GROUP SEQUENTIAL TESTS AND REPEATED CONFIDENCE INTERVALS

by

Christopher Jennison*
and Bruce W. Turnbull

To appear as Chapter 12 in Handbook of Sequential Analysis (B.K. Ghosh and P.K. Sen, eds.). Marcel Dekker, New York.

*School of Mathematics, University of Bath, Bath, BA2 7AY, U.K.

This research was supported by Grant GM 28364 from the U.S. National Institutes of Health.
GROUP SEQUENTIAL TESTS AND REPEATED CONFIDENCE INTERVALS

by

Christopher Jennison
and Bruce W. Turnbull

ABSTRACT

Group sequential procedures allow early stopping without the need for continuous monitoring. They have been used in acceptance sampling for many years but their impact in clinical trials did not occur until the 1970's. We describe group sequential one-sided and two-sided tests, their adaptation to allow for unpredictable group sizes, and methods for calculating confidence intervals, P-values and approximately unbiased point estimates on termination. Finally, we describe the repeated confidence interval approach to group sequential testing: here, the role of rigid stopping rule is relaxed and secondary or external information may be considered in reaching a decision for early termination, but still with protection against the multiple looks effect that can arise from repeated analyses.

Keywords: Group sequential test, Interim analyses, One-sided test, Two-sided test, Repeated significance test, Unequal group sizes, Confidence interval, P-value, Point estimate, Repeated confidence intervals, Repeated P-values.

This research was supported by Grant GM 28364 from the U.S. National Institutes of Health.
1. Introduction and historical review

Group sequential procedures offer a compromise between fixed sample and fully sequential procedures in allowing the opportunity of early stopping without the need for continuous monitoring of accumulating data. Typically, most of the reduction in expected sample size afforded by a sequential test can be achieved by a group sequential test with as few as 5 or 10 groups and designs with even fewer groups might well be preferred in applications where interim inspections of data are time consuming or inconvenient. The predominance of fully sequential procedures in the theoretical development of sequential analysis owes much to the derivation of analytic expressions for the properties of certain sequential tests. However, efficient numerical methods for constructing and evaluating group sequential designs are now available and, because of their practical convenience, group sequential procedures should usually be regarded as the methods of choice.

Group sequential procedures have been in routine use in certain applications for many years. Their use in acceptance sampling dates back to the late 1920's and the two-stage plans of Dodge and Romig (1929). The multi-stage plans developed subsequently by the Columbia University Research Group (Freeman, Friedman, Mosteller and Wallis, 1948) went on to form the basis of the United States military standard, MIL-STD 105D (1963). The purpose of these plans is to reach a decision to accept or reject a batch of items on the basis of the proportion of defective items: a similar problem arises in the early stages of drug screening and in phase II clinical trials, where the aim is to identify drugs with at least a certain level of therapeutic efficacy. Two- and three-stage procedures for the case of normal response were discussed by Armitage and Schneiderman (1958), Schneiderman (1961), Dunnett (1961) and Roseberry and Gehan (1964). It is
noteworthy that, in this early work, repeated numerical integration was used to evaluate group sequential tests for normal data: this device now provides the key tool in the construction of many group sequential tests and forms the basis of methods designed to handle unpredictable group sizes and of methods for calculating significance levels and confidence intervals following a group sequential test. Because of the discreteness of the sample space, calculations for sums of binomial variables are relatively straightforward and there is a large literature on group sequential designs for drug screening experiments and Phase II clinical trials with binary response: see the papers by Schultz, Nichol, Elfring and Weed (1973), Elashoff and Beal (1976), Herson (1979), Lee, Staquet, Simon, Catane and Muggia (1979), Fleming (1982), Jennison and Turnbull (1983) and Chang, Therneau, Weiand and Cha (1987).

Our discussion in this paper will concentrate on methods for normal responses and we shall pay particular attention to applications in medical trials. That is not to say that the methods described are only applicable when independent identically distributed normal observations are recorded. Group sequential tests are based on sums of observations and, by the central limit theorem, such sums will usually have an approximately normal distribution. More generally, a test designed for normal responses may be applied whenever a sequence of test statistics has approximately the same joint distribution as a sequence of sums of independent normal variates. Thus, the tests we describe are directly applicable in any situation which gives rise to a sequence of statistics that can be embedded in a Brownian motion. One important example which meets this requirement is the sequence of logrank statistics for testing equality of two survival distributions; for others, see, for example, Sen (1981) and Whitehead (1983).
The use of sequential methods in clinical trials is increasing. One reason for this is the requirement of interim analyses, not just to check for a treatment difference with respect to the primary endpoint, but also to monitor the incidence of adverse side effects as well as the rate of patient accrual and compliance to the study protocol. Clearly, some safeguard is necessary to avoid the excessively high rate of false positive results which would arise from the repeated use of fixed sample size analyses. More importantly, the automatic inclusion of interim analyses in the study design affords an excellent opportunity to employ an efficient group sequential test, tailored to the objectives of a particular trial. The need for group sequential methods in this context is paramount: considerable effort is needed to update records and prepare data for an interim analysis and, for logistical reasons alone, analyses rarely take place more frequently than at six-monthly or yearly intervals.

Initially, the use of sequential methods in comparative clinical trials was confined to fully sequential methods. Various sequential plans and examples of their application are described by Armitage (1975). The shift to group sequential methods did not occur until the 1970's: Elfving and Schultz (1973) proposed group sequential designs for comparing two treatments with binary response; McPherson (1974) suggested that the repeated significance tests of Armitage, McPherson and Rowe (1969) might be used to analyze clinical trial data at a small number of interim analyses; Canner (1977) used Monte Carlo simulation to find critical values of a test statistic in a survival study with periodic analyses. However the major impetus for group sequential methods came from Pocock (1977); he gave clear guidelines for group sequential experimental designs and demonstrated the versatility of the group sequential approach by showing that the nominal significance levels of repeated significance tests for normal responses can be used
reliably for a variety of other responses. Shortly after Pocock's paper, O'Brien and Fleming (1979) proposed a different class of group sequential tests: these two papers together have formed the starting point of recent research.

The development of group sequential methods has progressed rapidly; our main objective in this paper is to review the advances of the last ten years. In Sections 2 and 3 we consider two-sided and one-sided tests respectively. In Section 4 we discuss methods for dealing with the important practical problem of unequal and unpredictable group sizes or, more generally, increments in information between successive analyses. In Section 5 we describe methods for calculating confidence intervals, significance levels and approximately unbiased point estimates following a group sequential test: together, these techniques provide a full statistical analysis upon termination of a sequential test, with proper allowance for the effect of the stopping rule on the sample size.

A recurrent problem in sequential analysis is the dominant role of the stopping rule. Most methods rely for their justification on adherence to a statistical stopping rule but it cannot be guaranteed that the monitoring committee of a clinical trial will abide by such a rule in practice. In Section 6 we describe a more flexible approach to sequential monitoring based on repeated confidence intervals. This approach, which combines aspects of sequential estimation and testing, allows a full exploration of the data at each interim analysis but does not depend on a rigidly enforced statistical stopping rule.

In an article of this length, some omissions are inevitable. We have not attempted to survey the extensive literature on specifically two-stage procedures; the interested reader should consult the review article on this topic by Hewett and Spurrier (1983). We have confined our attention to testing and estimation as
opposed to decision theoretic or selection problems. Throughout, computational
details have been omitted: in many cases, computer programs are available from
the authors of the papers cited.

2. Group sequential tests of a null hypothesis against a two-sided alternative

Our discussion in this section will center on the two-sample problem
described by Pocock (1977). Adaptation to a one-sample problem is
straightforward. In a clinical trial to compare two treatments, A and B,
treatment assignment is such that each consecutive group of 2n patients has n
on each treatment. Responses are normally distributed with known variance $\sigma^2$
and unknown means $\mu_A$ and $\mu_B$ for treatments A and B respectively. Let
$\bar{x}_{A_j}$ and $\bar{x}_{B_j}$ denote the observed mean responses on treatments A and B in
the j'th group of 2n patients. After k groups have been observed, the statistic

$$d_k = \frac{1}{k} \sum_{j=1}^{k} (\bar{x}_{A_j} - \bar{x}_{B_j})$$

is normal with mean $\theta = \mu_A - \mu_B$ and variance $2\sigma^2/(kn)$. Thus, for a single
analysis using these 2nk observations, the two-sided significance test of the null
hypothesis $H_0: \mu_A = \mu_B$ has significance level

$$P_k = 2[1 - \Phi \{|d_k| \sqrt{(kn/2\sigma^2)}\}],$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function.

There is a need for caution in interpreting such a significance level. For
example, if a maximum of 5 groups of 2n patients are possible, the probability
under $H_0$ of at least one of the significance levels $P_1, \ldots, P_5$ falling below 0.05 is
actually as high as 0.142. To compensate for repeated testing, Armitage, McPherson and Rowe (1969) proposed the use of repeated significance tests at a nominal level $\alpha'$ chosen to give an overall type I error $\alpha$. In the group sequential setting with a maximum of $K$ groups we require that under $H_0$

$$\Pr(P_k \leq \alpha' \text{ for some } k = 1, \ldots, K) = \alpha.$$  

Armitage, McPherson and Rowe (1969), McPherson and Armitage (1971) and Armitage (1975) dealt with fully sequential tests; although their results can be applied to the group sequential problem by treating each $(\bar{x}_{A_j} - \bar{x}_{B_j})$, $j = 1, 2, \ldots$, as a single observation, they do not include the low values of $K$ typically used in group sequential studies. McPherson (1974) and Pocock (1977) specifically address the group sequential problem and tabulate values of $\alpha'$ for $\alpha = 0.01$ and 0.05 and $K = 2, 3, \ldots, 10$. Pocock also provides a table of constants which can be used to calculate the value of $n$ needed to obtain power $1-\beta = 0.5, 0.75, 0.9, 0.95$ or 0.99 at a given non-zero value of $\theta = \mu_A - \mu_B$ for given $\alpha$, $K$ and $\sigma^2$.

As an illustrative numerical example, suppose $\sigma^2 = 1$ and a group sequential test with at most 4 groups is required for testing $H_0$: $\theta = 0$ against the two-sided alternative $\theta \neq 0$. A type I error $\alpha = 0.05$ is allowed and power $1-\beta = 0.9$ is to be guaranteed if $\theta = \pm 0.5$. From Table 2 of Pocock (1977) we obtain $n = 2 \times (1.763/0.5)^2 = 24.9$ and, rounding this to the nearest integer, we see that 50 patients, 25 on each treatment, are required in each group. Since there are at most 4 groups, the maximum number of patients that may be needed is 200. The form of the test is quite simple: one stops to reject $H_0$: $\theta = 0$ after the $k$'th group of patients if $|\bar{d}_k| \geq 2.361 \sqrt{2/(25k)}$, corresponding to a two-sided significance level for $H_0$ of 0.0182 or less; if $H_0$ has not been rejected after the 4th group of patients it is accepted.
Table 1 shows some properties of this test. By comparison, the fixed sample size test with the same type one error and power at \( \theta = 0.5 \) requires a total of 168 patients; thus, there are noticeable reductions in expected sample size for values of \( \theta \) sufficiently far from 0.

Pocock (1982) assessed the performance of repeated significance tests with constant nominal significance levels by considering the problem of minimizing expected sample size at a certain value of \( \theta \), subject to attaining power \( 1-\beta \) at this \( \theta \). Minimization was amongst all group sequential tests with a fixed number of equally sized groups, type I error \( \alpha \) and power \( 1-\beta \) at the specified \( \theta \); no constraint was placed on the common group size. Pocock found that for \( 1-\beta = 0.9 \) or 0.95, the expected sample size of the repeated significance test is within a few percent of the optimum and argued that comparisons at high values of \( 1-\beta \) are most important, since the ethical imperative to reach an early decision is greatest when the treatment difference is large.

It is still possible that other considerations will lead an experimenter to choose a different form of test. One argument for wider boundaries at early analyses is that, despite a low p-value, the broader community may be skeptical of conclusions based on only a few observations. Another issue is a possible constraint on the maximum sample size. It was tacitly assumed in the previous example that the experimenter's choice lay between a non-sequential study with 168 patients and a group sequential study with a maximum of 200 patients. Suppose now that the study concerns a rare disease and there is a bound on the total number of patients available. If this bound is below 200 the group sequential test is not feasible. On the other hand, if a total of 200 patients is available, the
experimenter could conduct a non-sequential test with this sample size and type I error 0.05 and achieve power 0.94 at \( \theta = 0.5 \). Thus, if the maximum sample size is fixed, the group sequential test obtains savings in expected sample size at the cost of a reduction in power.

O'Brien and Fleming (1979) proposed a test with stricter criteria for early stopping than the repeated significance test. Consequently, this test has only a slight loss of power relative to the non-sequential test with the same maximum sample size. Whereas the repeated significance test rejects \( H_0 \) after the \( k \)th group of \( 2n \) patients if

\[
|\bar{d}_k| \geq C_P \sqrt{\frac{2\sigma^2}{nk}} \quad k = 1, \ldots, K,
\]

the O'Brien and Fleming test rejects \( H_0 \) after the \( k \)th group if

\[
|\bar{d}_k| \geq \frac{C_B}{\sqrt{k}} \sqrt{\frac{2\sigma^2}{nk}}.
\]

The constants \( C_P \) and \( C_B \) depend on \( K \) and are chosen to achieve the overall type I error probability \( \alpha \). Values of \( C_P(K,\alpha) \) and \( C_B(K,\alpha) \) for \( \alpha = 0.05 \) and \( 0.1 \) and \( K = 2, 3, 4, 5 \) and 10 are given in Table 2.

(Table 2 about here)

For the O'Brien and Fleming test, the significance level needed to reject \( H_0 \) increases from a very low value at group 1 to a value very close to the overall level, \( \alpha \), at group \( K \). For our earlier example with 4 groups, \( \sigma^2 = 1 \) and \( \alpha = 0.05 \), we must now use \( C_B = 4.408 \) and \( n = 22 \) patients per treatment arm in each group are required for power 0.9 at \( \theta = 0.5 \); thus the maximum number of patients, 176, is only 8 more than the fixed sample size. A comparison of expected sample sizes over a range of values of \( \theta \) is presented later in Table 3.
Even stricter requirements for early stopping have been suggested by Haybittle (1971) and Peto et al (1976). These authors recommend stopping to reject $H_0$ before the final group of patients only if $|\bar{d}_k|$ is 3 standard deviations from zero; a fixed sample level $\alpha$ test may then be employed at the final analysis or a small correction included if desired.

Wang and Tsiatis (1987) present a family of tests which includes both the Pocock and the O'Brien and Fleming tests as well as intermediate designs. For the test indexed by parameter $\Delta$, $H_0$ is rejected after group $k$ if $|kd_k|$ exceeds a critical value proportional to $k^\Delta$; $\Delta = 0$ corresponds to the O'Brien and Fleming test and $\Delta = 0.5$ to the Pocock test. Wang and Tsiatis tabulate values of $\Delta$ which define optimal tests within this family in terms of minimizing expected sample size at specific values of $\theta$ subject to size and power conditions.

Fleming, Harrington and O'Brien (1984) present a family of tests including some similar to the O'Brien and Fleming test as well as others which allow somewhat greater opportunity for early stopping. Let $\pi_k$ denote the unconditional probability of rejecting $H_0$ at the $k$'th analysis: the tests are defined by specifying $\pi_1 = \ldots = \pi_{K-1} = \pi$ and $\pi_K = \alpha - (K-1)\pi$, where $\pi = \mu\alpha/(K-1)$ for a value of $\mu \in \{0.1, 0.2, \ldots, 0.5\}$. Defining a test in this way is particularly helpful if the observed group sizes are unpredictable and we shall return to these tests in Section 4.

As an illustration of the qualitative differences between the various tests described above, Table 3 shows standardized boundaries, maximum sample sizes and expected sample sizes for seven group sequential tests which guarantee the size and power requirements of our previous example in which $\sigma^2 = 1$. Notice that, as mentioned earlier, the Pocock test yields the lowest expected sample size at $\theta = \pm 0.5$, where the power is equal to 0.9.
Clearly, there is substantial freedom in choosing the group sequential test best suited to a particular study. The experimenter must assess the importance of reducing the sample size under different values of $\theta$ and also bear in mind the likelihood that each value of $\theta$ might arise. In choosing the number of groups in which to collect observations, reductions in expected sample size must be weighed against administrative ease. Pocock (1982) and McPherson (1982) investigate this issue for repeated significance tests and conclude that 4 or 5 groups will usually suffice.

So far, we have assumed each group of patients will be of the same size, $2n$, but there may be some merit in relaxing this constraint, particularly if the number of groups is small. Elashoff and Reedy (1984) consider this question for tests with just two groups. Another possible modification is the introduction of an inner boundary to allow early termination with acceptance of the null hypothesis: group sequential tests with this feature have been proposed by Gould and Pecore (1982) and by Whitehead and Stratton (1983, Sec. 5).

The numerical results in this section were obtained using methods for the evaluation of multiple normal integrals similar to those originally described by Armitage, McPherson and Rowe (1969). Although it may be straightforward to use similar methods to derive group sequential tests for other response distributions, tests for normal response can often be used directly. Pocock (1977) suggests using the nominal significance level $\alpha'$ calculated for a repeated significance test of normal data for other types of response and he demonstrates that this gives tests with approximately the desired type I error for binary and exponential responses and for the repeated $t$–test and the repeated Wilcoxon test. Other forms of group sequential test can be generalized in the same way: at
analysis k one determines the nominal significance level, $\alpha'_k$, that would be
required of normal data at this point and $H_0$ is rejected if the appropriate test
statistic has a significance level of $\alpha'_k$ or less.

3. Group sequential tests of a null hypothesis against a one-sided alternative

DeMets and Ware (1980) note that many clinical trials designed to assess
whether a new therapy is superior to a standard have a one-sided alternative. If
the new therapy is to be abandoned if no advantage is demonstrated in this one
trial, it is inappropriate and even unethical to continue the trial to determine
whether the new therapy is actually inferior to the standard (more generally, a
non-zero point of equivalence for the major response may be chosen to compensate
for subsidiary differences in, for example, side-effects, cost or ease of
administration; see Schwartz, Flamant and Lellouch, 1980, pp. 50–51). In the
example of Section 2, tests have three possible conclusions: rejection of the null
hypothesis of no treatment difference in favor of treatment A superior to
treatment B, rejection of the null hypothesis in favor of treatment B superior to
treatment A, and acceptance of the null hypothesis. But, if treatment A is the
new therapy, the last two conclusions will lead to the same action, abandonment of
treatment A, and they should therefore be combined.

DeMets and Ware (1980) consider the group sequential design with normal
response described in Section 2 but with treatment arms labelled E, denoting the
experimental therapy, and C, denoting the standard or control therapy. After
the k'th group of 2n patients, the test statistic

$$\bar{d}_k = \frac{1}{k}\sum_{j=1}^{k} (\bar{x}_{Ej} - \bar{x}_{Cj})$$
is normally distributed with mean $\theta = \mu_E - \mu_C$ and variance $2\sigma^2/(kn)$. In the one-sided formulation the simple null hypothesis, $\theta = 0$, is replaced by the composite hypothesis, $H_0: \theta \leq 0$ and the alternative is $H_A: \theta > 0$. DeMets and Ware (1980) propose group sequential tests with type I error $\alpha$ when $\theta = 0$ and power $1 - \beta$ at a specific positive value of $\theta$. They consider two methods which are modifications of two-sided repeated significance tests and a third motivated by Wald's (1947) sequential probability ratio test. In a subsequent paper, DeMets and Ware (1982) propose tests with more stringent requirements for very early stopping based on the O'Brien and Fleming two-sided test.

A common feature of DeMets and Ware's methods is their lack of symmetry. This gives rise to additional parameters in the test boundaries and apparent arbitrariness in the choice of tests. Clearly the problem as formulated does not regard the two treatments symmetrically. However, this does not prohibit some symmetry in the problem formulation: if $\beta = \alpha$, the problem of testing $H_0: \theta \leq 0$ against $H_A: \theta > 0$ with type I error $\alpha$ when $\theta = 0$ and type II error $\beta$ at $\theta = \Delta$ is symmetric about $\theta = \Delta/2$. We would recommend that a power of $1-\alpha$ be used in the problem specification in order to take advantage of the mathematical simplifications afforded by a symmetric formulation. It is instructive to remember that size and power requirements do nothing more than specify two points of a test's operating characteristic and empirical evidence suggests that any two sequential tests satisfying the same size and power conditions will have almost identical operating characteristics over the whole range of parameter values.

As an aside, we note that there are problems which are intrinsically asymmetric, since early stopping may only be appropriate under one conclusion. Gould (1983) remarks that, in trials for non-life-threatening conditions, early
stopping for negative results is desirable but if interim findings suggest positive
efficacy of a new therapy, a trial should continue to completion in order to provide
adequate information on safety, tolerability and secondary aspects of efficacy.

Before moving on to describe further tests we shall introduce some new
notation. We suppose that at a study's k'th interim analysis the pair of statistics
\((Z_k, V_k)\) summarizes response on the major endpoint. Here, \(V_k\) is a measure of
the observed information and, conditional on \(V_k\), \(Z_k \sim N(\theta V_k, V_k)\); also,
conditional on the sequence \(\{V_1, V_2, \ldots\}\), increments \(Z_k - Z_{k-1}\) \((k = 2, 3, \ldots)\) are
independent. An equivalent definition is that, conditional on the sequence
\(\{V_1, V_2, \ldots\}\), the joint distribution of \(\{Z_1, Z_2, \ldots\}\) is that of a standard Brownian motion with drift \(\theta\) observed at times \(\{V_1, V_2, \ldots\}\) in the Brownian motion
timescale. Our earlier example fits into this general formulation: setting
\(V_k = kn/(2\sigma^2)\) and \(Z_k = kn\bar{d}_k/(2\sigma^2)\) we obtain \(Z_k \sim N(\theta V_k, V_k)\). Whitehead
(1983, ch. 3) provides formulae for \(Z_k\) and \(V_k\) for the normal approximation to
a variety of response types and presents a powerful general theory, based on the
likelihood function, in which \(V_k\) represents observed Fisher information and \(Z_k\)
is the efficient score statistic.

Whitehead and Stratton (1983) adapt a fully sequential one-sided test to
the group sequential setting. In our new notation, they test \(H_0: \theta \leq 0\) against
\(H_A: \theta > 0\) with type I error \(\alpha\) at \(\theta = 0\) and power \(1-\alpha\) at \(\theta = \Delta\). It is
assumed that \(V_k = kI\) \((k = 1, 2, \ldots)\) for some constant \(I\), corresponding to equal
group sizes in the case of a simple normal response. The test has a triangular
continuation region: \(H_0\) is rejected at analysis \(k\) if \(Z_k > a + (\Delta/4)V_k\) and
accepted if \(Z_k < -a + (3\Delta/4)V_k\) where \(a = -(2/\Delta)\log(2\alpha) - 0.583 \sqrt{I}\). Putting
\(I = 0\) yields the continuation region of a fully sequential test which meets the size
and power requirements exactly (for an elegant proof of this fact see Lorden, 1976).
For a group sequential test, the term $0.583 \sqrt{I}$ (due to Siegmund, 1979) is needed to correct for overshoot. For a maximum number of groups, $K$, the required group size can be found by solving for $I$ the equation

$$a + (\Delta/4)V_K = -a + (3\Delta/4)V_K$$

which ensures that a decision will be reached at analysis $K$.

An attractive property of Whitehead and Stratton's triangular test is that the fully sequential version very nearly attains the minimum possible expected sample size at $\theta = \Delta/2$ for tests with the specified size and power. Similarly, the optimality of Wald's sequential probability ratio test in minimizing the expected sample size at $\theta = 0$ and $\Delta$ motivated DeMets and Ware's group sequential adaptation. However, a group sequential test with a small number of groups is determined by a few parameters, namely the boundary points at each analysis, and modern computers make it possible to search directly for an optimal group sequential test. Moreover, one is free to define the most suitable objective function to be minimized for a particular study and to choose the number of groups and maximum sample size by examining their effect on the minimized objective function. Jennison (1987) describes a method for finding optimal tests with up to 10 groups of observations for a range of optimality criteria. For earlier work on optimal two–stage procedures see Spurrier and Hewett (1975), Hald (1975), Colton and McPherson (1976) and references therein.

Using the above notation, Jennison (1987) considers tests of $H_0$: $\theta \leq 0$ against $H_A$: $\theta > 0$ with type I error $\alpha$ at $\theta = -\delta$ and power $1-\alpha$ at $\theta = \delta$ (although these tests are symmetric about $\theta = 0$ rather than $\theta = \Delta/2$, it is straightforward to adapt them to our earlier problem). A maximum number of analyses, $K$, and maximum total information, $V_{\text{max}}$, are specified and it is assumed that $V_k = (k/K)V_{\text{max}}$ for $k = 1, \ldots, K$. Jennison derives tests which
minimize one of four objective functions amongst all group sequential tests with the specified $K$ and $V_{\text{max}}$ which meet the size and power requirements. The value of the objective function for any one test is obtained by numerical integration and minimization is by a numerical search over the space of possible boundary vectors. The objective functions are $F_1 = E(V|\theta = 0)$, $F_2 = E(V|\theta = \delta)$, $F_3 = E(V|\theta = 2\delta)$ and

$$F_4 = \frac{1}{5} \sum_{i=0}^{4} E(V|\theta = i\delta/2),$$

where $V$ denotes the value of $V_k$ on termination. In studies with immediate response $E(V)$ will generally be proportional to the expected sample size.

[Table 4 about here]

Results for objective function $F_2$ and $\alpha = 0.05$ are shown in Table 4 with entries presented as fractions of the information, $V_f = \{\Phi^{-1}(1-\alpha)/\delta\}^2$, required by a fixed sample test. Note that the optimal value of $V_{\text{max}}$ increases with $K$ but even for $K = 10$ there is little advantage in taking $V_{\text{max}}$ larger than 1.1 or 1.2 times the fixed sample size. In the limit as $K \to \infty$ the optimal test for objective function $F_2$ is known to be the sequential probability ratio test, which attains an expected sample size under $\theta = \pm \delta$ of 0.490 times the fixed sample size. Thus, it is apparent that most of the available reductions in expected sample size are achieved by group sequential tests with as few as 5 or 10 groups.

[Table 5 about here]

The boundaries of a set of Jennison's tests are listed in Table 5. The tests have error rates $\alpha = 0.05$ at $\theta = \pm 1.64485$, the value for which the fixed sample test requires a single unit of information, i.e., $V_f = 1$. All the tests have
\[ V_{\text{max}} = 1.1 \] and, therefore, \[ V_k = (k/K)1.1 \] for \( k = 1, \ldots, K \); they stop to reject \( H_0 \) at analysis \( k \) if \( Z_k > c_k \) and stop to accept \( H_0 \) if \( Z_k < -c_k \). Adaptation to a general value of \( \delta \) is straightforward: for a test with error rates \( \alpha = 0.05 \) at \( \theta = \pm \delta \), take \[ V_{\text{max}} = 1.1 \times (1.64485/\delta)^2, \quad V_k = (k/K)V_{\text{max}}, \] and replace \( c_k \) in the stopping rule by \( c'_{k} = (1.64485/\delta)c_k \). These particular tests were chosen for their good performance over a range of parameter values: each test achieves a value of \( F_4 \) within 0.001 of the minimum possible for the same value of \( K \) and \( V_{\text{max}} = 1.1 \). A parametric description of these tests by means of an "error spending" function will be given in Section 4. The value \( V_{\text{max}} = 1.1 \) allows most of the possible reductions in expected sample size for a minimal increase in maximum sample size beyond that of the fixed sample test.

We conclude this section with an illustrative numerical example. In the problem of comparing a new therapy with a standard, described at the beginning of this section, suppose that it is required to test \( H_0: \theta = 0 \) against \( H_A: \theta > 0 \) with type I error 0.05 at \( \theta = 0 \) and power 0.95 at \( \theta = 0.4 \). Suppose also that \( \sigma^2 = 1 \) and observations are to be taken in 5 groups. The symmetric version of this problem is to test \( \theta' = -0.2 \) against \( \theta' = 0.2 \) where \( \theta' = \theta - 0.2 \). To use the boundary for \( K = 5 \) from Table 5 we require \[ V_{\text{max}} = 1.1 \times (1.64485/0.2)^2 = 74.40. \] Since the information after \( k \) groups of \( 2n \) patients is \( V_k = kn/(2\sigma^2) = kn/2 \) and \( V_{\text{max}} = V_5 \), we require \( 5n/2 = 74.40 \), i.e., \( n = 29.76 \). On rounding, we see that 30 patients must be allocated to each treatment arm within each group and \( V_k = 15k \) for \( k = 1, \ldots, 5 \). Since \( \delta = 0.2 \), the boundary values of Table 5 must be multiplied by a factor of 1.64485/0.2 and these values must then be centered about \( 0.2V_k \) to obtain critical values for \( Z_k = kn\delta_k/2 \). Simple computation shows the pairs of lower and upper boundary points for \( Z_k \), \( k = 1, \ldots, 5 \), to be \( 3.0 \pm 7.08, 6.0 \pm 7.04, 9.0 \pm 6.29, 12.0 \pm 4.62 \) and \( 15.0 \pm 0 \).
The fixed sample test with error probability 0.05 at $\theta = 0$ and $\theta = 0.4$ requires 272 patients. By comparison, this sequential test has a maximum possible sample size of 300, but an expected sample size of only 220 at $\theta = 0.2$ and 169 at $\theta = 0$ or 0.4. The value of $F_4$, the average of the expected sample sizes at $\theta = 0.2, 0.3, 0.4, 0.5$ and 0.6, is 168 which is, by construction, the minimum possible value for a test with 5 groups of 60 patients meeting the error probability requirements.

4. Unequal group sizes

It is often not possible to achieve equal numbers of observations or, more generally, equal increments in information between analyses. In a clinical trial, interim analyses are often arranged at equally spaced intervals in calendar time but the rate of accrual of information may not be constant throughout the study period and, thus, increments in information between analyses will be unequal. For example, in a two treatment survival study, the observed information is approximately proportional to the total number of deaths to date: if the individual failure rates are approximately constant over time, the rate of accrual of information will increase initially as more patients enter the study and then decline in the followup phase after patient entry has ceased.

[Table 6 about here]

We start by considering two-sided tests. In his original paper, Pocock (1977) suggested that small variations in group sizes might be ignored and the nominal significance levels appropriate to equally sized groups employed at each analysis. Table 6 shows the actual type I error probabilities for examples with normal response in which this strategy is applied both to Pocock and to O’Brien and Fleming two-sided tests with five groups of observations. Throughout this section, we continue to use the general notation of Section 3. Thus, the $V_k$. 
k = 1, ..., 5, are proportional to the total numbers of observations in the first k groups. The results are satisfactory for sequences \( V_1, ..., V_5 \) close to the sequence 1,2,3,4,5 but the final two cases demonstrate that problems can sometimes arise. Note that a simple rescaling argument shows that the Type I error probability is the same for all sequences with the same set of ratios \( V_1: V_2: \ldots : V_5 \).

Slud and Wei (1982) present an exact solution to this problem in which the total type I error is partitioned between analyses. For a study with \( K \) analyses, probabilities \( \pi_1, ..., \pi_K \), summing to \( \alpha \), are specified and critical values for the statistics \( Z_k \), \( k = 1, ..., K \), found such that the unconditional probability of wrongly rejecting the null hypothesis at analysis \( k \) is equal to \( \pi_k \). These critical values are calculated successively using numerical integration; the \( k \)'th value depends on \( V_1, ..., V_k \) but not on the as yet unobserved \( V_{k+1}, ..., V_K \).

A similar approach is proposed by Lan and DeMets (1983) and extended by Kim and DeMets (1987a). Whereas Slud and Wei specify the probabilities \( \pi_1, ..., \pi_K \) at the outset, Lan and DeMets spend type I error at a prespecified rate in the information scale. Before implementing the Lan and DeMets method, a maximum information level, \( V_{\text{max}} \), must be determined: this could be the information needed to achieve a certain power or an estimate of the maximum information that will eventually be observed. The type I error is then partitioned according to an "error spending" function, \( f(t) \), where \( f(t) \) is nondecreasing, \( f(0) = 0 \) and \( f(t) = \alpha \) for \( t \geq 1 \). The error allocated to analysis \( k \) is \( \pi_k = f(V_k/V_{\text{max}}) - f(V_{k-1}/V_{\text{max}}) \), for \( k = 1, ..., K \), and critical values for the \( Z_k \) are computed successively as in Slud and Wei's method. Both methods can be used to define symmetric two-sided tests of a null hypothesis \( H_0: \theta = 0 \) or asymmetric tests with different error probabilities for rejection in favor of \( \theta < 0 \) and \( \theta > 0 \); in the latter case separate partitions, \( \pi_1, ..., \pi_K \), or error spending
functions, \( f(t) \), must be specified for the two types of rejection. The Lan and DeMets method has flexibility in that the number of analyses need not be specified in advance. However, it is necessary to specify the target information level, \( V_{\text{max}} \), and this may be troublesome if there is uncertainty about the rate of patient recruitment or the overall response or failure rate.

In choosing the partition \( \tau_1, \ldots, \tau_K \), for Slud and Wei’s method or the error spending function for Lan and DeMets’ method, one should consider the same efficiency criteria as were applied to tests for equally sized groups in Section 2. Note that the methods of Fleming, Harrington and O’Brien (1984), described in Section 2, are defined in terms of error probabilities \( \tau_1, \ldots, \tau_K \) and, thus, provide a useful family from which to choose a Slud and Wei type procedure. Alternatively, the family of functions \( f(t) = \alpha t^\rho \) (\( \rho > 0 \)) provides a good range of Lan and DeMets procedures and includes boundaries roughly the same as the Pocock and the O’Brien and Fleming tests at \( \rho = 0.8 \) and \( \rho = 3 \) respectively.

[Table 7 about here]

It is still possible to plan a test with specified power when group sizes are unequal. At the planning stage it is sufficient to consider equally sized groups, since the actual power, conditional on the observed sequence \( V_1, \ldots, V_K \), is robust to fluctuations in group size as long as the final value, \( V_K \), is close to the target, \( V_{\text{max}} \). Table 7 shows the actual type I error and power, conditional on various sequences \( V_1, \ldots, V_5 \), achieved by Slud and Wei and Lan and DeMets two-sided tests. Since the Slud and Wei test allocates error probability equally between the five analyses and the Lan and DeMets test has a linear error spending function, the two tests coincide if the \( V_k \) are equally spaced with \( V_5 = V_{\text{max}} \). Both tests are designed to achieve power 0.9 at \( \theta = 0.489 \), a requirement that is met if \( V_k = 10k \) for \( k = 1, \ldots, 5 \). Results for sequences 2 and 3 demonstrate that, even
when group sizes are far from equal, power is preserved almost exactly as long as $V_{\text{max}} = 50$. However, the last two examples show that, as in a fixed sample study, power is reduced if insufficient data are obtained and increased if more data than expected are collected. Note that for sequence 4 the Lan and DeMets test fails to use the full type I error, since $V_{5}$ falls short of $V_{\text{max}}$: if, however, further data were available, such a study could be continued and further analyses conducted until the information level reached $V_{\text{max}}$. A shortcoming of the Lan and DeMets test is apparent in sequence 5: since $V_{4}$ is close to $V_{\text{max}}$, very little type I error is left for the final analysis with a consequent loss of power relative to the Slud and Wei test.

Fleming, Harrington and O'Brien (1984) suggests a modification to their procedure which allows the number of analyses to be increased if information is observed to accumulate too slowly. Suppose a total of $K$ analyses are planned with allocation of type I error $\pi_1 = ... = \pi_{K-1} = \mu \alpha/(K-1)$ and $\pi_K = (1-\mu)\alpha$ for some chosen value of $\mu$. If, after $k$ analyses, it is decided to increase $K$ to $K'$, the error allocated to the final analysis is preserved by setting $\pi_{K'} = (1-\mu)\alpha$ and the rest of the type I error is redistributed amongst the other remaining analyses through $\pi_{k+1} = ... = \pi_{K'-1} = \mu \alpha (K-k-1)/\{(K' - k - 1)(K-1)\}$. A similar modification could be applied to the Lan and DeMets method by redefining $V_{\text{max}}$ and $f(t)$ at an intermediate point in the trial.

Fleming, Harrington and O'Brien (1984) point out that it is allowable to modify a sequential design on the basis of observed information levels but warn against altering a design on the basis of observed response. Their comments apply quite generally to the whole spectrum of group sequential tests and their warning should not be taken lightly. Properties of tests are guaranteed conditionally for any sequence $\{V_1, V_2, ...\}$, under the assumption that the sequence of statistics
\{Z_1, Z_2, \ldots\} has the specified multivariate normal distribution. The sequence
\{V_1, V_2, \ldots\} may be a realization of a random process, for example, group sizes
may have a Poisson distribution, but it must in no way depend on the sequence
\{Z_1, Z_2, \ldots\}. The following examples illustrate how the type I error can be affected
if \( V_k \) is allowed to depend on \( Z_1, \ldots, Z_{k-1} \).

Example 1. An incorrect Lan and DeMets test.

Observations are to be taken in three equally sized groups, each
contributing one unit of information. Thus, statistics \( Z_1 \sim N(\theta, 1), Z_2 \sim N(2\theta, 2) \)
and \( Z_3 \sim N(3\theta, 3) \) will be available. A size 0.05 Lan and DeMets two-sided test
with \( f(t) = 0.05 \min(t,1) \) and \( V_{\text{max}} = 3 \) rejects \( H_0 \) if \( |Z_1| > 2.394, \)
\( |Z_2| > 3.244 \) or \( |Z_3| > 3.810 \). Suppose now that the experimenter decides to
omit the second analysis and pool the second and third groups of observations if
\( |Z_1| < 1.2, \) in which case the critical value for \( |Z_3| \) is taken to be that of the
size 0.05 Lan and DeMets test based on \( Z_1 \) and \( Z_3 \) only, namely 3.596. The
overall test is as follows:

if \( |Z_1| > 2.394 \) reject \( H_0 \),
if \( 1.2 < |Z_1| < 2.394 \) take two further groups of observations and reject
\( H_0 \) if \( |Z_2| > 3.244 \) or \( |Z_3| > 3.810 \),
if \( |Z_1| < 1.2 \) take a single large group of observations and reject \( H_0 \) if
\( |Z_3| > 3.596 \).

Numerical integration shows that this test has type I error probability 0.0525, a
small but significant increase over the intended value 0.05. A similar four stage
procedure in which groups 2 and 3 are pooled if \( |Z_1| < 1.2 \) and groups 3 and 4
are pooled if \( |Z_2| < 1.2 \sqrt{2} \) has type I error 0.0522.
Example 2. An incorrect Slud and Wei test.

In the original design, two equally sized groups of observations are to be taken, each contributing one unit of information. Thus, statistics $Z_1 \sim N(0, 1)$ and $Z_2 \sim N(2\theta, 2)$ will be available. A size 0.05 Slud and Wei two-sided test with $\pi_1 = \pi_2 = 0.025$ rejects $H_0$ if $|Z_1| > 2.241$ or $|Z_2| > 3.005$. Suppose now that the experimenter decides to take fewer observations in the second group if $|Z_1|$ is large and more if $|Z_1|$ is low, using, in each case, the critical value for a Slud and Wei test with the observed value of $V_2$. One scheme of this type is as follows:

- if $|Z_1| > 2.241$ reject $H_0$,
- if $1.5 \leq |Z_1| \leq 2.241$ take half a group of observations in stage 2, $V_2 = 1.5$ and $H_0$ is rejected if $|Z_2| > 2.542$,
- if $1.0 < |Z_1| < 1.5$ take a single group of observations in stage 2, $V_2 = 2$ and $H_0$ is rejected if $|Z_2| > 3.005$,
- if $|Z_1| \leq 1.0$, take a double group of observations in stage 2, $V_2 = 3$ and $H_0$ is rejected if $|Z_2| > 3.753$.

The type I error for the above procedure is 0.058. If a value $V_2 = 4$ is used when $|Z_1| < 1.0$ the error probability rises to 0.060.

The first example is realistic in that one might expect a researcher to examine accumulating data more frequently when there are signs of a significant treatment difference. The strategy of Example 2 is a more blatant attempt to increase the probability of rejecting the null hypothesis: if $|Z_1|$ is small, a large value of $V_2$ reduces the correlation between $Z_1$ and $Z_2$ and increases the
chance of rejecting $H_0$ at the second analysis; for large $|Z_1|$, a small value of $V_2$ increases the correlation between $Z_1$ and $Z_2$ and, again, increases the chance of rejecting $H_0$. Just how large a type I error can be produced by a data dependent choice of $\{V_1, V_2, \ldots\}$ is unknown but one might expect the scope for manipulating a sequential design to increase with the number of analyses. Our two examples indicate the potential for serious inflation of type I error and add support to the warnings of others.

Methods for group sequential one-sided tests with unpredictable group sizes have also been developed. Jennison (1987) adapts the Lan and DeMets approach to this problem and uses a parametric family of error spending functions to specify tests which are nearly optimal according to his various criteria. For a test of $H_0: \theta \leq 0$ against $H_A: \theta > 0$ with error probabilities $\alpha$ at $\theta = -\delta$ and $\delta$, let $V_f = \{\Phi^{-1}(1-\alpha)/\delta\}^2$ denote the information required by a fixed sample test. A target information level $V_{\text{max}} = t_{\text{max}} V_f$ is specified and the symmetric group sequential boundary is constructed so that the probability of an error under $\theta = -\delta$ at or before the $k$'th analysis is $f(V_k/V_f)$, where

$$f(t) = \begin{cases} 
  \frac{\alpha}{1 + \exp\left\{- \left[ - \frac{b_1}{t} + \frac{b_2}{t_{\text{max}}-t} + b_3 t + b_4 \right]\right\}} & 0 < t < t_{\text{max}} \\
  \alpha & t \geq t_{\text{max}}
\end{cases}$$
The parameter $t_{\text{max}}$ governs the maximum required sample size and $b_1, b_2, b_3,$ and $b_4$ determine the shape of the test boundary. At the design stage it is convenient to assume a fixed number of analyses, $K$, and equally spaced information levels $V_k = (k/K)V_{\text{max}}, k = 1, ..., K$. Then, given $b_1, b_2$ and $b_3$, a value can be chosen for $b_4$ so that the upper and lower boundary values at analysis $K$ coincide at 0 and the test must terminate at this point.

[Table 8 about here]

Jennison tabulates values of $b_1, b_2$ and $b_3$ which determine nearly optimal tests for the criteria discussed in Section 3. The tests of Table 5 are of this form with $t_{\text{max}} = 1.1$, $b_1 = 0.3$, $b_2 = 0.08$, $b_3 = 1.8$ and $b_4$ as shown in Table 8. Jennison discusses the performance of these tests under various observed sequences $\{V_1, V_2, \ldots\}$ and concludes that their optimality properties are highly robust to variations in the observed information levels.

Whitehead (1983, Sec. 4.7) uses a quite different method to adapt the triangular test to unequal group sizes. The test is as described in Section 3 except that at analysis $k$, the term $0.583 \sqrt{I}$ in the critical values for $Z_k$ is replaced by $0.583\sqrt{(V_k - V_{k-1})}$. Whilst preserving the simple analytic definition of the test, this adjustment appears to be highly effective in maintaining error probabilities close to the desired value $\alpha$.

5. Analysis following a group sequential test

The language used to describe sequential tests suggests that, as soon as a sequential study is terminated, decisions are made and actions taken. Whilst industrial acceptance sampling and control chart procedures generally conform to this pattern, the same cannot be said of most clinical trials. Only rarely are the results of a single Phase III study accepted as conclusive evidence in favor of a new
therapy, and thus a more complete analysis is required than the "accept" or "reject" decision of a hypothesis test. In this section we shall describe frequentist methods for calculating confidence intervals, significance levels and point estimates following a group sequential test.

The distinctive feature of the problem is the data—dependent sample size. In non—sequential studies, sample size is usually ancillary to the parameter of interest and inferences can be made by conditioning on the observed sample size; in a sequential study, the parameter of interest can have a dramatic effect on the sample size distribution, the observed sample size is then far from ancillary and it would be inappropriate to make probability statements conditionally on an observed sample size. In our previous notation, the sequence \( \{V_1, V_2, \ldots \} \) is ancillary to the parameter \( \theta \) as long as the value of each \( V_k \), \( k = 2, 3, \ldots \), is not influenced by the previous observations, \( Z_1, \ldots, Z_{k-1} \), as discussed in Section 4. Inference will therefore be made conditionally on the realization of the sequence \( \{V_1, V_2, \ldots \} \). If early termination occurs, later values of the sequence are unobserved but we shall see that it is still possible to calculate confidence intervals and significance levels in this case.

5.1. Confidence intervals for a binomial parameter.

Reference was made in Section 1 to the large literature on group sequential tests for a binary response. A typical multistage plan is specified by the maximum number of stages, \( K \), the numbers of observations to be taken at each stage, \( n_k \), \( k = 1, \ldots, K \), and critical numbers \( a_k < b_k \), \( k = 1, \ldots, K \). We define \( m_k = n_1 + \ldots + n_k \), the number of observations taken in the first \( k \) stages. Each observation takes the value 1 with probability \( p \) and 0 with probability \( 1-p \) and the test statistic after stage \( k \) is \( S_k \), the sum of the first \( m_k \)
observations. For \( k = 1, \ldots, K \), early termination takes place at stage \( k \) if \( S_k \leq a_k \) or \( S_k \geq b_k \), otherwise the study continues until stage \( K \). Note that, although \( a_K \) and \( b_K \) are needed to determine the outcome of the test when it terminates at stage \( K \), they do not affect the sample space and, hence, do not enter into the calculation of a confidence interval upon termination; thus, the same general method applies to both one-sided and two-sided tests. We denote the stage at which sampling stops by \( N \) and the final value of the test statistic by \( S_N \). The pair \((N, S_N)\) is sufficient for the binomial parameter \( p \) and its sample space, \( \Omega \), is the set of all pairs of values \((n, s)\) which can be attained under the given sampling plan.

A convenient way to construct a \((1-\alpha)\) level confidence interval for \( p \) is as follows. First, define an ordering of the sample space, \( \Omega \), in which higher values of \((N, S_N)\) are typical of higher values of \( p \). We shall write \((n', s') \succ (n, s)\) to denote that \((n', s')\) is higher than \((n, s)\) in this ordering. Now, for each value \( p_0 \in [0, 1] \) find the lowest value \((n_u, s_u)\) and the highest value \((n_\ell, s_\ell)\) such that

\[
P\{(N, S_N) \leq (n_u, s_u) \mid p = p_0\} \leq \alpha/2
\]

and

\[
P\{(N, S_N) \geq (n_\ell, s_\ell) \mid p = p_0\} \leq \alpha/2.
\]

The acceptance region \( \{(n, s) : (n_\ell, s_\ell) \preceq (n, s) \preceq (n_u, s_u)\} \) defines a two-sided test of \( H_0: p = p_0 \) with type I error at most \( \alpha \). By the standard argument, this family of tests can be inverted to give a conservative \((1-\alpha)\) level confidence interval for \( p \) consisting of the set of all values \( p_0 \) for which the hypothesis \( p = p_0 \) is accepted. If \( P\{(N, S_N) \preceq (n, s) \mid p = p_0\} \) is an increasing function of \( p_0 \) for each \((n, s) \in \Omega\) we say that the distributions on \( \Omega \) are stochastically ordered.
with respect to \( p \in [0,1] \). In this case, \((n^*, s^*)\) and \((n_u, s_u)\) are both increasing functions of \( p_0 \) and the confidence interval for \( p \) when the observed value of \((N, S_N)\) is \((n^*, s^*)\) is \((p_L, p_U)\) where

\[
P\{(N, S_N) \gtrless (n^*, s^*) \mid p = p_L\} = \alpha/2
\]

and

\[
P\{(N, S_N) \gtrless (n^*, s^*) \mid p = p_U\} = \alpha/2.
\]

For each \( k = 1, \ldots, K \), the range of possible values for \( S_k \) is finite and the left hand sides of (5.1) can be evaluated without difficulty. Solutions \( p_L \) and \( p_U \) can be found by, for example, a bisection search.

Armitage (1958) derived confidence intervals for a binomial parameter following a fully sequential test. He ordered \( \Omega \) in terms of increasing values of the maximum likelihood estimate of \( p \), \( S_N/m_N \), and proved the stochastic ordering property analytically. Jennison and Turnbull (1983) derived confidence intervals following a group sequential test using a different ordering of the sample space. Their ordering depends firstly on the boundary crossed, i.e., whether \( S_N \leq a_N \) or \( S_N \geq b_N \), secondly on the number of stages, \( N \), and thirdly on the value of \( S_N \). More formally, an enumeration of the elements of \( \Omega \) in ascending order is:

\[
(1,0), (1,1), \ldots, (1, a_1)
\]

\[
(2, a_1+1), (2, a_1+2), \ldots, (2, a_2)
\]

\[
\ldots
\]

\[
(K, a_{K-1}+1), (K, a_{K-1}+2), \ldots, (K, b_{K-1}+n_K)
\]

\[
(K-1, b_{K-1}), (K-1, b_{K-1}+1), \ldots, (K-1, b_{K-2}+n_{K-1})
\]

\[
\ldots
\]

\[
(1, b_1), (1, b_1+1), \ldots, (1, n_1).
\]

(5.2)
Bather (1988) proves that the distributions on $\Omega$ are stochastically ordered for this ordering and, thus, the endpoints of a $1-\alpha$ level confidence interval can be found by solving (5.1).

Atkinson and Brown (1985) use the ordering (5.2) to derive confidence intervals following two and three stage plans in which early stopping is only possible for low values of $S_k$. Chang and O'Brien (1986) remark that ordering (5.2) may be unreasonable in its treatment of outcomes which greatly overshoot a boundary, i.e., $S_N \ll a_N$ or $S_N \gg b_N$. They define confidence sets by inverting a family of likelihood ratio tests (occasionally these sets are not actually intervals) and note that the average length of their intervals is less than that of Jennison and Turnbull's intervals. However, the Chang and O'Brien intervals also have lower coverage probabilities which even fall below the nominal level $1-\alpha$ for some parameter values.

Other approaches to this problem are possible: Duffy and Santner (1987) adapt several methods for one-sample binomial confidence intervals to the multistage setting. Since the sample space, $\Omega$, does not possess the monotone likelihood ratio property with respect to $p$, there is no simple rule to provide a definitive choice between methods. However, if confidence intervals were calculated as if the data had been obtained from a non-sequential study, the desired confidence level would not be maintained and it is, therefore, imperative that a method which makes proper allowance for the sequential sampling rule be used.

5.2 Confidence intervals for a normal mean

The problem of calculating a confidence interval for a normal mean following a sequential test is broadly similar to the binomial problem but the
The continuous nature of the sample space makes computations more tedious. An analytic treatment is possible for certain fully sequential tests. Siegmund (1978, 1985) derives confidence intervals following repeated significance tests and truncated sequential probability ratio tests; Whitehead and Jones (1979) derive confidence intervals following the triangular test described previously in Section 3. For group sequential tests, direct numerical computation allows a more general treatment. Tsiatis, Rosner and Mehta (1984) present the basic methodology and note that it can also be applied to data-dependent boundaries. Applications to boundaries defined through a Lan and DeMets (1983) error spending function are studied in detail by Kim and DeMets (1987b).

We shall use the notation introduced in Section 3 to describe the general problem with test statistic \( Z_k \sim N(\theta \nu_k, \nu_k) \) at the \( k \)th analysis. Suppose that lower and upper boundary points are \( (a_k, b_k) \) for \( k = 1, \ldots, K \): for a two-sided test \( a_K < b_K \) whilst \( a_K = b_K \) for a one-sided test. We denote the stage of termination by \( N \) and the value of the test statistic at this point by \( Z_N \). The sample space is the set of all possible values for the pair \((N, Z_N)\). The ordering of the sample space analogous to ordering (5.2) for the binomial problem is:

\[
(1, -\infty), \ldots, (1, a_1) \\
(2, -\infty), \ldots, (2, a_2) \\
\ldots \\
(K-1, -\infty), \ldots, (K-1, a_{K-1}) \\
(K, -\infty), \ldots, (K, +\infty) \\
(K-1, b_{K-1}), \ldots, (K-1, +\infty) \\
\ldots \\
(1, b_1), \ldots, (1, +\infty). \tag{5.3}
\]
The $(1-\alpha)$ level confidence interval for $\theta$ when $(N,Z_N)$ takes the value $(n^*,z^*)$ is $(\theta_L,\theta_U)$ where

\[ P\{ (N,Z_N) \geq (n^*,z^*) \mid \theta = \theta_L \} = \alpha/2 \]

and

\[ P\{ (N,Z_N) \leq (n^*,z^*) \mid \theta = \theta_U \} = \alpha/2. \quad (5.4) \]

The left hand sides of (5.4) may be computed using numerical integration and solutions $\theta_L$ and $\theta_U$ found by, for example, a bisection search.

Tsiatis, Rosner and Mehta note that this procedure depends only on the values of boundary points prior to stopping. Thus, if the sequence $\{V_1, V_2, \ldots\}$ is not known in advance and a Slud and Wei (1982) or Lan and DeMets (1983) method is used to calculate boundary points $(a_k,b_k)$ conditional on $\{V_1, \ldots, V_k\}$ for $k = 1, 2, \ldots$, a confidence interval for $\theta$ can still be calculated when the study terminates early and the subsequent information levels are unobserved. As in the binomial problem, other orderings may be used in place of (5.3). For example, outcomes could be ordered in terms of increasing values of the maximum likelihood estimate of $\theta$, $Z_N/V_N$. A drawback of such an ordering is the dependence of the confidence interval on boundary points following termination, which precludes its use when information levels are unpredictable.

5.3. Significance levels

Consider a sequential test of the null hypothesis $H_0: \theta = \theta_0$ with type I error probability $\alpha$. Whenever this test rejects $H_0$ one can say that $H_0$ is rejected at significance level $\alpha$. However, it is preferable to report a smaller significance level when the evidence against $H_0$ is particularly strong, e.g., when
$H_0$ is rejected at an early stage. The derivation of significance levels following a sequential test parallels that of the confidence intervals described in Sections 5.2 and 5.3. An ordering of the sample space is specified and the significance level for $H_0: \theta = \theta_0$ is defined to be the probability, under $\theta = \theta_0$, of an observation as extreme or more extreme than that actually observed, with the usual interpretation of "extreme" for one-sided and two-sided significance levels.

Analytic formulae for significance levels following certain fully sequential tests have been obtained by Siegmund (1978, 1985) and by Whitehead and Jones (1979). The group sequential problem has been studied numerically by Fairbanks and Madsen (1982) for normal response and by Madsen and Fairbanks (1983) for exponentially distributed response. To illustrate the case of normal response, consider the group sequential test described in Section 5.2 and the sample space ordering (5.3). When $(N,Z_N)$ takes the value $(n^*, z^*)$, the one-sided significance level for $H_0: \theta = \theta_0$ with alternative $\theta > \theta_0$ is

$$P\{(N,Z_N) \geq (n^*, z^*) | \theta = \theta_0\},$$

the one-sided significance level of $H_0: \theta = \theta_0$ with alternative $\theta < \theta_0$ is

$$P\{(N,Z_N) \leq (n^*, z^*) | \theta = \theta_0\}$$

and the two-sided significance level is twice the smaller of these quantities. Note that significance levels can be found following either a one-sided or two-sided test
and for any value of \( \theta_0 \). As in the non-sequential case, the set of values of \( \theta_0 \) with two-sided significance levels greater than \( \alpha \) forms a \( 1-\alpha \) level confidence interval for \( \theta \).

5.4. Point estimates

It is well-known that maximum likelihood estimates calculated after a sequential test are biased. A two-sided test of \( H_0: \theta = 0 \) terminates early if the sample mean is large in absolute value, leading to a positive bias in the maximum likelihood estimate of \( \theta \) when \( \theta > 0 \) and a negative bias when \( \theta < 0 \). To our knowledge, there has been no work directed specifically at bias correction in group sequential tests but, in this section, we shall outline methods for fully sequential tests which can be adapted to the group sequential setting.

Whitehead (1983, Sec. 5.3) observes that, for a continuous response variable, a lower (or upper) 50\% confidence limit for a parameter \( \theta \) is a median unbiased estimate of \( \theta \), i.e., the median of its distribution is equal to the true parameter value. This construction can also be used after a group sequential test. In the example of Section 5.2 a median unbiased estimate for the normal mean, \( \theta \), is obtained by solving for \( \theta_L \) the first of equations (5.4) with \( \alpha = 1 \).

The problem of finding an estimator whose mean is approximately equal to the true parameter value is more difficult. Let \( \hat{\theta} \) denote the maximum likelihood estimate of \( \theta \) and suppose \( E(\hat{\theta}) = \theta + b(\theta) \). Whitehead (1986) evaluates the bias function \( b(\theta) \) analytically for the sequential probability ratio test and the previously described triangular test. In order to reduce the bias in \( \hat{\theta} \) by subtracting an estimate of its bias, he proposes the estimator \( \tilde{\theta} \), obtained by
solving

$$\hat{\theta} = \hat{\theta} - b(\hat{\theta}).$$

Analytic expressions for $b(\theta)$ are not generally available. However, for group sequential tests, direct computation of $b(\theta)$ by numerical integration is possible and this method of bias reduction could be applied; whether it performs adequately in general is a topic for further investigation.

6. Repeated confidence intervals

So far, we have discussed statistical stopping rules and described testing and estimation procedures associated with them. However, as mentioned in Section 1, in certain applications, strict adherence to a rigid stopping rule may be neither possible nor desirable. For example, in a medical trial, factors such as side-effects, financial cost, quality of patients' lives and the availability of new promising treatments unknown at the start of the trial will influence the decision to terminate a trial, as well as statistical issues on the primary endpoint. If the adherence to the stopping rule is not guaranteed, then formally none of the results of the previous sections are directly applicable. Rather, what is required is a more flexible approach in which inferential statements concerning the primary endpoint can be made at points in time during the trial, these statements being valid whatever stopping rule is to be employed.

Suppose a total of $K$ analyses are planned during a study. A $1-\alpha$ level sequence of repeated confidence intervals (RCI's) for the parameter $\theta$ is a sequence $\{I_k; k = 1, \ldots, K\}$ where $I_k$ is calculated from the information available at analysis $k$ and
\[ P_{\theta} \{ \theta \in I_k \; : \; k = 1, \ldots, K \} = 1 - \alpha. \]

The interval \( I_k \) provides a statistical summary of the information about the parameter \( \theta \) at the \( k \)'th analysis, automatically adjusted to compensate for repeated looks at the accumulating data. As such, it can be presented to a monitoring committee to be considered with all other relevant information when discussing early termination of the study. Even though RCI's for a major endpoint may influence the decision to terminate the study, it is important to realize that the overall coverage probability of the sequence of RCI's is guaranteed independently of any stopping rule, other than the imposition of a maximum number of analyses, \( K \).

The historical antecedent of repeated confidence intervals is the confidence sequence introduced by Robbins (1970), an infinite sequence of intervals, \( I_n \), \( n = 1, 2, \ldots \), where \( I_n \) is calculated from the first \( n \) observations and, for a 1\( - \alpha \) level sequence,

\[ P_{\theta} \{ \theta \in I_n \; \text{for all} \; n = 1, 2, \ldots \} = 1 - \alpha. \]

The theory of confidence sequences has had little practical impact, one reason being that the intervals turn out to be extremely wide and unacceptable to investigators accustomed to fixed sample intervals. However, sequences of repeated confidence intervals are finite in number and tend to be much narrower. The development of repeated confidence intervals is due to Jennison and Turnbull (1984, 1985 and 1989); similar ideas have also been discussed by Lai (1984). It should be stressed that we are concerned with a quite different problem from that of constructing a confidence interval upon termination of a test, described in Section 5.
Repeated confidence intervals are formed by inverting a family of group sequential tests. We shall confine our attention to the normal case, retaining the notation introduced in Section 3, where at the \( k \)’th interim analysis, \( V_k \) denotes the current measure of observed information and, conditional on \( V_k \),

\[
Z_k \sim \left( \theta V_k, V_k \right) \quad \text{and increments } \quad Z_k - Z_{k-1}, \quad k = 2,3,\ldots, \quad \text{are independent.}
\]

Construction of a sequence of RCI’s starts with a two-sided group sequential test of the hypothesis \( H_0: \theta = \theta_0 \) with \( K \) analyses and type I error probability \( \alpha \).

Such a test rejects \( H_0 \) at analysis \( k \) if

\[
|Z_k - \theta_0 V_k| \geq c_k(V_k), \quad \text{for } k = 1,\ldots,K,
\]

where the critical value \( c_k(V_k) \) depends on the form of test being used. For the Pocock test, described in Section 2,

\[
c_k(V_k) = C_p(K,\alpha) \sqrt{V_k}
\]

whilst, for the O’Brien and Fleming test,

\[
c_k(V_k) = C_B(K,\alpha)\sqrt{\frac{V_k}{k}},
\]

which is constant with respect to \( k \) if the \( V_k \) are equally spaced; alternatively, any other test from Section 2 could be used here or, if increments in information between analyses are unpredictable, one of the methods described in Section 4 may be applied. Varying the value of \( \theta_0 \) produces a family of group sequential tests and the \( k \)’th repeated confidence interval, \( I_k \), is defined to be the set of values of \( \theta_0 \) currently accepted by their specific tests. Formally,

\[
I_k = \{ \theta: |Z_k - \theta V_k| < c_k(V_k) \}
\]

\[
= \left( \left\{ Z_k - c_k(V_k) \right\}/V_k, \quad \left\{ Z_k + c_k(V_k) \right\}/V_k \right)
\]

\[
= (\theta_k, \bar{\theta}_k), \quad \text{say.}
\]

It follows immediately from that fact that the underlying test has size \( \alpha \) that the sequence \( \{I_k; \quad k = 1,\ldots,K\} \) has the desired RCI property, namely
\[ P_\theta \left\{ \theta \in I_k; \ k = 1, \ldots, K \right\} = 1 - \alpha. \]

Note that if we ignored the fact that we were performing multiple analyses, the fixed sample confidence interval for \( \theta \) would be \( \left( \frac{Z_k}{\sqrt{V_k}} \pm \Phi^{-1}(1-(\alpha/2))/\sqrt{V_k} \right. \). A measure of the cost of "snooping" at the data by performing interim analyses can be constructed by examining the width of the final interval, \( I_K \), relative to that of the fixed sample interval, assuming no interim analyses were to be performed. This ratio is given by \( \sqrt{V_K} \Phi^{-1}(V_K)/(\Phi^{-1}(1-(\alpha/2)). \) We can see that performing interim analyses and making the proper adjustment in the width of the final stated interval does not involve a very great cost. For example, using Table 2, for 90\% intervals (\( \alpha = 0.1 \)) and \( K = 5 \) looks, the ratio of adjusted to unadjusted width of the final interval is 2.12/1.645 = 1.29 for the Pocock boundary, and only 1.75/1.645 = 1.07 for the O'Brien and Fleming boundary. For \( K = 10 \) looks, the ratios are 1.38 and 1.10 respectively. Using such information, group sizes can be chosen such that \( I_K \) is of some prespecified width. Note that although the O'Brien and Fleming boundary based intervals are narrower at the final planned analysis, the Pocock boundary based intervals are narrower at the early looks. Of course if interim analyses are to be carried out, the fixed sample intervals are invalid in any case and some adjustment must be made.

RCI's can be used simply as data summaries at interim analyses. However, their major role will normally be as a source of information in the process of deciding when to terminate a trial. Suppose that \( \theta \) measures the difference with respect to a major endpoint between a new treatment and a standard and a positive value of \( \theta \) denotes that the new treatment is superior. If the initial objective of a study is to test \( H_0: \ \theta = 0 \) against the two-sided alternative \( H_A: \ \theta \neq 0 \), a stopping rule based on the sequence of RCI's for \( \theta \) is to terminate
with rejection of $H_0$ at stage $k$ if $I_k$ fails to contain $\theta = 0$, for $k = 1, \ldots, K$, and to accept $H_0$ if the study continues to stage $K$ without rejection. In this case it is easy to see that the original size $\alpha$ group sequential test upon which the RCIs were based has been recovered exactly. The RCIs can be considered as adjuncts to this test indicating which other values of $\theta$ are plausible given the data, both at intermediate and final analyses. Suppose, however, that $\theta_k > 0$ for some $k < K$ but the new treatment has exhibited unexpected adverse side-effects and, although the above stopping rule calls for termination, the monitoring committee decides to continue the study. The advantage of the repeated confidence interval approach is that, since the defining property of the sequence of RCIs does not depend on adherence to the formal stopping rule, RCIs may still be calculated at subsequent analyses and used in the decision making process. For example, the study might be terminated in favor of the new treatment at a later analysis if the lower end point of the RCI is sufficiently greater than zero that the benefits with respect to the major endpoint are considered to outweigh the adverse side-effects.

RCIs can be used in a similar way for one-sided testing problems. Suppose that in the above example the original objective was to test $H_0: \theta \leq 0$ against $H_A: \theta > 0$ with error probabilities at most $\alpha$ at $\theta = 0$ and $\theta = \delta$. A stopping rule for this problem can be defined in terms of a sequence of $1-2\alpha$ level RCIs. The study is terminated at stage $k$ to accept $H_0$ if $\theta_k < \delta$ or to accept $H_A$ if $\theta_k > 0$; in order to ensure termination at the $K$'th analysis, the group sizes should be chosen so that the final interval width, $2c_K/V_K$, is no greater than $\delta$. It follows from the fact that $P\{\theta_k < \theta \text{ for all } k = 1, \ldots, K\}$ and $P\{\theta_k > \theta \text{ for all } k = 1, \ldots, K\}$ are both almost exactly equal to $\alpha$ that the error probabilities at $\theta = 0$ and $\theta = \delta$ are no greater than $\alpha$. In fact, a small amount of conservatism
occurs: if the true value of $\theta$ is 0 and $\hat{\theta}_k > 0$ for some $k$, a type I error will not be made if $\hat{\theta}_{k'} < \delta$ for some $k' < k$ and thus the test has already terminated to accept $H_0$. Again, the intention is that RCI's should be used to provide guidelines for termination rather than a strict stopping rule. If the study continues past an interim analysis at which $\hat{\theta}_k > 0$ or $\hat{\theta}_k < \delta$, later RCI's are still valid and can be used in reaching subsequent decisions. It is, however, of interest to study the properties of one-sided tests derived from RCI's when the stopping rule is strictly applied, since a comparison with other group sequential tests allows assessment of the statistical efficiency of this approach. Jennison and Turnbull (1989) show that tests derived from either Pocock or O'Brien and Fleming type RCI's are highly efficient. Their conservatism is slight, with typical error probabilities around 0.045 rather than 0.05, and expected sample sizes are within a few percent of the minimum possible values for tests with the same group sizes and the same error probabilities.

As an illustration, we present the Pocock based and O'Brien and Fleming based derived tests for the example presented earlier at the end of Section 3. For the general two sample problem of DeMets and Ware (1980), up to $K$ groups of $2n$ patients are available and $V_k = kn/(2\sigma^2)$ for $k = 1, ..., K$. In order to ensure that the $K$'th interval has width $\delta$ a common group size $n = 8\sigma^2 C_p^2 (K,2\alpha)/(K \delta^2)$ is required if Pocock type RCI's are used and $n = 8\sigma^2 C_B^2 (K,2\alpha)/(K^2 \delta^2)$ is required for the O'Brien and Fleming based RCI's. Values of $C_p$ and $C_B$ can be obtained from Table 2 but note that the type I error for the Pocock or O'Brien and Fleming test is $2\alpha$, thus 90% RCI's are needed to give a derived test with error probabilities of 0.05. In terms of the $Z_k$, the upper boundary of the derived test is $Z_k = c_k$ and the lower boundary is $Z_k = \delta V_k - c_k$. Unlike the two-sided case, the derived test is now quite different.
from the parent test upon which the RCI's are based. In the numerical example of Section 3, we had \( \sigma^2 = 1, \ \delta = 0.4 \) and \( K = 5 \). For the Pocock derived test, substitution of \( C_P(5, 0.1) = 2.122 \) into the above formulae gives \( n = 45.03 \)
which we round to 45; thus, \( V_k = 22.5 \ k, \ c_k = 2.122 \sqrt{V_k} = 10.06 \sqrt{k} \) and the boundary points of the derived test are \( 4.5 \pm 5.56, 9.0 \pm 5.23, 13.5 \pm 3.93, 18.0 \pm 2.13 \) and \( 22.5 \pm 0 \). This test has a maximum sample size of 450 patients and error probabilities 0.044 at \( \theta = 0 \) and 0.4. The expected sample size is 233 when \( \theta = 0.2 \) and 171 when \( \theta = 0 \) or 0.4; the value of \( F_4 \), the average of the expected sample sizes for \( \theta = 0.2, 0.3, 0.4, 0.5 \) and 0.6 is 173. For the O'Brien and Fleming derived test, \( C_B(5, 0.1) = 3.915 \) and \( n = 30.65 \) which we round to 31; thus \( V_k = 15.5 \ k, \ c_k = 15.41 \) for all \( k \) and the boundary points of the derived test are \( 3.1 \pm 12.31, 6.2 \pm 9.21, 9.3 \pm 6.11, 12.4 \pm 3.01 \) and \( 15.5 \pm 0 \). This test has a maximum sample size of 310 patients and error probabilities 0.045 at \( \theta = 0 \) and 0.4. The expected sample size is 235 when \( \theta = 0.2 \) and 191 when \( \theta = 0 \) or 0.4; the value of \( F_4 \) is 190.

In addition to the one-sided and two-sided hypothesis testing problems, a particularly useful application of tests derived from RCI's is in bioequivalence testing, where it is desired to control the Type I error probability for rejecting the null hypothesis \( H_0: \theta \neq 0 \) in favor of the alternative \( H_A: \theta = 0 \). A derived test of this hypothesis can be constructed by rejecting \( H_0 \) only if at some stage the RCI is wholly contained in some specified "region of bioequivalence" \( (-\delta^*, \delta^*) \). See Jennison and Turnbull (1989) for more details.

Throughout this section and this paper, we have concentrated on the prototype problem of normal responses with known variance but, as for other group sequential methods, the normal theory has wide applicability. Jennison and Turnbull (1984) derive RCI's for the hazard ratio in a two-sample problem with
censored survival data. In Jennison and Turnbull (1985), the methods are applied to the problem of monitoring the median of a survival distribution, again with censored lifetime data. Jennison and Turnbull (1989) treat several problems concerning discrete data, namely, success probabilities for binomial data, odds ratios in stratified designs and matched pairs designs, and coefficients in an odds ratio regression model. They also derive RCI's for a normal mean when the variance is unknown by inverting an exact group sequential t-test and obtain RCI's for a multivariate normal mean by inverting a group sequential $\chi^2$ test.

It should be noted that a theory of "repeated P-values" can be developed analogously to that of repeated confidence intervals. At the k'th analysis, a two-sided repeated P-value for the null hypothesis $H_0: \theta = \theta_0$ is defined as $P_k = \max \{ \alpha: \theta_0 \in I_k \}$, where $I_k$ is the current $(1-\alpha)$-level RCI. In other words $P_k$ is that value of $\alpha$ for which the k'th $(1-\alpha)$-level RCI, $I_k$, contains the null value, $\theta_0$, as one of its endpoints. The construction ensures that the overall probability under $H_0$ of ever seeing a repeated P-value less than or equal to p is no more than p for any $0 \leq p \leq 1$, and equals p if all P-values are to be observed. Thus the repeated P-value can be quoted with the usual interpretation, yet with protection against the multiple looks effect. The concept of repeated P-values should not be confused with the significance levels constructed in Section 5.3. Repeated P-values, like RCI's, have the advantage of flexibility — their frequentist interpretation is valid at interim analyses and not dependent on any stopping rule.

This research was supported in part by Grant GM 28364 from the U.S. National Institutes of Health.
References


Table 1. Distribution of stopping time and expected sample size for a repeated significance test of $H_0: \theta = 0$ with four groups and power 0.9 at $\theta = \pm 0.5$

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>P{stop after group $j$}</th>
<th>E{Number of patients}</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0.0$</td>
<td>0.018 0.013 0.010 0.958</td>
<td>195.4</td>
</tr>
<tr>
<td>$\pm 0.25$</td>
<td>0.070 0.090 0.094 0.746</td>
<td>175.8</td>
</tr>
<tr>
<td>$\pm 0.5$</td>
<td>0.277 0.307 0.205 0.211</td>
<td>117.6</td>
</tr>
<tr>
<td>$\pm 0.75$</td>
<td>0.614 0.311 0.064 0.011</td>
<td>73.6</td>
</tr>
<tr>
<td>$\pm 1.0$</td>
<td>0.880 0.116 0.004 0.000</td>
<td>56.2</td>
</tr>
</tbody>
</table>
Table 2. *Constants $C_p(K, \alpha)$ and $C_B(K, \alpha)$ for two-sided group sequential tests with type I error $\alpha$ due to Pocock (1977) and O'Brien and Fleming (1979) respectively.

<table>
<thead>
<tr>
<th>K</th>
<th>$\alpha = 0.05$</th>
<th>$\alpha = 0.10$</th>
<th>$\alpha = 0.05$</th>
<th>$\alpha = 0.10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2.178</td>
<td>1.875</td>
<td>2.797</td>
<td>2.373</td>
</tr>
<tr>
<td>3</td>
<td>2.289</td>
<td>1.992</td>
<td>3.471</td>
<td>2.962</td>
</tr>
<tr>
<td>4</td>
<td>2.361</td>
<td>2.067</td>
<td>4.048</td>
<td>3.466</td>
</tr>
<tr>
<td>5</td>
<td>2.413</td>
<td>2.122</td>
<td>4.562</td>
<td>3.915</td>
</tr>
<tr>
<td>10</td>
<td>2.555</td>
<td>2.270</td>
<td>6.597</td>
<td>5.695</td>
</tr>
</tbody>
</table>

*These figures are exact and obtained by numerical integration. The constants $C_B$ are slightly different from those in Table 1 of Jennison and Turnbull (1984), which were obtained by simulation.
Table 3. A comparison of group sequential tests with four groups. In this example 
\( \sigma^2 = 1 \) and tests have type I error 0.05 at \( H_0: \theta = 0 \) and power 0.9* at 
\( \theta = 0.5 \). The standardized critical values \( c_1, \ldots, c_4 \) determine the tests: \( H_0 \) is 
rejected after group \( k \) if \( |\bar{d}_k| \geq c_k\sqrt{2\sigma^2/(kn)} \). \( N_{\text{max}} \) denotes the 
maximum number of subjects and \( E_{\theta}(N) \) the expected number of subjects 
when \( \mu_A - \mu_B = \theta \). A fixed sample test requires 168 subjects.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \mu = 0.2 )</td>
<td>( \mu = 0.5 )</td>
<td>( \Delta = 0.2 )</td>
<td>( \Delta = 0.4 )</td>
<td></td>
</tr>
<tr>
<td>( c_1 )</td>
<td>2.36</td>
<td>4.05</td>
<td>3.0</td>
<td>2.94</td>
<td>2.64</td>
</tr>
<tr>
<td>( c_2 )</td>
<td>2.36</td>
<td>2.86</td>
<td>3.0</td>
<td>2.88</td>
<td>2.56</td>
</tr>
<tr>
<td>( c_3 )</td>
<td>2.36</td>
<td>2.34</td>
<td>3.0</td>
<td>2.81</td>
<td>2.48</td>
</tr>
<tr>
<td>( c_4 )</td>
<td>2.36</td>
<td>2.02</td>
<td>1.98</td>
<td>1.99</td>
<td>2.09</td>
</tr>
<tr>
<td>( N_{\text{max}} )</td>
<td>200</td>
<td>176</td>
<td>176</td>
<td>176</td>
<td>184</td>
</tr>
<tr>
<td>( \theta = 0 )</td>
<td>195.4</td>
<td>174.9</td>
<td>175.4</td>
<td>175.1</td>
<td>181.7</td>
</tr>
<tr>
<td>( \theta = \pm 0.25 )</td>
<td>175.8</td>
<td>165.7</td>
<td>170.1</td>
<td>168.3</td>
<td>169.0</td>
</tr>
<tr>
<td>( E_{\theta}(N) ) ( \theta = \pm 0.5 )</td>
<td>117.6</td>
<td>131.2</td>
<td>138.8</td>
<td>133.4</td>
<td>122.7</td>
</tr>
<tr>
<td>( \theta = \pm 0.75 )</td>
<td>73.6</td>
<td>97.7</td>
<td>91.1</td>
<td>87.3</td>
<td>78.3</td>
</tr>
<tr>
<td>( \theta = \pm 1.0 )</td>
<td>56.2</td>
<td>79.3</td>
<td>62.5</td>
<td>61.0</td>
<td>56.9</td>
</tr>
</tbody>
</table>

*There are some small differences in power due to rounding of the sample size.
Table 4. Minimum expected sample size at $\theta = \pm \delta$ as a fraction of the fixed sample size needed for a test of $H_0: \theta \leq 0$ against $H_A: \theta > 0$ with type I error 0.05 at $\theta = -\delta$ and power 0.95 at $\theta = \delta$. Tabulated minima are amongst group sequential tests with $K$ groups and maximum sample size $V_{\text{max}}/V_f$ times that of the fixed sample test.

<table>
<thead>
<tr>
<th>K</th>
<th>$V_{\text{max}}/V_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.05</td>
</tr>
<tr>
<td>2</td>
<td>.745</td>
</tr>
<tr>
<td>3</td>
<td>.693</td>
</tr>
<tr>
<td>5</td>
<td>.652</td>
</tr>
<tr>
<td>10</td>
<td>.621</td>
</tr>
</tbody>
</table>
Table 5. Boundaries of group sequential one-sided tests derived by Jennison (1987). Tests are of $H_0: \theta \leq 0$ against $H_A: \theta > 0$ with error probabilities $\alpha = 0.05$ at $\theta = \pm 1.64485$ and are very nearly optimal with respect to minimizing objective function $F_4$. The test with $K$ analyses has $V_k = (k/K)^{1.1}$, and stops at analysis $k$ to reject $H_0$ if $Z_k > c_k$ and to accept $H_0$ if $Z_k \leq -c_k$.

<table>
<thead>
<tr>
<th>$c_1$</th>
<th>0.608</th>
<th>0.786</th>
<th>0.842</th>
<th>0.861</th>
<th>0.866</th>
<th>0.866</th>
<th>0.863</th>
<th>0.860</th>
<th>0.856</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_2$</td>
<td>0.0</td>
<td>0.613</td>
<td>0.783</td>
<td>0.856</td>
<td>0.889</td>
<td>0.903</td>
<td>0.908</td>
<td>0.909</td>
<td>0.907</td>
</tr>
<tr>
<td>$c_3$</td>
<td>0.0</td>
<td>0.0</td>
<td>0.591</td>
<td>0.765</td>
<td>0.854</td>
<td>0.903</td>
<td>0.929</td>
<td>0.944</td>
<td>0.950</td>
</tr>
<tr>
<td>$c_4$</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.562</td>
<td>0.739</td>
<td>0.837</td>
<td>0.895</td>
<td>0.932</td>
<td>0.955</td>
</tr>
<tr>
<td>$c_5$</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.532</td>
<td>0.713</td>
<td>0.815</td>
<td>0.880</td>
<td>0.923</td>
<td></td>
</tr>
<tr>
<td>$c_6$</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.502</td>
<td>0.688</td>
<td>0.792</td>
<td>0.862</td>
<td></td>
</tr>
<tr>
<td>$c_7$</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.474</td>
<td>0.664</td>
<td>0.771</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$c_8$</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.446</td>
<td>0.641</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$c_9$</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.420</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$c_{10}$</td>
<td>0.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 6. Type I error probabilities for tests with 5 groups using nominal significance levels from size 0.05 Pocock and O'Brien & Fleming tests with no adjustment for unequal group sizes.

<table>
<thead>
<tr>
<th>$V_1, \ldots, V_5$</th>
<th>Pocock</th>
<th>O'Brien &amp; Fleming</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0, 2.0, 3.0, 4.0, 5.0</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>1.2, 2.1, 2.8, 4.1, 5.3</td>
<td>0.049</td>
<td>0.051</td>
</tr>
<tr>
<td>1.1, 2.0, 2.8, 3.4, 4.6</td>
<td>0.048</td>
<td>0.051</td>
</tr>
<tr>
<td>2.1, 2.8, 3.4, 4.4, 4.9</td>
<td>0.041</td>
<td>0.046</td>
</tr>
<tr>
<td>0.3, 0.9, 2.0, 3.2, 5.4</td>
<td>0.060</td>
<td>0.057</td>
</tr>
</tbody>
</table>
Table 7. Size and power achieved by Slud & Wei and Lan & DeMets two-sided tests of $H_0: \theta = 0$. Test statistics $Z_k \sim N(\theta V_k, V_k)$ for $k = 1, \ldots, 5$, where $\{V_1, \ldots, V_5\}$ denote the observed information. The Slud & Wei test allocates type I error 0.01 to each analysis whereas the Lan & DeMets test is defined through the error spending function $f(t) = \alpha \min(t, 1)$ and target information level $V_{max} = 50$.

<table>
<thead>
<tr>
<th>$V_1, \ldots, V_5$</th>
<th>Slud &amp; Wei</th>
<th>Lan &amp; DeMets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>size</td>
<td>power at $\theta = 0.489$</td>
</tr>
<tr>
<td>1. 10, 20, 30, 40, 50</td>
<td>0.05</td>
<td>0.9</td>
</tr>
<tr>
<td>2. 4, 13, 28, 41, 50</td>
<td>0.05</td>
<td>0.895</td>
</tr>
<tr>
<td>3. 15, 18, 26, 38, 50</td>
<td>0.05</td>
<td>0.896</td>
</tr>
<tr>
<td>4. 8, 15, 25, 33, 42</td>
<td>0.05</td>
<td>0.839</td>
</tr>
<tr>
<td>5. 12, 25, 34, 47, 61</td>
<td>0.05</td>
<td>0.947</td>
</tr>
</tbody>
</table>
Values of $b_4$ for the error spending function used to define the group sequential tests of Table 5. All tests have $t_{\text{max}} = 1$, $b_1 = 0.3$, $b_2 = 0.08$ and $b_3 = 1.8$.

<table>
<thead>
<tr>
<th>$K$</th>
<th>$b_4$</th>
<th>$K$</th>
<th>$b_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-0.93918</td>
<td>7</td>
<td>-1.50000</td>
</tr>
<tr>
<td>3</td>
<td>-1.23132</td>
<td>8</td>
<td>-1.52499</td>
</tr>
<tr>
<td>4</td>
<td>-1.35147</td>
<td>9</td>
<td>-1.54621</td>
</tr>
<tr>
<td>5</td>
<td>-1.42063</td>
<td>10</td>
<td>-1.56391</td>
</tr>
<tr>
<td>6</td>
<td>-1.46601</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>