SCHOOL OF OPERATIONS RESEARCH
AND INDUSTRIAL ENGINEERING
COLLEGE OF ENGINEERING
CORNELL UNIVERSITY
ITHACA, NEW YORK 14853-3801

TECHNICAL REPORT NO. 1267

August 2000

**ON GROUP SEQUENTIAL TESTS
FOR DATA IN UNEQUALLY SIZED
GROUPS AND WITH UNKNOWN VARIANCE**

by

Christopher Jennison and Bruce W. Turnbull

On Group Sequential Tests for Data in Unequally Sized Groups and with Unknown Variance

Christopher Jennison

Department of Mathematical Sciences,
University of Bath, Bath BA2 7AY, U. K.

and

Bruce W. Turnbull

School of Operations Research and Industrial Engineering
Cornell University, Ithaca, New York 14853-3801, U. S. A.

Abstract

We present a method to generalise the scope of application of group sequential tests designed for equally sized groups of normal observations with known variance. Preserving the significance levels against which standardised statistics are compared leads to tests for unequally grouped data which maintain Type I error probabilities to a high degree of accuracy. The same approach can be followed when observations have unknown variance by setting critical values for Studentised statistics at percentiles of the appropriate t -distributions.

This significance level approach is equally applicable to group sequential one-sided tests and two-sided tests, possibly with early stopping permitted to accept the null hypothesis. In applications to equivalence testing, tests are required to maintain a specified power, rather than Type I error rate: such tests can be constructed by defining the standardised test statistics used in the significance level approach with respect to appropriately chosen hypotheses.

1 Introduction

Many authors have proposed group sequential tests for studies producing normally distributed observations with known variance, arriving in equally sized groups. However, these assumptions are too restrictive in many situations, particularly for the important application to clinical trials. In this paper we consider generalisations of these tests in other directions, first for normal data of known variance in unequal and unpredictably sized

groups, then when data are normal but of unknown variance. In this context, the most interesting challenges arise in one-sided tests, two-sided tests allowing early termination either to reject or accept the null hypothesis, and equivalence tests where the error rates at non-null parameter values are to be afforded the greatest protection.

A general setting for the methodology we shall describe is the group sequential analysis of a normal linear model when inferences concern a scalar parameter θ . Here, θ may be a linear combination of elements of the parameter vector or perhaps just a single component representing the treatment effect under investigation. Suppose a study has a maximum of K analyses and, for $k = 1, \dots, K$, we denote the maximum likelihood estimate of θ at analysis k by $\hat{\theta}^{(k)}$. Jennison & Turnbull (1997a) prove the general result that, for a normal linear model, the sequence $\{\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(K)}\}$ is multivariate normal with

$$\text{Cov}(\hat{\theta}^{(k_1)}, \hat{\theta}^{(k_2)}) = \text{Var}(\hat{\theta}^{(k_2)}) \quad \text{for } k_1 \leq k_2.$$

Let $\mathcal{I}_k = \{\text{Var}(\hat{\theta}^{(k)})\}^{-1}$, the information for θ at analysis k , and define the standardised statistics

$$Z_k = \hat{\theta}^{(k)} \{\text{Var}(\hat{\theta}^{(k)})\}^{-1/2} = \hat{\theta}^{(k)} \sqrt{\mathcal{I}_k}, \quad k = 1, \dots, K.$$

Then, it follows from the joint distribution of $\{\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(K)}\}$ stated above that

- (i) $(Z_1, \dots, Z_K)$ is multivariate normal,
 - (ii) $E(Z_k) = \theta \sqrt{\mathcal{I}_k}$, $k = 1, \dots, K$, and
 - (iii) $\text{Cov}(Z_{k_1}, Z_{k_2}) = \sqrt{(\mathcal{I}_{k_1}/\mathcal{I}_{k_2})}$, $1 \leq k_1 \leq k_2 \leq K$.
- (1)

We shall refer to (1) as the *canonical distribution* of the sequence of test statistics $\{Z_1, \dots, Z_K\}$ with information levels $\{\mathcal{I}_1, \dots, \mathcal{I}_K\}$ for θ . Jennison & Turnbull (1997a) and Scharfstein, Tsiatis & Robins (1997) prove that this joint distribution also holds approximately in a wide variety of other situations.

We shall define tests in terms of the standardised statistics $Z_1, \dots, Z_K$ but note that tests are sometimes presented in terms of the score statistics $S_k = Z_k \sqrt{\mathcal{I}_k}$, $k = 1, \dots, K$. The sequence of score statistics has the same joint distribution as a Brownian motion with drift θ observed at times $\mathcal{I}_1, \dots, \mathcal{I}_K$; see Lan & Zucker (1993) for a discussion of the inter-relations between the three processes, $\{\hat{\theta}^{(k)}\}$, $\{Z_k\}$ and $\{S_k\}$. If observations are independent, identically distributed, univariate $N(\theta, \sigma^2)$, and n_k denotes the cumulative sample size at analysis k , then S_k is the sum of the first n_k observations, $\hat{\theta}^{(k)} = S_k/n_k$, $\mathcal{I}_k = n_k/\sigma^2$, and $Z_k = \hat{\theta}^{(k)} \sqrt{\mathcal{I}_k} = S_k/\sqrt{(n_k \sigma^2)}$. In other situations, $\mathcal{I}_k$ will increase with n_k in a manner depending on the experimental design employed. Although we refer to the problem of “unequal group sizes”, we shall really address the more general problem of unequal increments in the sequence of information levels $\mathcal{I}_1, \dots, \mathcal{I}_K$.

Two approaches to handling unequal group sizes have been proposed in the context of two-sided tests with no early stopping to accept H_0 . One is the “significance level” method proposed by Pocock (1977) and the other is Lan & DeMets’ (1983) “error spending” approach. Pocock’s method is simple to apply and maintains the Type I error rate quite accurately as long as the maximum number of analyses, K , is fixed and variations in group size are not too extreme. The error spending approach achieves a specified Type I error probability exactly and allows the maximum number of analyses to be varied, depending on the rate at which information accrues. Thus, with suitable computer implementation, such as the programs available at <http://www.medsch.wisc.edu/landemets/>, Lan & DeMets’ method would appear to be the appropriate choice for two-sided tests of normal data with known variance. However, we have found Pocock’s significance level approach to be a better starting point for handling unequal and unpredictable group sizes in two-sided tests which also include an “inner wedge” to allow early stopping to accept H_0 . Pocock also applied the significance level approach to create group sequential two-sided t -tests and we have found this approach to extend neatly and effectively to group sequential one-sided t -tests and two-sided, inner wedge t -tests with early stopping to accept or reject H_0 .

We discuss two-sided group sequential tests with no early stopping to accept H_0 in Section 2 and illustrate the significance level method for handling unequal group sizes and unknown variance. In Section 3, we present the family of one-sided group sequential tests described by Emerson & Fleming (1989) and Pampallona & Tsiatis (1994) and show how the significance level method is able to extend the scope of application of these tests to unequal group sizes and unknown variance in a manner which keeps the Type I error rate close to its target value. In Section 4 we use the significance level method to adapt the two-sided, inner wedge tests of Pampallona & Tsiatis (1994) to unequal group sizes and unknown variance. We consider the same type of test in Section 5, but with applications to equivalence testing in mind, and this time we apply a modification of the significance level method which maintains target error rates accurately at specified non-null parameter values. We conclude in Section 6 with some remarks on the scope of application of the significance level approach.

2 Two-sided tests with no inner wedge

Consider a group sequential test with a maximum of K analyses in which standardised statistics $Z_1, \dots, Z_K$ follow the canonical distribution (1). Suppose it is required to test the null hypothesis $H_0: \theta = 0$ against a two-sided alternative with Type I error probability α and power $1 - \beta$ at $\theta = \pm\delta$. Note that the information for θ required by a fixed sample test with this Type I error probability is

$$\mathcal{I}_{f,2} = \{\Phi^{-1}(1 - \alpha/2) + \Phi^{-1}(1 - \beta)\}^2/\delta^2, \quad (2)$$

the subscript f in $\mathcal{I}_{f,2}$ standing for “fixed” and 2 denoting a two-sided test. Here $\Phi(x)$ denotes the cumulative distribution function of the standard normal distribution.

A two-sided test allowing early stopping to reject H_0 , but not to accept H_0 , can be written in the form:

$$\begin{aligned}
 &\text{After group } k = 1, \dots, K - 1 \\
 &\quad \text{if } |Z_k| \geq c_k \quad \text{stop, reject } H_0 \\
 &\quad \text{otherwise} \quad \text{continue to group } k + 1, \\
 &\text{after group } K \\
 &\quad \text{if } |Z_K| \geq c_K \quad \text{stop, reject } H_0 \\
 &\quad \text{otherwise} \quad \text{stop, accept } H_0.
 \end{aligned} \tag{3}$$

It is evident from (1) that the joint distribution of $\{Z_1, \dots, Z_K\}$ under $\theta = 0$ depends only on the ratios between the information levels $\{\mathcal{I}_1, \dots, \mathcal{I}_K\}$. Hence, critical values $c_1, \dots, c_K$ chosen to satisfy the Type I error condition for a sequence of equally spaced information levels will yield the same Type I error probability for all such sequences.

Assuming information levels will be equally spaced, the power condition is met by fixing the value of the maximum information level $\mathcal{I}_K$. The ratio of this maximum information level to $\mathcal{I}_{f,2}$, the information needed for a fixed sample test with the same Type I error probability and power, depends on α , β and K and the chosen sequence of critical values but not on the value of δ . Hence, it is convenient to tabulate the ratio $\mathcal{I}_K/\mathcal{I}_{f,2}$ along with constants determining $c_1, \dots, c_K$ for a particular type of test.

In situations where it is difficult to control information levels at each analysis precisely, some assumptions about their likely values must be made at the design stage. We shall suppose that the amounts of information accruing between successive analyses are expected to be roughly equal and so it is reasonable to design tests on the basis that $\mathcal{I}_k \propto k$. The attained Type I error probability, given the sequence of information levels actually occurring, will depend on the ratios between the $\mathcal{I}_k$, $k = 1, \dots, K$. The attained power is influenced most by the overall magnitude of the $\mathcal{I}_k$ s and in particular by the final information level, $\mathcal{I}_K$.

Two popular choices of group sequential test are Pocock’s (1977) repeated significance test in which c_k is constant over k and O’Brien & Fleming’s (1979) test in which c_k is proportional to $k^{-1/2}$ and so decreases as the analysis number increases. The Pocock test with K groups of data and Type I error rate α has critical values

$$c_k = C_P(K, \alpha), \quad k = 1, \dots, K,$$

and requires a maximum information level

$$\mathcal{I}_K = R_P(K, \alpha, \beta) \times \mathcal{I}_{f,2}$$

Table 1: Constants defining Pocock and O’Brien & Fleming two-sided tests with K groups of observations, Type I error probability $\alpha = 0.05$ and power $1 - \beta = 0.9$ at $\theta = \pm\delta$

K	Pocock		O’Brien & Fleming	
	$C_P(K, \alpha)$	$R_P(K, \alpha, \beta)$	$C_B(K, \alpha)$	$R_B(K, \alpha, \beta)$
3	2.289	1.151	2.004	1.016
5	2.413	1.207	2.040	1.026
10	2.555	1.271	2.087	1.037

to achieve power β at the specified alternative $\theta = \pm\delta$. Critical values for the O’Brien & Fleming test are

$$c_k = C_B(K, \alpha) \sqrt{K/k}, \quad k = 1, \dots, K,$$

and this test requires a maximum information level

$$\mathcal{I}_K = R_B(K, \alpha, \beta) \times \mathcal{I}_{f,2}.$$

We shall use Pocock and O’Brien & Fleming tests for $K = 3, 5$ and 10 , $\alpha = 0.05$ and $1 - \beta = 0.9$ as our illustrative examples. Constants determining these tests are given in Table 1; for more complete tabulations of such constants see, for example, Pocock (1977) or Jennison & Turnbull (1989).

If the test (3) is applied with

$$\mathcal{I}_k = (k/K)R \times \mathcal{I}_{f,2}, \quad k = 1, \dots, K,$$

where R is $R_P(K, \alpha, \beta)$ for a Pocock test or $R_B(K, \alpha, \beta)$ for an O’Brien & Fleming test, Type I error rate α and power $1 - \beta$ at $\theta = \pm\delta$ will be attained exactly. In practice, it is often not possible to guarantee such equally spaced information levels. The rule (3) can be still applied to the standardised statistics Z_k calculated from the data available at analyses $k = 1, \dots, K$. Since each Z_k has a marginal $N(0, 1)$ distribution under $H_0: \theta = 0$, the criterion for stopping to reject H_0 at analysis k is that a single hypothesis test based on Z_k should reject H_0 at a two-sided significance level less than or equal to $2\{1 - \Phi(c_k)\}$, hence we refer to this as the “significance level” approach to dealing with unequally spaced information levels.

Table 2 shows the Type I error probability and power achieved when information levels depart from their planned values. These results were obtained by numerical computation and are accurate to the three decimal places displayed. In each case, the initial design is for a Pocock or O’Brien & Fleming test with Type I error probability $\alpha = 0.05$ and power $1 - \beta = 0.9$ at $\theta = \pm\delta$. The intended information levels are $\mathcal{I}_k = (k/K)\mathcal{I}_{max}$, $k = 1, \dots, K$,

Table 2: Properties of two-sided tests designed for information levels $\mathcal{I}_k = (k/K)\mathcal{I}_{max}$ and applied with $\mathcal{I}_k = \pi(k/K)^r\mathcal{I}_{max}$, $k = 1, \dots, K$. The table shows Type I error probabilities and power attained by Pocock and O'Brien & Fleming tests with K analyses, designed to achieve Type I error rate $\alpha = 0.05$ and power $1 - \beta = 0.9$ at $\theta = \pm\delta$.

			Pocock		O'Brien & Fleming	
	r	π	Type I error	Power	Type I error	Power
$K = 3$						
	0.70	0.9	0.046	0.868	0.048	0.867
		1.0	0.046	0.902	0.048	0.899
		1.1	0.046	0.927	0.048	0.925
	1.00	0.9	0.050	0.866	0.050	0.867
		1.0	0.050	0.900	0.050	0.900
		1.1	0.050	0.926	0.050	0.925
	1.30	0.9	0.053	0.864	0.051	0.868
		1.0	0.053	0.899	0.051	0.900
		1.1	0.053	0.925	0.051	0.925
$K = 5$						
	0.70	0.9	0.045	0.867	0.047	0.866
		1.0	0.045	0.902	0.047	0.899
		1.1	0.045	0.928	0.047	0.924
	1.00	0.9	0.050	0.866	0.050	0.867
		1.0	0.050	0.900	0.050	0.900
		1.1	0.050	0.927	0.050	0.925
	1.30	0.9	0.054	0.864	0.052	0.868
		1.0	0.054	0.899	0.052	0.901
		1.1	0.054	0.925	0.052	0.926
$K = 10$						
	0.70	0.9	0.043	0.866	0.046	0.864
		1.0	0.043	0.901	0.046	0.898
		1.1	0.043	0.928	0.046	0.923
	1.00	0.9	0.050	0.865	0.050	0.867
		1.0	0.050	0.900	0.050	0.900
		1.1	0.050	0.927	0.050	0.925
	1.30	0.9	0.056	0.863	0.053	0.869
		1.0	0.056	0.898	0.053	0.901
		1.1	0.056	0.926	0.053	0.926

where $\mathcal{I}_{max}$ is the multiple of $\mathcal{I}_{f,2}$ needed to meet the power requirement, but observed information levels are $\mathcal{I}_k = \pi(k/K)^r \mathcal{I}_{max}$, $k = 1, \dots, K$. The final information level, $\mathcal{I}_K = \pi \mathcal{I}_{max}$, is controlled by π and we see that this is the key factor in determining power. The attained power lies within 0.002 of its nominal level in all cases where $\pi = 1$ and increases as π , and hence $\mathcal{I}_{max}$, increases. When $\pi = 0.9$, power is close to the value 0.868 attained by a fixed sample test with 0.9 times the required information, $\mathcal{I}_{f,2}$, and when $\pi = 1.1$ power is close to the value 0.925 achieved by a fixed sample test with information $1.1 \times \mathcal{I}_{f,2}$.

The exponent r governs the spacing of the information levels. The $\mathcal{I}_k$ are equally spaced if $r = 1$, bunched together towards $\mathcal{I}_{max}$ if $r < 1$, and bunched towards zero when $r > 1$. As an example, values $\mathcal{I}_k = 10k$, $k = 1, \dots, 10$, when $r = 1$ become $\mathcal{I}_k = \{20, 32, 43, 53, 62, 70, 78, 86, 93, 100\}$ if $r = 0.7$, and $\mathcal{I}_k = \{5, 12, 21, 30, 41, 51, 63, 75, 87, 100\}$ if $r = 1.3$. The Type I error probability is reduced below its nominal value for $r < 1$ and increased for $r > 1$ but these effects are slight, the attained Type I error rate lying between 0.043 and 0.056 in all cases in Table 2.

These results demonstrate that approximate treatment of unequal group sizes through the significance level approach will be quite satisfactory in many applications. The O'Brien & Fleming test is particularly robust to variations in early group sizes as it allocates very little Type I error probability to the first few analyses.

Pocock (1977) proposed using the same significance level approach to construct group sequential t -tests for normal observations of unknown variance. Suppose data follow a normal linear model and, as before, θ is a scalar parameter. We wish to create a group sequential test of the hypothesis $H_0: \theta = 0$ with two-sided Type I error probability close to α . If observations' variances and covariances are known up to a factor σ^2 , the sequence of maximum likelihood estimates $\{\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(K)}\}$ is still multivariate normal with $Cov(\hat{\theta}^{(k_1)}, \hat{\theta}^{(k_2)}) = Var(\hat{\theta}^{(k_2)})$ for $k_1 \leq k_2$ but now $Var(\hat{\theta}^{(k)}) = V_k \sigma^2$, $k = 1, \dots, K$, where the V_k are known and σ^2 is unknown. Denote the usual unbiased estimates of σ^2 by s_k^2 , $k = 1, \dots, K$, so each $s_k^2 \sim \sigma^2 \chi_{\nu_k}^2 / \nu_k$ marginally, where ν_k is the number of degrees of freedom available for estimating σ^2 at analysis k . In general ν_k will equal the number of observations accrued by analysis k minus p , the length of the parameter vector in the linear model. If σ^2 were known, the z -statistic for testing $H_0: \theta = 0$ at analysis k would be

$$Z_k = \frac{\hat{\theta}^{(k)}}{\sqrt{\{V_k \sigma^2\}}}$$

and the t -statistic for testing H_0 is, therefore,

$$T_k = \frac{\hat{\theta}^{(k)}}{\sqrt{\{V_k s_k^2\}}}.$$

In adapting a known-variance test of form (3), we use the same two-sided significance levels, $2\{1 - \Phi(c_k)\}$, but apply these to the t -statistics $T_1, \dots, T_K$. Since T_k has, marginally, a t_{ν_k} distribution, we reject H_0 at analysis k if

$$|T_k| \geq t_{\nu_k, 1-\Phi(c_k)}, \quad k = 1, \dots, K,$$

where $t_{\nu, q}$ denotes the upper q tail-point of a t -distribution on ν degrees of freedom, i.e., $Pr\{T > t_{\nu, q}\} = q$ when $T \sim t_\nu$. If H_0 has not been rejected by analysis K , it is accepted. The sequence of critical values, $\{c_1, \dots, c_K\}$, can be that of a Pocock test, an O'Brien & Fleming test, or any other two-sided test for K group of observations with Type I error rate α . Since these critical values were derived for a sequence of z -statistics, this form of test will only satisfy the Type I error requirement approximately. Nevertheless, our investigations have shown that approximation to be remarkably accurate.

Table 3 shows empirical Type I error probabilities for group sequential two-sided tests defined in this way, based on simulations with 50,000 replications. In choosing the degrees of freedom ν_k , we have supposed data arrive in equally sized groups of 3, 4 or 5 observations but in each analysis two degrees of freedom are used in estimating model parameters. The total number of observations at analysis K is, therefore, $\nu_K + 2$ and at each analysis $k = 1, \dots, K$ we have $(\nu_K + 2)(k/K)$ observations and $\nu_k = (\nu_K + 2)(k/K) - 2$ error degrees of freedom. This sequence of ν_k values might arise in the group sequential analysis of a linear regression model. However, it follows from Jennison & Turnbull's (1997b) results concerning the joint distribution of the sequence of pairs of estimates $\{(\hat{\theta}^{(1)}, s_1^2), \dots, (\hat{\theta}^{(K)}, s_K^2)\}$ and of the sequence of t -statistics $\{T_1, \dots, T_K\}$ that the same Type I error rates occur in any group sequential setting with the same error degrees of freedom, $\nu_1, \dots, \nu_K$, and $V_k^{-1} \propto k$, $k = 1, \dots, K$. It is evident from the table that departures from the desired Type I error rate are quite minor, the worst discrepancies arising when the degrees of freedom are low at early analyses. Since the O'Brien & Fleming test imposes wide boundaries at the first few analyses it is not surprising that this test has Type I error rates closer to 0.05 than the Pocock test.

As power depends on the value of σ^2 , it is not possible for a test with pre-specified group sizes to guarantee power at a given value of θ . What we can do is find a good approximation to the power at any given value of σ^2 and thereby check whether a design will achieve sufficient power at a stated alternative $|\theta| = \delta$ if σ^2 lies in some plausible range. The information for θ at the final analysis is $Var(\hat{\theta}^{(K)})^{-1} = (V_K \sigma^2)^{-1}$. By analogy with the case of known σ^2 , we can expect the group sequential t -test's power at a given alternative to be approximately that of a fixed sample t -test with $1/R(K, \alpha, \beta)$ times the maximal information, where R is taken to mean R_P for a Pocock test or R_B for an O'Brien & Fleming test. Thus, we should find the power of a fixed sample test with estimator $\hat{\theta}$ for which $Var(\hat{\theta}) = R(K, \alpha, \beta) V_K \sigma^2$. If there are ν_K degrees of freedom for estimating σ^2 , the fixed sample t -statistic has a t_{ν_K} distribution under H_0 and a $T_{\nu_K}(\delta/\sqrt{\{R(K, \alpha, \beta) V_K \sigma^2\}})$ distribution

Table 3: Properties of group sequential t -tests designed to test $H_0: \theta = 0$ with two-sided Type I error $\alpha = 0.05$. Degrees of freedom for σ^2 at analysis k are $\nu_k = (\nu_K + 2)(k/K) - 2$ and $Var(\hat{\theta}^{(k)}) = (K/k)Var(\hat{\theta}^{(K)})$, $k = 1, \dots, K$. The penultimate column shows actual power at the alternative for which the approximation (4) gives power 0.9 and the final column shows the power predicted for each test by formula (5).

Type I error probability and power were estimated by simulation using 50,000 replicates; standard errors are 0.001 for Type I error and 0.0015 for attained power.

Pocock test

K	ν_K	Type I error	Attained power	Known σ^2 power
3	7	0.060	0.855	0.967
3	10	0.056	0.867	0.950
3	13	0.054	0.879	0.940
5	13	0.058	0.870	0.940
5	18	0.055	0.882	0.929
5	23	0.054	0.885	0.923
10	28	0.058	0.881	0.919
10	38	0.055	0.888	0.914
10	48	0.056	0.891	0.911

O'Brien & Fleming test

K	ν_K	Type I error	Attained power	Known σ^2 power
3	7	0.052	0.896	0.967
3	10	0.054	0.899	0.950
3	13	0.052	0.899	0.940
5	13	0.053	0.896	0.940
5	18	0.054	0.900	0.929
5	23	0.052	0.896	0.923
10	28	0.055	0.899	0.919
10	38	0.053	0.899	0.914
10	48	0.052	0.900	0.911

if $\theta = \delta$. Here $T_\nu(\xi)$ denotes a non-central t -distribution with ν degrees of freedom and non-centrality parameter ξ . Assuming $\delta > 0$ and ignoring the negligible probability of rejecting H_0 with a large, negative t -statistic, this fixed sample test has power

$$Pr\{T_{\nu_K}(\delta/\sqrt{\{R(K, \alpha, \beta) V_K \sigma^2\}}) \geq t_{\nu_K, 1-\alpha/2}\}, \quad (4)$$

and this forms our approximation to the group sequential t -test's power when $\theta = \delta$ and the variance is σ^2 . If ν_K is large and we disregard the error in estimating σ^2 we obtain the cruder but simpler estimate of power

$$\Phi\left(\frac{\delta}{\sqrt{\{R(K, \alpha, \beta) V_K \sigma^2\}}} - \Phi^{-1}(1 - \alpha/2)\right). \quad (5)$$

Since $R(K, \alpha, \beta)$ itself depends on the power we are in the process of estimating, the above discussion appears somewhat circular. In fact, the factors $R_P(K, \alpha, \beta)$ and $R_B(K, \alpha, \beta)$ vary only slightly with β and it suffices to work with a factor appropriate to a plausible value of the power, $1 - \beta$, in the above calculation.

Table 3 shows the power attained by group sequential t -tests, defined in the above manner, in situations where formula (4) gives the value 0.9. We see from the table that this approximation provides an accurate estimate of power as long as the degrees of freedom are reasonably high. The over-estimation of power when degrees of freedom are low is mainly attributable to the use of ν_K in setting the target information levels, whereas t -tests at early analyses have fewer degrees of freedom than this. Accordingly, discrepancies in the power approximation are particularly small for O'Brien & Fleming tests, under which there is only a very low probability of stopping at the earliest analyses. The final column of Table 3 shows the approximation to the power obtained using (5) and, not surprisingly, this simpler formula gives good results when degrees of freedom are high but is unreliable when the final degrees of freedom are as low as 10 or 20. Again, the general theory of Jennison & Turnbull (1997b) can be used to deduce that the joint distribution of $\{T_1, \dots, T_K\}$ at an alternative $\theta = \delta$ depends only on the values of $\delta/\sqrt{\{V_1 \sigma^2\}}, \dots, \delta/\sqrt{\{V_K \sigma^2\}}$ and the degrees of freedom $\nu_1, \dots, \nu_K$, hence the attained powers in the table also apply to a more general class of problems.

3 One-sided tests

We now turn our attention to group sequential tests of a null hypothesis against a one-sided alternative. As in Sections 1 and 2, we assume standardised statistics $\{Z_1, \dots, Z_K\}$ following the canonical joint distribution (1) are available at analyses $k = 1, \dots, K$. We suppose it is required to test $H_0: \theta = 0$ against $H_A: \theta > 0$ with Type I error probability

α and power $1 - \beta$ at $\theta = \delta$. A fixed sample test based on a statistic $Z \sim N(\theta\sqrt{\mathcal{I}}, 1)$ requires an information level $\mathcal{I}$ equal to

$$\mathcal{I}_{f,1} = \{\Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta)\}^2 / \delta^2 \quad (6)$$

in order to attain power $1 - \beta$ when $\theta = \delta$. As for two-sided tests, we shall express the maximum information levels of group sequential one-sided tests as multiples of the fixed sample information level, $\mathcal{I}_{f,1}$.

A group sequential one-sided test with K analyses can be written as:

$$\begin{aligned} &\text{After group } k = 1, \dots, K - 1 \\ &\quad \text{if } Z_k \geq b_k \quad \text{stop, reject } H_0 \\ &\quad \text{if } Z_k \leq a_k \quad \text{stop, accept } H_0 \\ &\quad \text{otherwise} \quad \text{continue to group } k + 1, \\ &\text{after group } K \\ &\quad \text{if } Z_K \geq b_K \quad \text{stop, reject } H_0 \\ &\quad \text{if } Z_K < a_K \quad \text{stop, accept } H_0. \end{aligned} \quad (7)$$

Here a_K must be set equal to b_K in order to ensure that the test terminates at analysis K .

We shall confine our attention to “power family” tests for which

$$\begin{aligned} b_k &= \tilde{C}_1(K, \alpha, \beta, \Delta)(k/K)^{\Delta-1/2} \quad \text{and} \\ a_k &= \delta\sqrt{\mathcal{I}_k} - \tilde{C}_2(K, \alpha, \beta, \Delta)(k/K)^{\Delta-1/2}, \quad k = 1, \dots, K. \end{aligned} \quad (8)$$

This family of tests was introduced by Emerson & Fleming (1989) for the case $\alpha = \beta$ and extended to include asymmetric tests with $\alpha \neq \beta$ by Pampallona & Tsiatis (1994). The family is indexed by a parameter Δ which controls the shape of the stopping boundary. For a low value of Δ , the testing boundary is wide early on and then narrows rapidly, converging at a relatively low information level just a little greater than $\mathcal{I}_{f,1}$; increasing Δ produces continuation regions which are narrower at the first analyses but extend further along the information scale, with the final possible analysis at relatively high information levels. In order that $a_K = b_K$, a final information level

$$\mathcal{I}_K = \frac{\{\tilde{C}_1(K, \alpha, \beta, \Delta) + \tilde{C}_2(K, \alpha, \beta, \Delta)\}^2}{\delta^2} \quad (9)$$

is needed. Since these tests are designed under the assumption of equally spaced information levels, $\mathcal{I}_k = (k/K)\mathcal{I}_K$, $k = 1, \dots, K$, we can re-express the above formulae for a_k and b_k as

$$\begin{aligned} b_k &= \tilde{C}_1(K, \alpha, \beta, \Delta)(k/K)^{\Delta-1/2} \quad \text{and} \\ a_k &= \{\tilde{C}_1(K, \alpha, \beta, \Delta) + \tilde{C}_2(K, \alpha, \beta, \Delta)\}(k/K)^{1/2} \\ &\quad - \tilde{C}_2(K, \alpha, \beta, \Delta)(k/K)^{\Delta-1/2}, \quad k = 1, \dots, K. \end{aligned} \quad (10)$$

Table 4: Constants $\tilde{C}_1(K, \alpha, \beta, \Delta)$, $\tilde{C}_2(K, \alpha, \beta, \Delta)$ and $\tilde{R}(K, \alpha, \beta, \Delta)$ defining power family one-sided tests with parameter Δ , K groups of observations, Type I error probability $\alpha = 0.05$ and power $1 - \beta = 0.9$ at $\theta = \delta$

	$\Delta = 0$			$\Delta = 0.25$		
K	$\tilde{C}_1$	$\tilde{C}_2$	$\tilde{R}$	$\tilde{C}_1$	$\tilde{C}_2$	$\tilde{R}$
3	1.673	1.357	1.072	1.746	1.438	1.184
5	1.696	1.389	1.111	1.788	1.490	1.255
10	1.725	1.425	1.158	1.836	1.546	1.336

This definition, which avoids explicit mention of $\mathcal{I}_k$, will be used in adapting the tests to unequally spaced information levels.

We illustrate our methods for tests with $K = 3, 5$ and 10 analyses, and shape parameter $\Delta = 0$ and 0.25, designed to achieve Type I error rate $\alpha = 0.05$ and power $1 - \beta = 0.9$ at $\theta = \delta$. Constants $\tilde{C}_1(K, \alpha, \beta, \Delta)$ and $\tilde{C}_2(K, \alpha, \beta, \Delta)$ defining these tests are shown in Table 4. Values of $\tilde{R}(K, \alpha, \beta, \Delta)$, the ratio of the maximum information level to that of the fixed sample test are also shown. It follows from (6) and (9) that this ratio is

$$\tilde{R}(K, \alpha, \beta, \Delta) = \frac{\mathcal{I}_K}{\mathcal{I}_{f,1}} = \frac{\{\tilde{C}_1(K, \alpha, \beta, \Delta) + \tilde{C}_2(K, \alpha, \beta, \Delta)\}^2}{\{\Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta)\}^2}.$$

Constants for other tests can be found in Emerson & Fleming (1989) and Pampallona & Tsiatis (1994). (These authors have expressed tests in terms of the score statistics $S_k = Z_k \sqrt{\mathcal{I}_k}$, whereas we have worked with the standardised statistics Z_k , $k = 1, \dots, K$, and consequently our $\tilde{C}_j(K, \alpha, \beta, \Delta)$ are $K^{\Delta-1/2}$ times the $C_j(K, \alpha, \beta, \Delta)$, $j = 1$ and 2, of Emerson & Fleming and Pampallona & Tsiatis.)

If information levels

$$\mathcal{I}_k = (k/K) \tilde{R}(K, \alpha, \beta, \Delta) \mathcal{I}_{f,1}, \quad k = 1, \dots, K, \quad (11)$$

arise as planned, the Type I error and power conditions will be satisfied exactly. Suppose instead that information levels differ from their planned values and our first priority is to maintain a Type I error rate as close as possible to α . Following the significance level approach, we apply rule (7) and compare the standardised statistics Z_k to the same sequence of critical values we would have used if information levels had equalled their planned values. In doing this, critical values a_k and b_k are obtained from equation (10) directly in terms of $\tilde{C}_1(K, \alpha, \beta, \Delta)$ and $\tilde{C}_2(K, \alpha, \beta, \Delta)$. Under H_0 : $\theta = 0$, the joint distribution of $\{Z_1, \dots, Z_K\}$ depends only on the ratios between $\{\mathcal{I}_1, \dots, \mathcal{I}_K\}$ and, since the critical values $(a_1, b_1), \dots, (a_K, b_K)$ lead to Type

I error probability α in the planned design (11), the Type I error rate is also α under any sequence of equally spaced information levels. Unequal increments in information levels cause perturbations in the attained Type I error rate whereas power at $\theta = \delta$ is affected primarily by the overall scale of observed information levels.

Table 5 shows Type I error probabilities and power achieved when information levels depart from their planned values. The observed sequences of information levels are the same as those considered for two-sided tests in Table 2: planned information levels are $\mathcal{I}_k = (k/K)\mathcal{I}_{max}$, $k = 1, \dots, K$, where $\mathcal{I}_{max} = \tilde{R}(K, \alpha, \beta, \Delta)\mathcal{I}_{f,1}$ and actual information levels are $\mathcal{I}_k = \pi(k/K)^r\mathcal{I}_{max}$. We see that the Type I error rate is controlled very closely, remaining within 0.002 of $\alpha = 0.05$ in all cases. The attained power is more variable, being affected both by the overall scale of the information levels and their spacing. Tests with $\Delta = 0$ have wide early boundaries, so changes in the values of the first few information levels have relatively small effects. In comparison, tests with $\Delta = 0.25$ offer a realistic opportunity for stopping at all stages and the parameter r , which controls spacing between information levels, has a more substantial effect on power.

There are other ways in which one might try to adapt these tests to unequal group sizes. In (8), it is natural to view the upper boundary as a repeated significance test of $\theta = 0$ against $\theta > 0$ and the lower boundary as a repeated significance test of $\theta = \delta$ against $\theta < \delta$. In testing $\theta = 0$, critical values for the Z_k are proportional to $k^{\Delta-1/2}$, and in testing $\theta = \delta$ one compares $Z_k - \delta\sqrt{\mathcal{I}_k}$, the standardised statistics with mean zero if $\theta = \delta$, with critical values proportional to $k^{\Delta-1/2}$, $k = 1, \dots, K$. Simple application of the stopping rule (7) with a_k and b_k obtained by substituting the observed information levels $\mathcal{I}_1, \dots, \mathcal{I}_K$ into (8) maintains these two sequences of significance levels for tests of $\theta = 0$ and $\theta = \delta$. However, the boundaries will only meet up with a_K equal to b_K if $\mathcal{I}_K$ is exactly equal to the target given in (9) and some modification to this approach is needed to ensure an unambiguous conclusion at the final analysis. More importantly, this approach maintains the Type I error rate much less accurately than our proposal described above.

Emerson & Fleming (1989) have suggested a way of handling general information sequences in which boundary values are interpolated on the score statistic scale. We have found this method to be quite effective in maintaining Type I error probability and power but perturbations in Type I error rates are greater than for our proposed method: in the examples of Table 5, Type I error probabilities for Emerson & Fleming's method vary about 0.05 for each combination of K and Δ and range from 0.045 to 0.056 overall.

Suppose now that observations follow a normal linear model with unknown variance and we wish to test $H_0: \theta = 0$ against the one-sided alternative $H_A: \theta > 0$ with Type I error rate close to α . As in Section 2, we suppose the sequence of estimates $\{\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(K)}\}$ has the canonical joint distribution (1) and so is multivariate normal with $Cov(\hat{\theta}^{(k_1)}, \hat{\theta}^{(k_2)}) =$

Table 5: Properties of power family one-sided tests designed for information levels $\mathcal{I}_k = (k/K)\mathcal{I}_{max}$ and applied with $\mathcal{I}_k = \pi(k/K)^r\mathcal{I}_{max}$, $k = 1, \dots, K$. The table shows Type I error probabilities and power attained by tests with shape parameter Δ and K analyses, designed to achieve Type I error probability $\alpha = 0.05$ and power $1 - \beta = 0.9$ at $\theta = \delta$.

		$\Delta = 0$		$\Delta = 0.25$	
r	π	Type I error	Power	Type I error	Power
$K = 3$					
0.70	0.9	0.050	0.883	0.050	0.894
	1.0	0.050	0.910	0.050	0.920
	1.1	0.050	0.932	0.050	0.940
1.00	0.9	0.050	0.871	0.050	0.872
	1.0	0.050	0.900	0.050	0.900
	1.1	0.050	0.923	0.050	0.922
1.30	0.9	0.049	0.853	0.049	0.838
	1.0	0.049	0.883	0.049	0.869
	1.1	0.049	0.907	0.049	0.893
$K = 5$					
0.70	0.9	0.050	0.887	0.049	0.900
	1.0	0.050	0.914	0.049	0.925
	1.1	0.050	0.935	0.049	0.944
1.00	0.9	0.050	0.871	0.050	0.872
	1.0	0.050	0.900	0.050	0.900
	1.1	0.050	0.923	0.050	0.922
1.30	0.9	0.049	0.847	0.050	0.829
	1.0	0.049	0.878	0.050	0.860
	1.1	0.049	0.903	0.050	0.885
$K = 10$					
0.70	0.9	0.049	0.891	0.048	0.906
	1.0	0.049	0.918	0.048	0.931
	1.1	0.049	0.938	0.048	0.949
1.00	0.9	0.050	0.871	0.050	0.872
	1.0	0.050	0.900	0.050	0.900
	1.1	0.050	0.923	0.050	0.922
1.30	0.9	0.050	0.841	0.050	0.819
	1.0	0.050	0.872	0.050	0.851
	1.1	0.050	0.898	0.050	0.877

$Var(\hat{\theta}^{(k_2)})$ for $k_1 \leq k_2$. Let $Var(\hat{\theta}^{(k)}) = V_k \sigma^2$, $k = 1, \dots, K$, where the V_k are known and σ^2 is unknown and denote the usual estimates of σ^2 by $s_k^2 \sim \sigma^2 \chi_{\nu_k}^2 / \nu_k$, $k = 1, \dots, K$, where ν_k is the number of degrees of freedom for estimating σ^2 at analysis k . The t -statistic for testing H_0 at analysis k is then

$$T_k = \frac{\hat{\theta}^{(k)}}{\sqrt{\{V_k s_k^2\}}}.$$

To adapt a known-variance test of the form (7), we replace each Z_k by T_k and set critical values for the T_k to preserve the marginal probability under H_0 (ignoring the effect of possible earlier stopping) that T_k should lie above or below the stopping boundary. This gives the rule:

$$\begin{aligned} &\text{After group } k = 1, \dots, K - 1 \\ &\quad \text{if } T_k \geq t_{\nu_k, 1-\Phi(b_k)} \quad \text{stop, reject } H_0 \\ &\quad \text{if } T_k \leq t_{\nu_k, 1-\Phi(a_k)} \quad \text{stop, accept } H_0 \\ &\quad \text{otherwise} \quad \text{continue to group } k + 1, \\ &\text{after group } K \\ &\quad \text{if } T_K \geq t_{\nu_K, 1-\Phi(b_K)} \quad \text{stop, reject } H_0 \\ &\quad \text{if } T_K < t_{\nu_K, 1-\Phi(a_K)} \quad \text{stop, accept } H_0. \end{aligned} \tag{12}$$

Here the a_k and b_k are as defined in (10) and, as before, $t_{\nu, q}$ denotes the upper q tail-point of a t -distribution on ν degrees of freedom.

Note that the “significance level” approach has been followed since rule (12) applies the same significance levels to the T_k as rule (7) applies to the Z_k . The comparison may seem a little strained since the lower boundary does not lie at extreme values of the null distribution of the Z_k s. However, there are good reasons to expect this construction to produce a test with Type I error rate close to the target α . First, if the degrees of freedom ν_k are high, our group sequential t -test closely resembles the z -test, which we have already seen to be effective in handling unpredictable information levels. Thus, we can expect the Type I error rate of our group sequential t -test to be close to its intended value if information levels are fairly equally spaced. This is a key property as it overcomes the dependence of information levels on the unknown σ^2 . Secondly, if the degrees of freedom are low, matching the marginal probabilities under H_0 of being above each boundary point to those of the z -test is a useful first step towards matching the overall probability under H_0 of crossing the upper boundary, which is of course the Type I error probability.

We can repeat the arguments of Section 2 to obtain an approximation to our test’s power at given values of θ and σ^2 . By analogy with the case of known σ^2 , for given $V_1, \dots, V_K$ we expect the group sequential t -test’s power at a particular alternative to be approximately that of a fixed sample t -test with $1/\tilde{R}(K, \alpha, \beta, \Delta)$ times the maximal information. So, we need to find the power of a fixed sample test with estimator $\hat{\theta}$ for which

$Var(\hat{\theta}) = \tilde{R}(K, \alpha, \beta, \Delta) V_K \sigma^2$. This gives the approximation to the power when $\theta = \delta$, corresponding to formula (4) for a two-sided test, of

$$Pr\{T_{\nu_K}(\delta/\sqrt{\{\tilde{R}(K, \alpha, \beta, \Delta) V_K \sigma^2\}}) \geq t_{\nu_K, 1-\alpha}\}, \quad (13)$$

where, as before, $T_{\nu}(\xi)$ denotes a non-central t -distribution with ν degrees of freedom and non-centrality parameter ξ . The simpler approximation for use when ν_K is large is

$$\Phi\left(\frac{\delta}{\sqrt{\{\tilde{R}(K, \alpha, \beta, \Delta) V_K \sigma^2\}}} - \Phi^{-1}(1 - \alpha)\right). \quad (14)$$

Table 6 shows the Type I error probabilities of our group sequential one-sided t -tests and the power they attain in situations where formula (13) gives the value 0.9. As in Table 3, the degrees of freedom are those which arise when groups consist of 3, 4 or 5 observations and two degrees of freedom are used in fitting the linear model at each analysis. Departures from the intended Type I error rates are minor and decrease as the degrees of freedom increase. Formula (13) is seen to provide a good estimate of power when degrees of freedom are reasonably high. Since tests with $\Delta = 0.0$ have less opportunity for early stopping than those with $\Delta = 0.25$, the slightly higher discrepancies in Type I error rates in the second case are as one might expect.

Since our definition of group sequential t -tests does not rely on equally spaced information levels, these tests may also be used with unequal group sizes or information increments. We have simulated tests with sequences of information levels of the form

$$\mathcal{I}_k = (k/K)^r \mathcal{I}_{max}, \quad k = 1, \dots, K, \quad (15)$$

to investigate the effects of such unequal spacing. Our findings are similar to those reported earlier in this section for the known variance one-sided testing problem. With $r = 0.7$ and 1.3 in (15), the Type I error rate and power of a test differ only slightly from those of the same test with $r = 1$, i.e., with equally spaced information levels and $\mathcal{I}_K = \mathcal{I}_{max}$. Thus, there is no difficulty and no evidence of any serious loss of accuracy in applying these tests with both an unknown variance and unequal information increments.

4 Two-sided inner wedge tests

We return to the two-sided testing problem of Section 2 but now consider tests which allow early stopping to accept, as well as to reject, the null hypothesis. Again we suppose standardised statistics $\{Z_1, \dots, Z_K\}$ following the canonical joint distribution (1) are available at analyses $k = 1, \dots, K$. Suppose we wish to test, group sequentially, the null hypothesis $H_0: \theta = 0$ against a two-sided alternative with Type I error probability α and power $1 - \beta$ at $\theta = \pm\delta$.

Table 6: Properties of group sequential one-sided t -tests designed to test $H_0: \theta = 0$ with Type I error $\alpha = 0.05$, based on power family tests for known variance with $\Delta = 0$ and 0.25. Degrees of freedom for σ^2 at analysis k are $\nu_k = (\nu_K + 2)(k/K) - 2$ and $Var(\hat{\theta}^{(k)}) = (K/k)Var(\hat{\theta}^{(K)})$, $k = 1, \dots, K$. The penultimate column shows actual power at the alternative for which the approximation (13) gives power 0.9 and the final column shows the power predicted for each test by formula (14).

Type I error and power were estimated by simulation using 50,000 replicates; standard errors are 0.001 for Type I error and 0.0015 for attained power.

$\Delta = 0$

K	ν_K	Type I error	Attained power	Known σ^2 power
3	7	0.057	0.912	0.947
3	10	0.057	0.907	0.934
3	13	0.055	0.908	0.926
5	13	0.055	0.908	0.926
5	18	0.055	0.907	0.919
5	23	0.053	0.903	0.915
10	28	0.054	0.905	0.912
10	38	0.054	0.904	0.909
10	48	0.052	0.905	0.907

$\Delta = 0.25$

K	ν_K	Type I error	Attained power	Known σ^2 power
3	7	0.060	0.913	0.947
3	10	0.054	0.911	0.934
3	13	0.057	0.906	0.926
5	13	0.058	0.907	0.926
5	18	0.054	0.903	0.919
5	23	0.055	0.906	0.915
10	28	0.055	0.903	0.912
10	38	0.053	0.903	0.909
10	48	0.052	0.901	0.907

In general, a group sequential test with possible early stopping to accept H_0 has the form:

$$\begin{aligned}
&\text{After group } k = 1, \dots, K - 1 \\
&\quad \text{if } |Z_k| \geq b_k \quad \text{stop, reject } H_0 \\
&\quad \text{if } |Z_k| < a_k \quad \text{stop, accept } H_0 \\
&\quad \text{otherwise} \quad \text{continue to group } k + 1, \\
&\text{after group } K \\
&\quad \text{if } |Z_K| \geq b_K \quad \text{stop, reject } H_0 \\
&\quad \text{if } |Z_K| < a_K \quad \text{stop, accept } H_0.
\end{aligned} \tag{16}$$

Setting $a_K = b_K$ ensures termination at analysis K . The sets of values $\{Z_k; |Z_k| \leq a_k\}$, $k = 1, \dots, K - 1$, form an “inner wedge” within the continuation region $\{Z_k; |Z_k| < b_k\}$ at which early stopping to accept H_0 occurs. If such early stopping is not desired at the first few analyses, this possibility can be precluded by setting $a_1 = 0$, $a_2 = 0$, etc. Early stopping to accept H_0 has the benefit of reducing the expected sample size when θ is close to zero.

We shall focus on the tests introduced by Pampallona & Tsiatis (1994) which we refer to as the power family of two-sided, inner wedge tests. Like the power family of one-sided tests, these tests are indexed by a parameter Δ which occurs as a power in formulae for the boundary points a_k and b_k , $k = 1, \dots, K$. Specifically

$$\begin{aligned}
b_k &= \tilde{C}_{W1}(K, \alpha, \beta, \Delta)(k/K)^{\Delta-1/2} \quad \text{and} \\
a_k &= \delta\sqrt{\mathcal{I}_k} - \tilde{C}_{W2}(K, \alpha, \beta, \Delta)(k/K)^{\Delta-1/2}, \quad k = 1, \dots, K.
\end{aligned} \tag{17}$$

A final information level

$$\mathcal{I}_K = \frac{\{\tilde{C}_{W1}(K, \alpha, \beta, \Delta) + \tilde{C}_{W2}(K, \alpha, \beta, \Delta)\}^2}{\delta^2}$$

ensures that $a_K = b_K$ and the test terminates properly.

These tests are designed under the assumption of equally spaced information levels, $\mathcal{I}_k = (k/K)\mathcal{I}_K$, $k = 1, \dots, K$, and so we can re-express the above formulae as

$$\begin{aligned}
b_k &= \tilde{C}_{W1}(K, \alpha, \beta, \Delta)(k/K)^{\Delta-1/2} \quad \text{and} \\
a_k &= \{\tilde{C}_{W1}(K, \alpha, \beta, \Delta) + \tilde{C}_{W2}(K, \alpha, \beta, \Delta)\}(k/K)^{1/2} \\
&\quad - \tilde{C}_{W2}(K, \alpha, \beta, \Delta)(k/K)^{\Delta-1/2}, \quad k = 1, \dots, K,
\end{aligned} \tag{18}$$

avoiding explicit mention of $\mathcal{I}_k$.

For illustration, we consider tests with $K = 3, 5$ and 10 analyses, and shape parameter $\Delta = 0$ and 0.25, designed to achieve Type I error rate $\alpha = 0.05$ and power $1 - \beta = 0.9$ at $\theta = \pm\delta$. Table 7 shows the constants

Table 7: Constants $\tilde{C}_{W1}(K, \alpha, \beta, \Delta)$, $\tilde{C}_{W2}(K, \alpha, \beta, \Delta)$ and $\tilde{R}_W(K, \alpha, \beta, \Delta)$ defining power family two-sided, inner wedge tests with shape parameter Δ . Tests are for K groups of observations and have Type I error probability $\alpha = 0.05$ and power $1 - \beta = 0.9$ at $\theta = \pm\delta$. The first opportunity for early stopping to accept H_0 occurs at analysis k^* .

K	$\Delta = 0$				$\Delta = 0.25$			
	$\tilde{C}_{W1}$	$\tilde{C}_{W2}$	$\tilde{R}_W$	k^*	$\tilde{C}_{W1}$	$\tilde{C}_{W2}$	$\tilde{R}_W$	k^*
3	1.971	1.353	1.051	2	2.037	1.422	1.139	1
5	1.990	1.385	1.084	3	2.073	1.477	1.199	2
10	2.013	1.428	1.127	5	2.119	1.521	1.261	4

$\tilde{C}_{W1}(K, \alpha, \beta, \Delta)$ and $\tilde{C}_{W2}(K, \alpha, \beta, \Delta)$ defining these tests. For each test, the value k^* is the index of the first analysis k for which $a_k > 0$ and, thus, early stopping to accept H_0 is possible. The table also shows values of

$$\tilde{R}_W(K, \alpha, \beta, \Delta) = \frac{\mathcal{I}_K}{\mathcal{I}_{f,2}} = \frac{\{\tilde{C}_{W1}(K, \alpha, \beta, \Delta) + \tilde{C}_{W2}(K, \alpha, \beta, \Delta)\}^2}{\{\Phi^{-1}(1 - \alpha/2) + \Phi^{-1}(1 - \beta)\}^2},$$

the ratio of the maximum information level to that of the corresponding fixed sample test. Constants for other tests can be found in Pampallona & Tsiatis (1994) but, since these authors have expressed tests in terms of the score statistics $S_k = Z_k \sqrt{\mathcal{I}_k}$ and we have worked with the standardised statistics Z_k , $k = 1, \dots, K$, our $\tilde{C}_{Wj}(K, \alpha, \beta, \Delta)$ are $K^{\Delta-1/2}$ times their $C_j(K, \alpha, \beta, \Delta)$, $j = 1$ and 2 .

If the intended information levels $\mathcal{I}_k = (k/K) \tilde{R}_W(K, \alpha, \beta, \Delta) \mathcal{I}_{f,2}$, $k = 1, \dots, K$, arise as planned, Type I error and power conditions will be met exactly. Suppose instead that observed information levels differ from their planned values and our priority is to maintain the Type I error rate close to its nominal value α . In following the significance level approach, we apply rule (16), comparing the standardised statistics Z_k to the original sequence of critical values. In doing this, we obtain values of a_k and b_k from (18) directly in terms of $\tilde{C}_{W1}(K, \alpha, \beta, \Delta)$ and $\tilde{C}_{W2}(K, \alpha, \beta, \Delta)$. Since $\{Z_1, \dots, Z_K\}$ follow the canonical distribution (1), their joint distribution under H_0 depends on the observed information levels $\{\mathcal{I}_1, \dots, \mathcal{I}_K\}$ only through their ratios and, hence, the Type I error rate is exactly α under any sequence of equally spaced information levels. Perturbations in the attained Type I error rate arise if information levels are unequally spaced and power at $\theta = \pm\delta$ is affected both by the spacing and the overall scale of observed information levels.

Table 8 shows achieved Type I error probabilities and power under sequences of information levels $\mathcal{I}_k = \pi(k/K)^r \mathcal{I}_{max}$, $k = 1, \dots, K$, when

Table 8: Properties of power family two-sided, inner wedge tests designed for information levels $\mathcal{I}_k = (k/K)\mathcal{I}_{max}$ and applied with $\mathcal{I}_k = \pi(k/K)^r\mathcal{I}_{max}$, $k = 1, \dots, K$, in a manner which aims to preserve the intended Type I error rate. The table shows Type I error probabilities and power attained by tests with shape parameter Δ and K analyses, designed to achieve Type I error probability $\alpha = 0.05$ and power $1 - \beta = 0.9$ at $\theta = \pm\delta$.

			$\Delta = 0$		$\Delta = 0.25$	
	r	π	Type I error	Power	Type I error	Power
$K = 3$						
	0.70	0.9	0.051	0.878	0.050	0.888
		1.0	0.051	0.909	0.050	0.917
		1.1	0.051	0.932	0.050	0.940
	1.00	0.9	0.050	0.868	0.050	0.868
		1.0	0.050	0.900	0.050	0.900
		1.1	0.050	0.925	0.050	0.924
	1.30	0.9	0.049	0.853	0.049	0.840
		1.0	0.049	0.886	0.049	0.874
		1.1	0.049	0.913	0.049	0.901
$K = 5$						
	0.70	0.9	0.050	0.882	0.049	0.895
		1.0	0.050	0.912	0.049	0.923
		1.1	0.050	0.935	0.049	0.944
	1.00	0.9	0.050	0.868	0.050	0.868
		1.0	0.050	0.900	0.050	0.900
		1.1	0.050	0.925	0.050	0.924
	1.30	0.9	0.049	0.847	0.050	0.830
		1.0	0.049	0.881	0.050	0.865
		1.1	0.049	0.908	0.050	0.893
$K = 10$						
	0.70	0.9	0.050	0.887	0.048	0.899
		1.0	0.050	0.917	0.048	0.927
		1.1	0.050	0.939	0.048	0.947
	1.00	0.9	0.050	0.868	0.050	0.868
		1.0	0.050	0.900	0.050	0.900
		1.1	0.050	0.925	0.050	0.924
	1.30	0.9	0.049	0.840	0.050	0.823
		1.0	0.049	0.874	0.050	0.859
		1.1	0.049	0.902	0.050	0.888

equally spaced values up to a maximum $\mathcal{I}_{max} = \tilde{R}_W(K, \alpha, \beta, \Delta) \mathcal{I}_{f,2}$ were planned. We see from the table that the Type I error rate remains within 0.002 of $\alpha = 0.05$ in all cases whereas power is more variable and is affected by the values of both π and r .

The inner wedge tests can also be adapted to handle normal observations with an unknown variance. Suppose data follow a normal linear model and we wish to test $H_0: \theta = 0$ against a two-sided alternative with Type I error probability close to α . As before, we suppose the sequence of estimates $\{\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(K)}\}$ is multivariate normal with $Cov(\hat{\theta}^{(k_1)}, \hat{\theta}^{(k_2)}) = Var(\hat{\theta}^{(k_2)})$ for $k_1 \leq k_2$, we write $Var(\hat{\theta}^{(k)}) = V_k \sigma^2$, $k = 1, \dots, K$, where the V_k are known and σ^2 is unknown, and denote the usual estimates of σ^2 on ν_k degrees of freedom by s_k^2 , $k = 1, \dots, K$. The t -statistic for testing H_0 at analysis k is

$$T_k = \frac{\hat{\theta}^{(k)}}{\sqrt{\{V_k s_k^2\}}}$$

and we adapt a known-variance test of the form (16) by replacing each Z_k by T_k and setting critical values at the same percentiles of the null distribution of T_k as they were originally in the null distribution of Z_k . The resulting rule is:

After group $k = 1, \dots, K - 1$
 if $|T_k| \geq t_{\nu_k, 1-\Phi(b_k)}$ stop, reject H_0
 if $|T_k| < t_{\nu_k, 1-\Phi(a_k)}$ stop, accept H_0
 otherwise continue to group $k + 1$,
 after group K
 if $|T_K| \geq t_{\nu_K, 1-\Phi(b_K)}$ stop, reject H_0
 if $|T_K| < t_{\nu_K, 1-\Phi(a_K)}$ stop, accept H_0 .

Here the a_k and b_k are as defined in (18). Note that $t_{\nu_k, 1-\Phi(a_k)}$ is positive if and only if $a_k > 0$ and so early stopping to accept H_0 is possible at analyses $k \geq k^*$, just as before.

The arguments originally used in Section 2 imply that these group sequential t -tests have Type I error rate close to α if information levels are evenly spaced and degrees of freedom ν_k are not too low. We can also follow a similar approach to estimate a test's power, obtaining the approximation

$$Pr\{T_{\nu_K}(\delta/\sqrt{\{\tilde{R}_W(K, \alpha, \beta, \Delta) V_K \sigma^2\}}) \geq t_{\nu_K, 1-\alpha/2}\} \quad (19)$$

as the analogue of (4). The cruder approximation to the power at $\theta = \pm\delta$ suitable for large values of ν_K becomes

$$\Phi\left(\frac{\delta}{\sqrt{\{\tilde{R}_W(K, \alpha, \beta, \Delta) V_K \sigma^2\}}} - \Phi^{-1}(1 - \alpha/2)\right). \quad (20)$$

Table 9 shows the Type I error probabilities of these group sequential inner wedge t -tests and the power they attain at $\theta = \pm\delta$ when (19) gives the

Table 9: Properties of two-sided, inner wedge t -tests designed to test $H_0: \theta = 0$ with Type I error rate $\alpha = 0.05$. Tests are based on power family tests for known variance with $\Delta = 0$ and 0.25. Degrees of freedom for σ^2 at analysis k are $\nu_k = (\nu_K + 2)(k/K) - 2$ and $Var(\hat{\theta}^{(k)}) = (K/k)Var(\hat{\theta}^{(K)})$, $k = 1, \dots, K$. The penultimate column shows actual power at the alternative for which the approximation (19) gives power 0.9 and the final column shows the power predicted for each test by formula (20).

Type I error and power were estimated by simulation using 50,000 replicates; standard errors are 0.001 for Type I error and 0.0015 for attained power.

$\Delta = 0$

K	ν_K	Type I error	Attained power	Known σ^2 power
3	7	0.057	0.914	0.967
3	10	0.057	0.908	0.950
3	13	0.054	0.909	0.940
5	13	0.056	0.911	0.940
5	18	0.055	0.908	0.929
5	23	0.055	0.904	0.923
10	28	0.057	0.907	0.919
10	38	0.055	0.905	0.914
10	48	0.054	0.905	0.911

$\Delta = 0.25$

K	ν_K	Type I error	Attained power	Known σ^2 power
3	7	0.060	0.912	0.967
3	10	0.056	0.919	0.950
3	13	0.057	0.906	0.940
5	13	0.059	0.912	0.940
5	18	0.055	0.907	0.929
5	23	0.056	0.909	0.923
10	28	0.057	0.908	0.919
10	38	0.055	0.906	0.914
10	48	0.054	0.904	0.911

value 0.9 as the approximate power. The degrees of freedom are those which arise when each group contains 3, 4 or 5 observations and two degrees of freedom are used in fitting the linear model at each analysis. Results follow the previous pattern for group sequential t -tests with only minor departures from the target of $\alpha = 0.05$ for the Type I error rate and formula (19) giving a reliable approximation to attained power when the degrees of freedom are fairly high. Once again, the definition of group sequential t -tests does not rely on equally spaced information levels, and so these tests may also be used with unequal group sizes or information increments.

5 Equivalence tests

The objective in equivalence testing is to demonstrate that two treatments produce essentially the same effect. Thus, if the parameter θ represents a treatment difference, one would like to show that θ is within a small distance of zero. Suppose that δ is the largest tolerable treatment difference, so it is acceptable to declare equivalence if $-\delta < \theta < \delta$ but not otherwise. Then we require a test with a high probability of declaring the two treatments to be equivalent if θ is close to zero and a high probability of *rejecting* equivalence if $\theta \leq -\delta$ or $\theta \geq \delta$.

A two-sided test of $H_0: \theta = 0$ can be converted into an equivalence test if we equate acceptance of H_0 with a declaration of equivalence and rejection of H_0 with a declaration of non-equivalence. The original test's Type I error rate becomes the probability of failing to declare equivalence when $\theta = 0$ and the treatments actually have identical effects. If the test's power at $\theta = \pm\delta$ is $1 - \beta$, then β becomes the probability of wrongly declaring equivalence when the treatment difference is just outside the allowable range. An incorrect conclusion of equivalence could lead to an unsuitable treatment being deemed equally effective as an existing standard and so it is this error rate that should be protected most vigorously. Thus, if one of the power family of inner wedge tests of Section 4 is to be used as a group sequential equivalence test, any adaptation to unequal group sizes or unknown variance should be chosen to maintain this test's power of $1 - \beta$ at $\theta = \pm\delta$ as closely as possible.

Consider once again the power family, inner wedge tests designed to achieve Type I error rate $\alpha = 0.05$ and power $1 - \beta = 0.9$ at $\theta = \pm\delta$, defined by the constants $\tilde{C}_{W1}(K, \alpha, \beta, \Delta)$ and $\tilde{C}_{W2}(K, \alpha, \beta, \Delta)$ in Table 7. One might well prefer to create an equivalence test from a two-sided test with a lower Type II error rate than $\beta = 0.1$ since this represents the key error rate in the equivalence problem. However, we shall carry on using these tests as illustrative examples for the sake of continuity. Suppose a group sequential study follows one of these designs with planned information levels

$$\mathcal{I}_k = (k/K) \tilde{R}_W(K, \alpha, \beta, \Delta) \mathcal{I}_{f,2}, \quad k = 1, \dots, K,$$

where $\mathcal{I}_{f,2}$ is as defined in (2). If observed information levels follow this pattern exactly, the two-sided test will have Type I error probability exactly $\alpha = 0.05$ and power exactly $1 - \beta = 0.9$ at $\theta = \pm\delta$ and the related equivalence test will have the corresponding error probabilities.

If observed information levels differ from their planned values, we are faced with the problem of adapting the group sequential two-sided test in a way that maintains power at $\theta = \pm\delta$ close to $1 - \beta = 0.9$ in order to preserve a probability $\beta = 0.1$ of wrongly accepting equivalence when $\theta = \pm\delta$. Consider first how we should use the significance level approach to maintain properties under $\theta = \delta$. We need to keep boundary points at fixed percentiles of the distribution of Z_k under $\theta = \delta$ and, since $Z_k - \delta\sqrt{\mathcal{I}_k}$ has a standard normal distribution if $\theta = \delta$, this objective can be achieved by expressing the boundary as a sequence of critical values for $Z_k - \delta\sqrt{\mathcal{I}_k}$ in a form which does not explicitly involve the information levels $\mathcal{I}_1, \dots, \mathcal{I}_K$. However, we must also simultaneously maintain properties under $\theta = -\delta$. Fortunately, for typical tests there is little probability of crossing the lower arms of the stopping boundary if $\theta = \delta$ or crossing the upper arms if $\theta = -\delta$ and so it suffices to define the upper arms by considering $\theta = \delta$ and the lower arms by considering $\theta = -\delta$.

For the case $\theta = \delta$, the two upper arms of the stopping boundary defined by (16) with a_k and b_k as specified in (17) are based on the comparison of

$$Z_k \text{ vs } \begin{cases} \tilde{C}_{W1}(K, \alpha, \beta, \Delta)(k/K)^{\Delta-1/2} & \text{and} \\ \delta\sqrt{\mathcal{I}_k} - \tilde{C}_{W2}(K, \alpha, \beta, \Delta)(k/K)^{\Delta-1/2}, \end{cases} \quad k = 1, \dots, K,$$

i.e., of

$$Z_k - \delta\sqrt{\mathcal{I}_k} \text{ vs } \begin{cases} -\delta\sqrt{\mathcal{I}_k} + \tilde{C}_{W1}(K, \alpha, \beta, \Delta)(k/K)^{\Delta-1/2} & \text{and} \\ -\tilde{C}_{W2}(K, \alpha, \beta, \Delta)(k/K)^{\Delta-1/2}, \end{cases} \quad k = 1, \dots, K.$$

Under the planned information levels $\mathcal{I}_k = (k/K)\{\tilde{C}_{W1}(K, \alpha, \beta, \Delta) + \tilde{C}_{W2}(K, \alpha, \beta, \Delta)\}^2/\delta^2$, $k = 1, \dots, K$, this comparison can be written as

$$Z_k - \delta\sqrt{\mathcal{I}_k} \text{ vs } \begin{cases} h_k & \text{and} \\ g_k, \end{cases} \quad k = 1, \dots, K,$$

where

$$\begin{aligned} h_k &= -\{\tilde{C}_{W1}(K, \alpha, \beta, \Delta) + \tilde{C}_{W2}(K, \alpha, \beta, \Delta)\}(k/K)^{1/2} \\ &\quad + \tilde{C}_{W1}(K, \alpha, \beta, \Delta)(k/K)^{\Delta-1/2} \quad \text{and} \\ g_k &= -\tilde{C}_{W2}(K, \alpha, \beta, \Delta)(k/K)^{\Delta-1/2}, \quad k = 1, \dots, K. \end{aligned} \tag{21}$$

Note that these formulae for g_k and h_k avoid explicit mention of the $\mathcal{I}_k$. Also, g_K is automatically equal to h_K . By symmetry, the two lower arms of the stopping boundary arise from comparing $Z_k + \delta\sqrt{\mathcal{I}_k}$ with $-g_k$ and

$-h_k$, $k = 1, \dots, K$. Adding the further feature that early stopping to accept H_0 should not occur before analysis k^* , the first occasion at which a_k given by (18) is greater than zero, we obtain the full stopping rule:

$$\begin{aligned}
&\text{After group } k = 1, \dots, k^* - 1 \\
&\quad \text{if } Z_k - \delta\sqrt{\mathcal{I}_k} \geq h_k \text{ or } Z_k + \delta\sqrt{\mathcal{I}_k} \leq -h_k \quad \text{stop, reject } H_0 \\
&\quad \text{otherwise} \quad \text{continue to group } k+1, \\
&\text{after group } k = k^*, \dots, K - 1 \\
&\quad \text{if } Z_k - \delta\sqrt{\mathcal{I}_k} \geq h_k \text{ or } Z_k + \delta\sqrt{\mathcal{I}_k} \leq -h_k \quad \text{stop, reject } H_0 \\
&\quad \text{if } Z_k - \delta\sqrt{\mathcal{I}_k} < g_k \text{ and } Z_k + \delta\sqrt{\mathcal{I}_k} > -g_k \quad \text{stop, accept } H_0 \\
&\quad \text{otherwise} \quad \text{continue to group } k+1, \\
&\text{after group } K \\
&\quad \text{if } Z_K - \delta\sqrt{\mathcal{I}_K} \geq h_K \text{ or } Z_K + \delta\sqrt{\mathcal{I}_K} \leq -h_K \quad \text{stop, reject } H_0 \\
&\quad \text{if } Z_K - \delta\sqrt{\mathcal{I}_K} < g_K \text{ and } Z_K + \delta\sqrt{\mathcal{I}_K} > -g_K \quad \text{stop, accept } H_0,
\end{aligned} \tag{22}$$

where g_k and h_k are as defined in (21). It is straightforward to check that the condition to stop and reject H_0 at analysis k simplifies to

$$\begin{aligned}
|Z_k| &\geq h_k + \delta\sqrt{\mathcal{I}_k} = \tilde{C}_{W1}(K, \alpha, \beta, \Delta)(k/K)^{\Delta-1/2} + \delta\sqrt{\mathcal{I}_k} \\
&\quad - \{\tilde{C}_{W1}(K, \alpha, \beta, \Delta) + \tilde{C}_{W2}(K, \alpha, \beta, \Delta)\}(k/K)^{1/2}
\end{aligned}$$

and the condition to stop and accept H_0 reduces to

$$|Z_k| < g_k + \delta\sqrt{\mathcal{I}_k} = \delta\sqrt{\mathcal{I}_k} - \tilde{C}_{W2}(K, \alpha, \beta, \Delta)(k/K)^{\Delta-1/2}.$$

To apply this adaptation of a power family, inner wedge test one simply follows rule (22), substituting in the observed information levels $\mathcal{I}_1, \dots, \mathcal{I}_K$. Table 10 shows the Type I error probability and power achieved by this adaptation of tests with $K = 3, 5$ and 10 analyses and shape parameter $\Delta = 0$ and 0.25 under sequences of information levels of the form $\mathcal{I}_k = \pi(k/K)^r \mathcal{I}_{\max}$, where $\mathcal{I}_{\max} = \tilde{R}_W(K, \alpha, \beta, \Delta) \mathcal{I}_{f,2}$. We see that, as intended, power remains close to its nominal value, 0.9. The two-sided tests originally planned here are the same as those considered in Section 4 and the observed information sequences are identical to those occurring in the examples of Table 8. However, we have worked towards a different objective in adapting the tests to unequal group sizes and the resulting tests have quite different properties. Whereas the tests in Table 8 showed variability in their power at $\theta = \pm\delta$, it is the attained Type I error rate of our new tests, representing the probability of failing to declare equivalence when $\theta = 0$, which is particularly variable.

It remains to consider adaptation of power family, inner wedge tests to the case of unknown variance in a way which maintains power at $\theta = \pm\delta$ as closely as possible. As before, we assume a multivariate normal sequence of estimates $\{\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(K)}\}$ with $\text{Cov}(\hat{\theta}^{(k_1)}, \hat{\theta}^{(k_2)}) = \text{Var}(\hat{\theta}^{(k_2)})$ for $k_1 \leq k_2$, and $\text{Var}(\hat{\theta}^{(k)}) = V_k \sigma^2$, $k = 1, \dots, K$, where the V_k are known and σ^2 is

Table 10: Properties of power family two-sided, inner wedge tests designed for information levels $\mathcal{I}_k = (k/K)\mathcal{I}_{max}$ and applied with $\mathcal{I}_k = \pi(k/K)^r\mathcal{I}_{max}$, $k = 1, \dots, K$, in a manner which aims to preserve the power at $\theta = \pm\delta$. The table shows Type I error probabilities and power attained by tests with shape parameter Δ and K analyses, designed to achieve Type I error $\alpha = 0.05$ and power $1 - \beta = 0.9$ at $\theta = \pm\delta$.

		$\Delta = 0$		$\Delta = 0.25$	
r	π	Type I error	Power	Type I error	Power
$K = 3$					
0.70	0.9	0.069	0.902	0.060	0.899
	1.0	0.047	0.902	0.040	0.899
	1.1	0.032	0.902	0.026	0.898
1.00	0.9	0.073	0.900	0.073	0.905
	1.0	0.050	0.900	0.050	0.900
	1.1	0.034	0.900	0.034	0.897
1.30	0.9	0.081	0.899	0.095	0.907
	1.0	0.057	0.899	0.069	0.907
	1.1	0.040	0.899	0.050	0.907
$K = 5$					
0.70	0.9	0.066	0.903	0.057	0.904
	1.0	0.044	0.903	0.037	0.904
	1.1	0.030	0.903	0.025	0.904
1.00	0.9	0.073	0.900	0.073	0.901
	1.0	0.050	0.900	0.050	0.900
	1.1	0.034	0.900	0.034	0.900
1.30	0.9	0.085	0.899	0.103	0.909
	1.0	0.060	0.899	0.074	0.903
	1.1	0.042	0.899	0.054	0.900
$K = 10$					
0.70	0.9	0.063	0.904	0.053	0.905
	1.0	0.042	0.904	0.035	0.905
	1.1	0.028	0.904	0.023	0.905
1.00	0.9	0.073	0.900	0.072	0.900
	1.0	0.050	0.900	0.050	0.900
	1.1	0.034	0.900	0.035	0.900
1.30	0.9	0.090	0.900	0.109	0.904
	1.0	0.064	0.898	0.080	0.900
	1.1	0.045	0.898	0.059	0.898

unknown. We denote by s_k^2 the usual estimate of σ^2 on ν_k degrees of freedom at analysis k . In the known variance problem, we used the z -statistics $Z_k - \delta\sqrt{\mathcal{I}_k}$ and $Z_k + \delta\sqrt{\mathcal{I}_k}$ which had standard normal distributions under the hypotheses $\theta = \delta$ and $\theta = -\delta$, respectively. With σ^2 unknown, we now use the t -statistics

$$T_k^1 = \frac{\hat{\theta}^{(k)} - \delta}{\sqrt{\{V_k s_k^2\}}} \quad \text{and} \quad T_k^2 = \frac{\hat{\theta}^{(k)} + \delta}{\sqrt{\{V_k s_k^2\}}}$$

which have t_{ν_k} distributions under the same two hypotheses. To create a group sequential t -test, we replace $Z_k - \delta\sqrt{\mathcal{I}_k}$ by T_k^1 and $Z_k + \delta\sqrt{\mathcal{I}_k}$ by T_k^2 in rule (22) and convert the g_k and h_k into percentiles of the t_{ν_k} distribution rather than the standard normal distribution. The test is then:

$$\begin{aligned} &\text{After group } k = 1, \dots, k^* - 1 \\ &\quad \text{if } T_k^1 \geq t_{\nu_k, 1-\Phi(h_k)} \text{ or } T_k^2 \leq -t_{\nu_k, 1-\Phi(h_k)} \quad \text{stop, reject } H_0 \\ &\quad \text{otherwise} \quad \text{continue to group } k+1, \\ &\text{after group } k = k^*, \dots, K - 1 \\ &\quad \text{if } T_k^1 \geq t_{\nu_k, 1-\Phi(h_k)} \text{ or } T_k^2 \leq -t_{\nu_k, 1-\Phi(h_k)} \quad \text{stop, reject } H_0 \\ &\quad \text{if } T_k^1 < t_{\nu_k, 1-\Phi(g_k)} \text{ and } T_k^2 > -t_{\nu_k, 1-\Phi(g_k)} \quad \text{stop, accept } H_0 \\ &\quad \text{otherwise} \quad \text{continue to group } k+1, \\ &\text{after group } K \\ &\quad \text{if } T_K^1 \geq t_{\nu_K, 1-\Phi(h_K)} \text{ or } T_K^2 \leq -t_{\nu_K, 1-\Phi(h_K)} \quad \text{stop, reject } H_0 \\ &\quad \text{if } T_K^1 < t_{\nu_K, 1-\Phi(g_K)} \text{ and } T_K^2 > -t_{\nu_K, 1-\Phi(g_K)} \quad \text{stop, accept } H_0, \end{aligned} \tag{23}$$

Note that, since $g_K = h_K$, termination at analysis K is guaranteed.

This construction addresses power at $\theta = \pm\delta$ and the Type I error rate is left to depend on the unknown value of σ^2 . The following approximation to the Type I error probability under a given value of σ^2 provides a means to check that this error rate is acceptably low for values of σ^2 in a plausible range. Consider a fixed sample test based on estimates $\hat{\theta} \sim N(\theta, V\sigma^2)$ and $s^2 \sim \sigma^2 \chi_\nu^2 / \nu$ which achieves power $1 - \beta$ at $\theta = \pm\delta$ by rejecting $\theta = 0$ if

$$\frac{\hat{\theta} - \delta}{\sqrt{\{Vs^2\}}} \geq -t_{\nu, \beta} \quad \text{or} \quad \frac{\hat{\theta} + \delta}{\sqrt{\{Vs^2\}}} \leq t_{\nu, \beta}.$$

The Type I error rate is the probability that

$$|\hat{\theta}| \geq \delta - t_{\nu, \beta} \sqrt{\{Vs^2\}}$$

if $\hat{\theta} \sim N(0, V\sigma^2)$, which simplifies to

$$2 \times \Pr\{T_\nu(\delta/\sqrt{\{V\sigma^2\}}) \leq t_{\nu, \beta}\}.$$

Setting $\nu = \nu_K$ in this expression and, by analogy with the case of known variance, replacing V by $R(K, \alpha, \beta, \Delta)$ times V_K , we obtain the approximation

$$2 \times \Pr\{T_{\nu_K}(\delta/\sqrt{\{R(K, \alpha, \beta, \Delta) V_K \sigma^2\}}) \leq t_{\nu_K, \beta}\} \tag{24}$$

to the group sequential t -test's Type I error probability.

Table 11 shows the power at $\theta = \pm\delta$ of group sequential, inner wedge t -tests and attained Type I error probability when the approximation (24) gives a value of 0.05. As before, the degrees of freedom are those which arise when groups contain 3, 4 or 5 observations and two degrees of freedom are used in fitting the linear model at each analysis. The table shows that both power and Type I error rate agree well with their predicted values.

The definition of these group sequential t -tests does not rely on equally spaced information levels, and the tests may also be used with unequal group sizes or information increments. We have carried out further simulations of such t -tests with different final information levels and information accruing unequally between analyses. The results of these simulations followed the pattern seen in Table 10 when information levels were perturbed in the case of known variance: power at $\theta = \pm\delta$ stayed close to its target of $1 - \beta = 0.9$ and Type I error probability varied noticeably with the final information level.

6 Discussion

We have presented a framework for generalising the application area of group sequential tests designed for equally sized groups of observations of known variance. This approach is not limited to group sequential tests with particular types of boundary and the tests we have presented should be regarded as illustrative examples. A natural extension in the case of two-sided tests with no inner wedge is to the family of tests proposed by Wang & Tsiatis (1987), which includes the Pocock and O'Brien & Fleming tests as special cases.

Just as we have used the significance level approach to construct group sequential t -tests, the same method can also be applied to other response distributions. To do this, one interprets boundary points of the original tests as significance levels of normally distributed test statistics and then conducts tests appropriate to the observed data's response distribution using the same sequence of significance levels. This approach facilitates the use of accurate small sample approximations to test statistics' distributions and it has proved effective in a number of instances including sequential χ^2 tests (Jennison & Turnbull, 1991) and sequential analysis of a hazard ratio in censored survival data (Jennison, 1992).

The example of equivalence testing in Section 5 demonstrates how one can choose to give priority to the error probability at a particular value of the parameter of interest, θ^* say, by interpreting all boundary points in terms of a significance test of the hypothesis $\theta = \theta^*$. This is particularly useful when it is difficult to arrange data collection to achieve pre-specified information levels and yet it is desirable to control closely one particular error probability.

Table 11: Properties of two-sided, inner wedge t -tests designed to test $H_0: \theta = 0$ with power $1 - \beta = 0.9$ at $\theta = \pm\delta$. Tests are based on power family tests for known variance with $\Delta = 0$ and 0.25. Degrees of freedom for σ^2 at analysis k are $\nu_k = (\nu_K + 2)(k/K) - 2$ and $Var(\hat{\theta}^{(k)}) = (K/k)Var(\hat{\theta}^{(K)})$, $k = 1, \dots, K$, where $Var(\hat{\theta}^{(K)})$ is such that the approximation (24) gives Type I error 0.05 at $\theta = 0$ when power at $\theta = \pm\delta$ is 0.9.

Type I error and power were estimated by simulation using 50,000 replicates; standard errors are 0.001 for Type I error and 0.0015 for attained power.

$\Delta = 0$

K	ν_K	Type I error	Power at $\theta = \pm\delta$
3	7	0.052	0.896
3	10	0.049	0.895
3	13	0.050	0.898
5	13	0.048	0.893
5	18	0.049	0.895
5	23	0.047	0.895
10	28	0.049	0.895
10	38	0.052	0.898
10	48	0.047	0.898

$\Delta = 0.25$

K	ν_K	Type I error	Power at $\theta = \pm\delta$
3	7	0.049	0.880
3	10	0.048	0.888
3	13	0.048	0.889
5	13	0.049	0.893
5	18	0.051	0.893
5	23	0.050	0.899
10	28	0.049	0.897
10	38	0.050	0.898
10	48	0.049	0.896

7 References

- Emerson, S. S. & Fleming, T. R. (1989) Symmetric group sequential designs. *Biometrics*, **45**, 905–923.
- Jennison, C. (1992) Bootstrap tests and confidence intervals for a hazard ratio when the number of observed failures is small, with applications to group sequential survival studies. In *Computing Science and Statistics, Vol. 23*. Eds. C. Page and R. LePage, Springer-Verlag, New York, p.89–97.
- Jennison, C. and Turnbull, B. W. (1989) Interim analyses: the repeated confidence interval approach (with discussion). *J. Roy. Statist. Soc. B*, **51** 305–361.
- Jennison, C. and Turnbull, B. W. (1991) Exact calculations for sequential t , χ^2 and F tests. *Biometrika*, **78**, 133–141.
- Jennison, C. & Turnbull, B. W. (1997a) Group sequential analysis incorporating covariate information. *J. Amer. Statist. Assoc.*, **92**, 1330–1341.
- Jennison, C. & Turnbull, B. W. (1997b) Distribution theory for group sequential t , χ^2 and F tests for general linear models. *Sequential Analysis*, **16**, 295–317.
- Lan, K. K. G. & DeMets, D. L. (1983) Discrete sequential boundaries for clinical trials. *Biometrika*, **70**, 659–663.
- Lan, K. K. G. & Zucker, D. M. (1993). Sequential monitoring of clinical trials. *Statistics in Medicine*, **12**, 753–766.
- O’Brien, P. C. & Fleming, T. R. (1979) A multiple testing procedure for clinical trials. *Biometrics*, **35**, 549–556.
- Pampallona, S. & Tsiatis, A. A. (1994) Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *J. Statist. Planning and Inference* **42**, 19–35.
- Pocock, S. J. (1977) Group sequential methods in the design and analysis of clinical trials. *Biometrika*, **64**, 191–199.
- Scharfstein, D.O., Tsiatis, A.A. and Robins, J.M. (1997). Semiparametric efficiency and its implication on the design and analysis of group sequential studies. *J. Amer. Statist. Assoc.*, **92**, 1342–1350.
- Wang, S. K. & Tsiatis, A. A. (1987) Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics*, **43**, 193–200.