B) + (m_k - X_k) \log(1 - \Delta - p_B) + Y_k \log p_B + (n_k - Y_k) \log(1 - p_B),
\]

and the corresponding profile likelihood for \( \Delta \) is

\[
\hat{l}_k(\Delta) = \sup_{\max\{0, -\Delta\} < p_B < \min\{1, 1 - \Delta\}} l_k(\Delta, p_B).
\]

A likelihood-ratio sequential procedure for testing \( H_0 \) is to reject \( H_0 \) at stage \( k \) if

\[
2\{l_k(\hat{\Delta}_k, \hat{p}_B) - \hat{l}_k(\Delta_0)\} \geq c^2_k, \quad \text{where } \{c_k; k = 1, \ldots, K\} \text{ are the same as those of the corresponding maximum-likelihood group sequential test.}
\]

We have conducted extensive numerical studies to evaluate the small-sample properties of the aforementioned approximate group sequential tests. As shown in Table 2, the actual Type I error of the maximum-likelihood procedure with pooled variance estimators (ML1)

<table>
<thead>
<tr>
<th>( m = n = n_t )</th>
<th>( p_A = p_B )</th>
<th>Pocock-type</th>
<th>Slud–Wei-type</th>
<th>O'Brien–Fleming-type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( p_A = p_B )</td>
<td>ML1</td>
<td>ML2</td>
<td>LR</td>
</tr>
<tr>
<td>( 5 )</td>
<td>.1</td>
<td>.065</td>
<td>.035</td>
<td>.033</td>
</tr>
<tr>
<td></td>
<td>.2</td>
<td>.027</td>
<td>.088</td>
<td>.076</td>
</tr>
<tr>
<td></td>
<td>.3</td>
<td>.043</td>
<td>.117</td>
<td>.092</td>
</tr>
<tr>
<td></td>
<td>.4</td>
<td>.048</td>
<td>.137</td>
<td>.095</td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>.047</td>
<td>.146</td>
<td>.094</td>
</tr>
<tr>
<td>( 10 )</td>
<td>.1</td>
<td>.026</td>
<td>.035</td>
<td>.063</td>
</tr>
<tr>
<td></td>
<td>.2</td>
<td>.042</td>
<td>.064</td>
<td>.064</td>
</tr>
<tr>
<td></td>
<td>.3</td>
<td>.049</td>
<td>.072</td>
<td>.065</td>
</tr>
<tr>
<td></td>
<td>.4</td>
<td>.051</td>
<td>.076</td>
<td>.068</td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>.051</td>
<td>.080</td>
<td>.070</td>
</tr>
<tr>
<td>( 15 )</td>
<td>.1</td>
<td>.034</td>
<td>.054</td>
<td>.072</td>
</tr>
<tr>
<td></td>
<td>.2</td>
<td>.048</td>
<td>.072</td>
<td>.062</td>
</tr>
<tr>
<td></td>
<td>.3</td>
<td>.051</td>
<td>.071</td>
<td>.055</td>
</tr>
<tr>
<td></td>
<td>.4</td>
<td>.053</td>
<td>.070</td>
<td>.053</td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>.053</td>
<td>.073</td>
<td>.054</td>
</tr>
</tbody>
</table>

Note: The total number of observations is \( 6m \). ML1 and ML2 refer to maximum-likelihood tests with pooled and unpooled variance estimators, respectively. All calculations are based on the exact distributions of the test statistics. The results are symmetric around the common null value of .5 success probability. For the Slud–Wei-type test, \( \pi_1 = .01 \), \( \pi_2 = .01 \), and \( \pi_3 = .03 \), i.e., \( \alpha_1 + \alpha_2 = .01 \), \( \alpha_2 + \alpha_3 = .02 \), and \( \alpha_1 + \alpha_3 = .05 \). For all exact tests, \( c(\Delta, \Delta) = -c(\Delta, \Delta) \) for all \( k \).
is much smaller than that of the maximum-likelihood procedure with unpoled variance estimators (ML2). This is to be expected because $Z_{q}(k)$ is always smaller than $Z_{q}(k, 0)$ when $m_{k} = n_{k}$ (Suissa and Shuster, 1985). The actual Type I error of the likelihood-ratio procedure (LR) tends to lie between those of ML1 and ML2. The excess of the actual Type I error over the nominal level can be quite substantial for the ML2 and LR procedures. In addition, for these two procedures, the Pocock test tends to be worse than the O’Brien–Fleming and Slud–Wei test (with $\pi_{1} = .01$, $\pi_{2} = .01$, and $\pi_{3} = .03$) because the former spends more Type I error in early stages where cumulative sample sizes are small. Finally, note that both ML1 and ML2 almost always have their lowest attained Type I error when the common null success probability is extreme, whereas LR usually has its highest attained value in this case.

It can be shown that the correlation coefficient between $\hat{\Delta}_{k}$ and $\hat{\Delta}_{l}$ is

$$\rho(\hat{\Delta}_{k}, \hat{\Delta}_{l}) = \left( \frac{p_{A}q_{A}/m_{k} + p_{B}q_{B}/n_{k}}{p_{A}q_{A}/m_{l} + p_{B}q_{B}/n_{l}} \right)^{1/2}, \quad \text{for } l = 1, \ldots, K \text{ and } k = 1, \ldots, l.$$  

Thus, $\rho(\hat{\Delta}_{k}, \hat{\Delta}_{l}) (k \neq l)$ is free of the unknown parameters $p_{A}$ and $p_{B}$ if and only if $m_{k}/n_{k} = m_{l}/n_{l}$; see also O’Brien and Fleming (1979). Since most trials use some form of constrained randomization, it is likely that the ratio of group sizes between two treatments will be nearly constant.

A common solution to the problem of unequal assignment ratios $m_{k}/n_{k} (k = 1, \ldots, K)$ is to pretend that these ratios were equal and use the critical values tabulated for equal group sizes. As Table 3 indicates, however, this practice tends to increase the already inflated Type I error of the ML2 procedure in small and moderate-sized trials.

We now turn to the problem of constructing exact group sequential testing procedures, which always preserve the preassigned Type I error. First of all, there are several ways of eliminating the nuisance parameter, say, $p_{B}$. For example, one may obtain Fisher-type exact tests for testing the null hypothesis that the odds ratio $\theta = p_{A}(1 - p_{B})/(p_{B}(1 - p_{A}))$ is

<table>
<thead>
<tr>
<th>$m_{1}$</th>
<th>$n_{1}$</th>
<th>$m_{2}$</th>
<th>$n_{2}$</th>
<th>$m_{3}$</th>
<th>$n_{3}$</th>
<th>$p_{A} = p_{B}$</th>
<th>Pocock</th>
<th>Slud–Wei</th>
<th>O’Brien–Fleming</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE1</td>
<td>MLE2</td>
<td>LR</td>
<td></td>
<td></td>
<td></td>
<td>MLE1</td>
<td>MLE2</td>
<td>LR</td>
<td></td>
</tr>
<tr>
<td>5 15 10 20 25 30 .1 .043 .077 .054 .036 .097 .070 .038 .095 .069</td>
<td>.2 .045 .174 .073 .048 .139 .067 .048 .110 .064</td>
<td>.3 .049 .173 .067 .049 .145 .063 .053 .099 .058</td>
<td>.4 .054 .137 .066 .051 .128 .061 .052 .102 .051</td>
<td>.5 .057 .119 .068 .054 .119 .063 .051 .105 .050</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 15 20 20 25 35 .1 .039 .065 .071 .051 .061 .066 .042 .067 .070</td>
<td>.2 .050 .157 .057 .052 .103 .048 .044 .064 .058</td>
<td>.3 .049 .171 .056 .049 .134 .050 .047 .078 .055</td>
<td>.4 .050 .137 .064 .051 .121 .060 .052 .095 .053</td>
<td>.5 .051 .119 .064 .052 .108 .065 .053 .100 .053</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 5 10 20 25 25 .1 .028 .038 .055 .044 .065 .085 .052 .058 .085</td>
<td>.2 .041 .101 .085 .047 .115 .060 .051 .090 .064</td>
<td>.3 .047 .127 .098 .056 .123 .064 .052 .087 .066</td>
<td>.4 .055 .145 .097 .059 .125 .062 .057 .094 .062</td>
<td>.5 .056 .152 .094 .058 .127 .061 .052 .096 .056</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note:** ML1 and ML2 refer to maximum-likelihood tests with pooled and unpoled variance estimators, respectively. The critical values for equal group sizes are used. The calculations of the Type I error are based on the exact distributions of the test statistics. The results are symmetric around the common null value of .5 success probability. For the Slud–Wei-type test, $\pi_{1} = .01$, $\pi_{2} = .01$, and $\pi_{3} = .03$, i.e., $\alpha_{1} + \alpha_{2} = .01$, $\alpha_{2} + \alpha_{3} = .02$, and $\alpha_{3} + \alpha_{1} = .03$. 

\[1402\]
unity by conditioning on the realizations of \( \{X_k + Y_k; k = 1, \ldots, K\} \). An alternative approach is to use the maximization method (Basu, 1977) to derive exact unconditional tests for testing \( \theta = \theta_0 \) or \( \Delta = \Delta_0 \). The maximization method caters to the worst possible value of the nuisance parameter \( p_B \) by maximizing the null power function of the selected test over the domain of \( p_B \). Here, we show how to perform exact unconditional tests for testing \( H_0; \Delta = \Delta_0 \). The ideas can also be applied to the cases of exact unconditional and conditional tests for testing \( \theta = \theta_0 \).

Let \( Z_k = Z_{\alpha}(k, 0) \) \( (k = 1, \ldots, K) \) with 0/0 being 0. A two-sided exact unconditional group sequential test is to reject \( H_0 \) at the \( k \)th look if \( z_k \geq \tilde{c}(k, \Delta_0) \) or \( z_k \leq \tilde{c}(k, \Delta_0) \), where \( z_k \) is the observed value of \( Z_k \). The calculations of the critical values \( \tilde{c}(k, \Delta_0), \tilde{c}(k, \Delta_0); k = 1, \ldots, K \) are described below. We loosely label the exact test according to their analogous approximate tests.

**Pocock-type:**

\[
\tilde{c}(k, \Delta) = \inf \left\{ c: \inf_{p_B} \Pr_{\Delta}|Z_1 < c, \text{ for all } 1 \leq l \leq K \geq 1 - \tilde{\alpha}_K \right\},
\]

\[
\tilde{c}(k, \Delta) = \sup \left\{ c: \inf_{p_B} \Pr_{\Delta}|Z_1 > c, \text{ for all } 1 \leq l \leq K \geq 1 - \tilde{\alpha}_K \right\};
\]

**O’Brien–Fleming-type:**

\[
\tilde{c}(K, \Delta) = \inf \left\{ c: \inf_{p_B} \Pr_{\Delta}|Z_1 < (K/l)^{1/2}c, \text{ for all } 1 \leq l \leq K \geq 1 - \tilde{\alpha}_K \right\},
\]

\[
\tilde{c}(K, \Delta) = \sup \left\{ c: \inf_{p_B} \Pr_{\Delta}|Z_1 > (K/l)^{1/2}c, \text{ for all } 1 \leq l \leq K \geq 1 - \tilde{\alpha}_K \right\},
\]

\[
\tilde{c}(k, \Delta) = (K/k)^{1/2}\tilde{c}(K, \Delta),
\]

\[
\tilde{c}(k, \Delta) = (K/k)^{1/2}\tilde{c}(K, \Delta);
\]

**Slud–Wei-type:**

\[
\tilde{c}(k, \Delta) = \inf \left\{ c: \inf_{p_B} \Pr_{\Delta}|Z_1 < \tilde{c}(1, \Delta), \ldots, Z_{k-1} < \tilde{c}(k-1, \Delta), Z_k < c \geq 1 - \tilde{\alpha}_K \right\},
\]

\[
\tilde{c}(k, \Delta) = \sup \left\{ c: \inf_{p_B} \Pr_{\Delta}|Z_1 > \tilde{c}(1, \Delta), \ldots, Z_{k-1} > \tilde{c}(k-1, \Delta), Z_k > c \geq 1 - \tilde{\alpha}_K \right\};
\]

where \( k = 1, \ldots, K ; \tilde{\alpha}_1 \leq \tilde{\alpha}_2 \leq \cdots \leq \tilde{\alpha}_K; \tilde{\alpha}_1 \leq \tilde{\alpha}_2 \leq \cdots \leq \tilde{\alpha}_K; \tilde{\alpha}_K + \alpha_K = \alpha \); and the domain of \( p_B \) for minimizations is \([\max \{0, -\Delta\}, \min \{1, 1 - \Delta\}]\). The probabilities are evaluated from the exact joint distribution of \( \{Z_k; k = 1, \ldots, K\} \), which is determined by the probability law of a binomial variable. An efficient algorithm for this calculation is described in Section 3. If \( K = 1, m_1 = n_1, \text{ and } \Delta_0 = 0 \), then the above procedures reduce to the unconditional test of Suisse and Shuster (1985).

We have included in Table 2 the actual Type I error of the exact unconditional group sequential procedure for testing no treatment effect. As long as the common null value of the two success probabilities is not close to zero or unity, the attained significance level is remarkably close to the nominal one. Thus, in practice the nuisance parameter is not much of a nuisance. It is interesting to note that the exact procedure is only slightly more
conservative than the ML1 procedure. In addition, the exact test seems somewhat more conservative for the Pocock-type boundary than for the other two types of boundaries. The choice of \( Z_k (k = 1, \ldots, K) \) as the test criteria for deriving the exact unconditional tests is somewhat arbitrary. Other statistics could also be used. Our empirical experience, however, indicates that \( Z_k (k = 1, \ldots, K) \) are less conservative than other likely candidates such as unstandardized differences \( \hat{\Delta}_k (k = 1, \ldots, K) \) and odds ratio statistics.

As an example, let us compare the exact and approximate group sequential tests as they might have been used in the Mayo Clinic trial. A maximum of three tests (i.e., \( K = 3 \)) is assumed. The trial itself compared two drugs (prednisone + VCR versus prednisone) in the treatment of leukemia. Success was considered to be remission, which in this case occurs relatively soon following treatment or not at all. To be consistent with O'Brien and Fleming (1979), we will employ one-sided procedures for detecting a positive VCR effect, which are obvious modifications of the aforementioned two-sided tests. However, we usually prefer two-sided monitoring procedures to protect from unexpected effects. For simplicity, we consider only the Slud–Wei-type tests with \( \alpha_1 = \alpha_2 = \alpha_3 = 0, \alpha_1 = .01, \alpha_2 = .02, \) and \( \alpha_3 = .05 \).

In this example, the observed values of the \( Z \) statistics with pooled variance estimators are \( z_{p}(1) \approx .7859, z_{p}(2) \approx 1.9450, \) and \( z_{p}(3) \approx 2.3467, \) and those with unpooleed variance estimators are \( z_{1} \approx .7338, z_{2} \approx 1.7760, \) and \( z_{3} \approx 2.1183. \) The critical values for the one-sided maximum-likelihood sequential tests at the three stages are about 2.3264, 2.2193, and 1.7392, respectively. Using these tests, one would have concluded at the end of the trial that the effect of VCR was beneficial. On the other hand, the critical values for the one-sided exact test are \( \tilde{c}(1, 0) \approx 3.2404, \tilde{c}(2, 0) \approx 2.7634, \) and \( \tilde{c}(3, 0) \approx 2.1223, \) which are substantially higher than the corresponding values of the maximum-likelihood tests. Formally, the null hypothesis would not have been rejected at the .05 level if the exact method had been utilized, which reflects the conservativeness of the exact test.

### 2.2 Exact Repeated Confidence Intervals

The repeated confidence interval (RCI) approach advocated by Robbins (1970), Jennison and Turnbull (1984), and Lai (1984) combines aspects of sequential estimation and testing and allows a full exploration of data at each interim analysis. The RCIs can be obtained by inverting proper group sequential tests. In particular, the (two-sided) maximum-likelihood and likelihood-ratio RCIs with \((1 - \alpha)\) nominal confidence coefficient are, respectively, \( (\hat{\Delta}_k - c_k SE_k, \hat{\Delta}_k + c_k SE_k) \) \((k = 1, \ldots, K)\), where \( SE_k = (\hat{p}_{\alpha k}(1 - \hat{p}_{\alpha k})/m_k + \hat{p}_{\beta k}(1 - \hat{p}_{\beta k})/n_k)^{1/2} \) (Jennison and Turnbull, 1989), and \( |\Delta| = 2[\hat{t}_k(\Delta, \hat{p}_{\alpha k}) - \hat{t}_k(\Delta)] < c_k^2 \) \((k = 1, \ldots, K)\) (Lin and Wei, 1989). Obviously, the pooled variance estimator cannot be used for interval estimation.

Table 2 also indicates that the actual coverage probability of the approximate RCIs, especially that of the maximum-likelihood ones, is generally much smaller than the nominal confidence coefficient for zero \( \Delta \). Additional studies have revealed similar results for nonzero \( \Delta \)'s; see also Lin and Wei (1989). As one might expect, the discrepancies between the actual and nominal coverage probabilities of the maximum-likelihood RCIs can be very alarming for large values of \( \Delta \). For example, under the same group sequential designs as those described in Table 2 and with \( p_{\alpha} = .9 \) and \( p_{\beta} = .1 \), the exact coverage probabilities of the maximum-likelihood RCIs with the Pocock boundary are about .646, .827, and .897 for \( m_1 = 5, 10, \) and 15, respectively, at the .95 nominal confidence level.

The exact group sequential tests studied in the last section can be easily inverted to obtain exact unconditional RCIs for \( \Delta \). The confidence interval for \( \Delta \) at the \( k \)th look consists of \( |\Delta| : \hat{c}(k, \Delta) < z_k < \tilde{c}(k, \Delta) \) \((k = 1, \ldots, K)\). The overall coverage probability of the \( K \) intervals is at least \((1 - \alpha)\). If \( K = 1 \) and the test criterion \( Z_1 \) is replaced by \( \hat{\Delta}_1 \), then
the resulting unconditional confidence interval becomes the “tail method” interval of Santner and Snell (1980).

We now return to the example of the Mayo Clinic trial. Let \( q_1 = q_2 = q_3 = 0, \alpha_1 = .01, \alpha_2 = .02, \) and \( \alpha_3 = .05. \) Then the lower bounds of the one-sided exact RCIs are \( \Delta_1 \approx -.3872, \Delta_2 \approx -.1130, \) and \( \Delta_3 \approx -.0034. \) The corresponding values of the maximum-likelihood RCIs are approximately \(-.3101, -.0624, \) and \(.0426, \) respectively.

2.3 Exact Confidence Interval Following a Group Sequential Test

A confidence interval for the parameter of interest may be derived upon termination of a group sequential trial if an appropriate stopping rule has been strictly enforced. Jennison and Turnbull (1983), Atkinson and Brown (1985), and Duffy and Santner (1987) developed exact procedures for constructing such confidence intervals for a single binomial parameter. Tsiatis, Rosner, and Mehta (1984), Kim and DeMets (1987), and Rosner and Tsiatis (1988) proposed similar methods in the context of Gaussian sampling, which can be used to obtain approximate confidence intervals for the difference \( \Delta \) between two binomial parameters.

We now outline the basic idea of estimation following sequential tests introduced by Siegmund (1978). This idea has been used in most of the papers mentioned above and provides the basis for constructing our exact confidence interval following a group sequential trial. Points on the stopping boundary are ordered in a counterclockwise direction, which corresponds to the notion that early stopping is evidence of a larger value of \( |\Delta| \). A lower \((1 - \tilde{\alpha})\) confidence bound for \( \Delta \) is the smallest value of \( \Delta \) that gives probability at least \( \tilde{\alpha} \) to the event that the test terminates at a boundary point at least as large as that actually observed in this order relation. An upper \((1 - \varrho)\) confidence bound is defined similarly.

The result of a group sequential test can be summarized by the bivariate vector \((W_T, T)\), where \( T \) denotes the number of tests that have been performed when the trial ends, and \( W_T \) denotes the value of the test statistic at the \( T \)th test. Suppose that a \( K \)-stage group sequential test with the lower and upper critical values \( \zeta^*_t \)'s and \( \zeta_t \)'s is utilized in the trial. Then the order relation on the sample space \((W_T, T)\) is defined in the following way. The value \((w_t, t_1)\) is greater than or equal to \((w_{t_2}, t_2)\) if and only if one of the following is true:

(i) \( t_1 = t_2 \) and \( w_{t_1} \geq w_{t_2} \) for \( t_1 = 1, \ldots, K; \)
(ii) \( t_1 < t_2 \) and \( w_{t_1} \geq \zeta_{t_2} \) for \( t_1 = 1, \ldots, K - 1; t_2 = 1, \ldots, K; \) and
(iii) \( t_1 > t_2 \) and \( w_{t_2} \leq \zeta_{t_2} \) for \( t_1 = 1, \ldots, K; t_2 = 1, \ldots, K - 1. \)

Let \((w_t, t)\) be the observed value of \((W_T, T)\). Then the lower and upper limits of the approximate confidence interval for \( \Delta \) are the unique solutions to the equations \( \Pr_{\beta}(W_T, T \geq (w_t, t)) = \tilde{\alpha}, \) and \( \Pr_{\beta}(W_T, T \leq (w_t, t)) = \varrho, \) respectively. Here, the computations of the probabilities are based on the normal approximation to the joint distribution of the \( W_k \)'s (Tsiatis et al., 1984). Note that it is more cumbersome to ascertain the asymptotic distribution of \((W_T, T)\) if the stopping rule is defined on test statistics with pooled variance estimators.

By comparison, the lower and upper bounds of the exact unconditional confidence interval for \( \Delta \) are, respectively,

\[
\Delta = \inf_{-1 \leq \Delta \leq 1} \left[ \Delta: \sup_{\beta} \Pr_{\Delta}(W_T, T \geq (w_t, t)) \geq \tilde{\alpha} \right],
\]

and

\[
\bar{\Delta} = \sup_{-1 \leq \Delta \leq 1} \left[ \Delta: \sup_{\beta} \Pr_{\Delta}(W_T, T \leq (w_t, t)) \geq \varrho \right].
\]
where the domain of \( p_B \) for maximizations is \([\max\{0, -\Delta\}, \min\{1, 1 - \Delta\}]\), and the probabilities are determined by the exact distribution of \((W_T, T)\).

Numerical studies have been carried out to evaluate the small-sample properties of the aforementioned approximate and exact confidence intervals following sequential tests. Some typical results are shown in Table 4. The actual coverage probability of the approximate interval is almost always below the nominal confidence level. The discrepancy tends to increase with the size of \( \Delta \), and can be quite disturbing even for a moderate-sized trial. The stopping rule does not seem to have much impact on the coverage probability of the approximate confidence interval, especially when \( \Delta \) is large. The exact confidence interval is somewhat conservative, but its attained coverage probability tends to be closer to the nominal level than that of the approximate interval. In Table 4, the only configuration where the approximate confidence interval achieves nominal coverage level is \( m_1 = n_1 = 5; p_A = p_B = .1 \). This is where the exact interval is the most conservative.

The lack of a well-defined stopping rule precluded the construction of a valid confidence interval for \( \Delta \) following the Mayo Clinic trial. For purpose of illustration, suppose that the one-sided three-stage Slud–Wei-type exact unconditional test with \( \bar{a}_1 = .01, \bar{a}_2 = .02, \) and \( \bar{a}_3 = .05 \) [i.e., \( c(1, 0) = 3.2404, c(2, 0) = 2.7634, \) and \( c(3, 0) = 2.1223 \) had been used. Based on this stopping rule, \( \{W_T, T \geq (w_t, t)\} \) consists of the following events: (i) \( Z_1 \geq 3.2404 \), (ii) \( Z_1 < 3.2404 \) and \( Z_2 \geq 2.7634 \), and (iii) \( Z_1 < 3.2404, Z_2 < 2.7634, \) and \( Z_3 \geq 2.1183 \). Then the lower bound of the one-sided exact unconditional confidence interval for \( \Delta \) upon termination of the trial would have been \(-.0033\). The corresponding value of the approximate confidence interval would have been \( .0510 \). As one might expect, these two lower bounds, especially the approximate one, are somewhat higher than the corresponding values of the RCIs at the final stage.

It should be pointed out that the order relation defined by Siegmund (1978) is only one

### Table 4

<table>
<thead>
<tr>
<th>( m_1 = n_1 )</th>
<th>( p_A )</th>
<th>( p_B )</th>
<th>Pocock test</th>
<th>Slud–Wei test</th>
<th>O'Brien–Fleming test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Approx.</td>
<td>Exact</td>
<td>Approx.</td>
<td>Exact</td>
<td>Approx.</td>
</tr>
<tr>
<td>5</td>
<td>.9</td>
<td>.5</td>
<td>.879</td>
<td>.979</td>
<td>.879</td>
</tr>
<tr>
<td></td>
<td>.7</td>
<td>.3</td>
<td>.851</td>
<td>.967</td>
<td>.851</td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>.3</td>
<td>.895</td>
<td>.951</td>
<td>.895</td>
</tr>
<tr>
<td></td>
<td>.3</td>
<td>.1</td>
<td>.903</td>
<td>.973</td>
<td>.903</td>
</tr>
<tr>
<td></td>
<td>.3</td>
<td>.3</td>
<td>.917</td>
<td>.969</td>
<td>.924</td>
</tr>
<tr>
<td></td>
<td>.1</td>
<td>.1</td>
<td>.968</td>
<td>.985</td>
<td>.954</td>
</tr>
<tr>
<td>10</td>
<td>.9</td>
<td>.5</td>
<td>.919</td>
<td>.966</td>
<td>.918</td>
</tr>
<tr>
<td></td>
<td>.7</td>
<td>.3</td>
<td>.898</td>
<td>.969</td>
<td>.896</td>
</tr>
<tr>
<td></td>
<td>.5</td>
<td>.3</td>
<td>.917</td>
<td>.957</td>
<td>.917</td>
</tr>
<tr>
<td></td>
<td>.3</td>
<td>.1</td>
<td>.930</td>
<td>.964</td>
<td>.932</td>
</tr>
<tr>
<td></td>
<td>.3</td>
<td>.3</td>
<td>.926</td>
<td>.956</td>
<td>.938</td>
</tr>
<tr>
<td></td>
<td>.1</td>
<td>.1</td>
<td>.946</td>
<td>.967</td>
<td>.937</td>
</tr>
</tbody>
</table>

Note: The maximum-likelihood test uses the unpooled variance estimator. The results are symmetric with respect to reflection around .5 success probability. For the Slud–Wei test, \( \pi_1 = .01, \pi_2 = .01, \) and \( \pi_3 = .03 \). The calculation of the coverage probability for the approximate confidence interval is based on the exact distribution of the confidence set when \( m_1 = n_1 = 5 \) and on 10,000 Monte Carlo replications when \( m_1 = n_1 = 10 \). For the exact interval, each entry is calculated from 1,000 Monte Carlo replications. The standard error of a Monte Carlo estimate is about .002 for 10,000 replications and .007 for 1,000 replications. The random number generator of Press et al. (1986, pp. 196–197) is used.
of several possible relations that could be used. Chang and O’Brien (1986) and Duffy and Santner (1987) obtained confidence intervals for binary outcomes following a group sequential trial by ordering the sample space using the likelihood principle.

3. Remarks

In summary, the exact approach is generally recommended for small and moderate-sized sequential trials when it is desired to test nonzero $\Delta_0$ and/or to construct confidence intervals. If one is interested only in testing the hypothesis that $\Delta = 0$, then the maximum-likelihood sequential test with pooled estimators of variance may be more preferable than the exact procedure, especially when an extreme null success probability is anticipated.

In our computations, the binomial distribution is generated through an efficient network algorithm instead of factorial calculations, which is described as follows. Consider a network of $(n + 1)$ stages, labelled $0, 1, \ldots, n$. At stage $i$ ($i = 0, 1, \ldots, n$), there is a set of nodes of the form $(i, S_i)$, where $S_i$ is the number of successes out of $i$ Bernoulli trials with common success probability $p$. Let $L_i$ be the probability of obtaining the event $(i, S_i)$, and let $\Omega_i$ be the set of all triplets $(i, S_i, L_i)$. Each record $(i, S_i, L_i)$ generates two records $(i + 1, S_i + \delta_{i+1}, L_{i+1})$ with $\delta_{i+1} = 0$ or $1$ and $L_{i+1} = L_i p^{S_i/(1-p)^{1-S_i}}$. The records with the same node $(i + 1, S_{i+1})$ are merged into one. Thus, starting with $\Omega_0 = \{0, 0, 1\}$ and moving recursively through stages $1, \ldots, n - 1$, one obtains $\Omega_n$, which determines the distribution of $S_n$.

The exact unconditional testing and estimation procedures involve the minimization (or maximization) of certain probability functions over the domain of $p_B$. Those functions usually have multiple local minima (or maxima), and the minimization techniques similar to that of Susska and Shuster (1985) should be employed.

Wei et al. (1990) proposed a number of inferential procedures for single-stage randomized play-the-winner designs (Wei and Durham, 1978), which include the complete randomization as a special case. Their exact methods, which are also based on an efficient network algorithm, can be easily modified to accommodate other unconventional designs. By using the ideas presented in Section 2 and Wei et al. (1990), one should be able to derive exact inferential procedures for group sequential trials with any treatment allocation rules.

ACKNOWLEDGEMENTS

This research was supported by NIH grants. The authors are grateful to two referees for their helpful comments and constructive suggestions on earlier versions of the paper. They would also like to thank Professors Jennison and Turnbull for providing their 1989 paper before its publication.

RéSUMÉ

Cet article concerne les essais cliniques comparant deux traitements sur une réponse binaire, où l’on examine périodiquement les données pour pouvoir détecter plus tôt une différence entre traitements. Les méthodes séquentielles pour de tels essais utilisent l’approximation normale asymptotique de la distribution conjointe, sur l’ensemble des analyses intermédiaires, des estimateurs de la différence entre traitements. Au travers de diverses simulations, nous montrons que, pour des échantillons de taille réduite ou même modérée, les approximations de ces procédures peuvent conduire à des risques de première espèce augmentés et à des intervalles de confiance trop étroits (surtout lorsque l’on ne fait pas l’hypothèse d’homogénéité des variances). C’est pourquoi nous proposons des méthodes exactes à la fois pour les tests et pour les intervalles de confiance. Ces nouvelles procédures peuvent être adaptées à toute règle d’attribution des traitements. Un exemple est ensuite présenté sur des données réelles.
REFERENCES


Received May 1989; revised January and May 1990; accepted August 1990.