S+SeqTrial:
Technical Overview

Scott S. Emerson, M.D., Ph.D.

7055 54th Avenue NE

Seattle, Washington 98115

September 6, 2007

Contents

0 Conventions

This document presents a technical overview of the methods implemented in the C and S-Plus code that comprises the module S+SeqTrial. In this document the following conventions are used:

1. Parentheses () are used to denote arguments to a function, elements of a vector, or endpoints of an open interval; square brackets [] are used to designate order of arithmetic operations, elements of a matrix, or endpoints of a closed interval; curly brackets { } are used to designate order of arithmetic operations (in alternation with the square brackets) or elements of a set. Hence, $y(t)$ shall mean a function y evaluated at t, while $y[t + u]$ would mean to multiply a variable y by the quantity $t + u$.

2. $X \sim \mathcal{N}(\mu, \sigma^2)$ is used to signify that X is a random variable distributed according to a normal distribution having mean μ and variance σ^2. $\Phi(x)$ denotes the cumulative distribution function for the standard normal distribution.

3. $Pr(A)$ denotes the probability of event A; $Pr(A|X)$ shall denote the probability of event A after conditioning on the observation of random variable X; $Pr(A;\mu)$ shall denote the probability of event A when the parameter is μ.

4. The letters a, b, c, and d when used as a subscript shall denote a parameter that is in some way related to the corresponding boundary of a group sequential test.

5. The letters S, $\overline{X}$, Z, P, B, C, H, and E when used as a subscript shall denote one of the scales for test statistics. The letter T shall be used to represent any choice of these test statistics.

6. An asterisk, $*$, used as a superscript shall denote a quantity measured under the standardizing transformation. Note that all quantities measured under the standardizing transformation are denoted by appending an asterisk as a superscript to the symbol used on the untransformed scale with the notable exception of the standardized mean, which is denoted by δ.

7. An asterisk, $*$, used as a subscript shall usually indicate a general formula that might apply to several different subscripted parameters. for instance, an asterisk might be used as a subscript to stand for any of the letters a, b, c, or d when it is desired to draw parallels among the formulas for the four boundaries, to stand for any of the statistics when it is desired to draw parallels among the various boundary scales, or to stand for '+', '−', or '0' when it is desired to draw parallels among the various hypotheses.

1 Fundamental model, test statistics, and standardizing transformation

Suppose we potentially have measurements

$$X_i \sim \mathcal{N}(\mu, \sigma^2) \quad i = 1, 2, \ldots, N. \tag{1.1}$$

We assume that all observations are independent. We further assume that σ^2 is known and that μ is an unknown parameter measuring treatment effect and which is to be estimated and/or tested. Let $N_1, N_2, \ldots, N_J$ be sample sizes such that $N_1 > 0, N_j < N_{j+1}$ for $j = 1, \ldots, J-1$ and $N_J = N$. We consider the testing of the null hypothesis

$$H_0 : \mu = \mu_0. \tag{1.2}$$

In the simplest clinical trial setting, X_i represents the measurement of treatment response in the ith sampling unit, and μ^2 reflects the variability of each sampling unit. The unknown parameter μ is the population average treatment response. $N_1, \ldots, N_J$ represents sample sizes at which the data might be statistically analyzed. For notational convenience, we define the group sizes accrued between analyses as $n_1 = N_1$ and $n_j = N_j - N_{j-1}$, for $j = 1, \ldots, J$.

More generally, N_j/σ^2 measures the statistical information accrued at various stages during the study. More general settings are described in section 3, and the issues that arise when estimating σ^2 are discussed in section 12.

1.1 Frequentist Test Statistics

For $j = 1, \ldots, J$, define statistics

$$\begin{aligned}
\text{(partial sum) } S_j &= \sum_{i=1}^{N_j} X_i \\
\text{(sample mean) } \overline{X}_j &= \frac{S_j}{N_j} \\
\text{(normalized statistic) } Z_j &= \sqrt{N_j}\frac{[\overline{X}_j - \mu_0]}{\sigma} \\
\text{(fixed sample P value) } P_j &= 1 - \Phi(Z_j) = 1 - \int_{-\infty}^{Z_j} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du
\end{aligned} \tag{1.3}$$

where $\Phi(x)$ is the cumulative distribution function for the standard normal distribution.

The above statistics should be recognizable as those that would typically be used in hypothesis testing. It should be noted that for our purposes those statistics are essentially equivalent. that is, because N_j, μ_0, and σ^2 are all assumed to be known quantities, and because $\Phi(x)$ is a known function, converting any one of those statistics into another is straightforward.

Note that for a fixed (nonrandom) N_j when the data are not sampled according to a stopping rule, the above statistics would have distributions

$$\begin{aligned}
S_j &\sim \mathcal{N}(N_j\mu, N_j\sigma^2) \\
\overline{X}_j &\sim \mathcal{N}\left(\mu, \frac{\sigma^2}{N_j}\right) \\
Z_j &\sim \mathcal{N}\left(\sqrt{N_j}\frac{[\mu - \mu_0]}{\sigma}, 1\right)
\end{aligned} \tag{1.4}$$

and P_j would be the upper one-sided P value in a fixed sample test of $H_0 : \mu = \mu_0$.

By the independent structure of the observations, the increment $S_{j+1} - S_j$ is independent of S_j. this in turn suggests that for $j = 1, \ldots, J-1$

$$\begin{aligned} \mathrm{cov}(S_j, S_{j+1}) &= N_j \sigma^2 \\ \mathrm{cov}(\overline{X}_j, \overline{X}_{j+1}) &= \frac{\sigma^2}{N_{j+1}} \end{aligned} \tag{1.5}$$

1.2 Bayesian Statistics

In a Bayesian setting, we are interested in statistics based on the posterior distribution of μ, which is based on the observed data and some prespecified prior distribution of the mean parameter. For convenience, we will consider only the conjugate prior distribution. thus, we will assume a prior distribution $\mu \sim \mathcal{N}(\zeta, \tau^2)$. the posterior distribution of μ conditioned on the observations $X_1, \ldots, X_{N_j}$ is then

$$\mu \mid (X_1, \ldots, X_{N_j} \sim \mathcal{N}\left(\frac{N_j \tau^2 \overline{X}_j + \sigma^2 \zeta}{N_j \tau^2 + \sigma^2}, \frac{\sigma^2 \tau^2}{N_j \tau^2 + \sigma^2}\right).$$

Statistics of interest might include the posterior probabilities that the mean mu is greater than the null hypothesis μ_0 or prespecified alternative hypotheses μ_+ and μ_- (see section 2). In general, then, we can define a statistic for the posterior probability that the mean μ is greater than some hypothesized value μ_*. We define statistics

$$\begin{aligned} B_j(\zeta, \tau^2, \mu_*) &= Pr(\mu \geq \mu_* \mid (X_1, \ldots, X_{N_j})) \\ &= 1 - \Phi\left(\frac{\mu_*[N_j \tau^2 + \sigma^2] - N_j \tau^2 \overline{X}_j - \sigma^2 \zeta}{\sigma \tau \sqrt{N_j \tau^2 + \sigma^2}}\right) \end{aligned} \tag{1.6}$$

A special case that is of occasional interest is the noninformative prior corresponding to the limit as $\tau^2 \to \infty$. In this setting, the Bayesian posterior probability reduces to

$$B_j(\zeta, \tau^2 = \infty, \mu_*) = Pr(\mu \geq \mu + * \mid (X_1, \ldots, X_{N_j})) = 1 - \Phi\left(\sqrt{N_j}\frac{\mu_* - \overline{X}_j}{\sigma}\right)$$

which is similar in form (but not interpretation) to the fixed sample P value.

These Bayesian statistics are for our purposes equivalent to the frequentist statistics specified in eqn (1.3), as again conversions among the various statistics involve only known quantities.

1.3 Measures of Futility

In the setting of group sequential trials, it is often of interest to consider various measures of the futility of continuing the study. A common goal of such measures is estimating the probability that the test statistic at the Jth analysis might exceed some threshold, where the calculation of the probability is conditioned on the observation at the jth analysis. In what follows, we consider the use of $\overline{X}_j$ as the test statistic and define $t_{\overline{x}J}$ as the threshold of interest for that test statistic at the Jth analysis. As noted above, these measures of futility could also be specified based on any of the statistics defined in this section, with a suitable transformation of the threshold as defined in section 1.5 and discussed in section 4.3. The value of the conditional probability is independent of which test statistic is used, providing the corresponding transformation of the threshold is used.

using the independence of the individual observations, the conditional distribution of $\overline{X}_J$ given $\overline{X}_j$ is found to be

$$\overline{X}_J \mid \overline{X}_j \sim \mathcal{N}\left(\mu + \frac{N_j}{N_J}[\overline{X}_j - \mu],\ \frac{[N_J - N_j]}{N_J}\frac{\sigma^2}{N_J}\right).$$

Computing probabilities based on the above distribution will not result in a statistic, as the distribution depends on the unknown parameter μ. We can, however, compute the probabilities under hypothesized values for μ. Obvious candidates for such computations might be the null hypothesis μ_0, either of the alternative hypotheses μ_+ or μ_-, or the maximum likelihood estimate $\hat{\mu} = \overline{X}_j$ of μ at the jth analysis. We can then define statistics for a specified threshold $t_{\overline{X}j}$ and specified value of $\mu = \mu_*$:

$$\begin{aligned} C_j(t_{\overline{X}J}, \mu_*) &\equiv Pr(\overline{X}_J > t_{\overline{X}J} \mid \overline{X}_j; \mu = \mu_*) \\ &= 1 - \Phi\left(\frac{N_J[t_{\overline{X}J} - \mu_*] - N_j[\overline{X}_j - \mu_*]}{\sigma\sqrt{N_J - N_j}}\right) \end{aligned} \tag{1.7}$$

Note that when the conditional probabilities are computed using the observed maximum likelihood estimate $\overline{X}_j$ for μ, we obtain

$$\begin{aligned} C_j(t_{\overline{X}J}, \mu_* = \overline{X}_j) &\equiv Pr(\overline{X}_J > t_{\overline{X}J} \mid \overline{X}_j; \mu = \overline{X}_j) \\ &= 1 - \Phi\left(\frac{N_J[t_{\overline{X}J} - \overline{X}_j]}{\sigma\sqrt{N_J - N_j}}\right) \end{aligned} \tag{1.8}$$

An alternative approach is to use a Bayesian prior distribution for μ to compute its posterior distribution based on the observation of $\overline{X}_j$, and then to compute a predictive probability by averaging the conditional probabilities of exceeding the threshold as μ ranges over that posterior distribution. Using this approach with a normal prior distribution $\mu \sim \mathcal{N}(\zeta, \tau^2)$ yields a posterior distribution $\lambda(\mu|\overline{X}_j)$ that is normal as given in section (1.2) above. We then compute the marginal conditional distribution of $\overline{X}_J$ given $\overline{X}_j$ as a normal distribution having mean $\{[N_J\tau^2+\sigma^2]N_j\overline{X}_j + [N_J-N_j]\sigma^2\zeta\}/\{N_j[N_j\tau^2+\sigma^2]\}$ and variance $\sigma^2[N_J-N_j][N_J\tau^2+\sigma^2]/\{N_j^2[N_j\tau^2+\sigma^2]\}$ and survival function

$$\begin{aligned} H_j(t_{\overline{X}J}, \zeta, \tau^2) &\equiv \int Pr(\overline{X}_J > t_{\overline{X}J} \,|\, \overline{X}_j, \mu)\, \lambda(\mu \,|\, \overline{X}_j)\, d\mu \\ &= 1 - \Phi\left(\frac{N_J[N_j\tau^2 + \sigma^2][t_{\overline{X}J} - \overline{X}_j] + \sigma^2[N_J - N_j][\overline{X}_j - \zeta]}{\sigma\sqrt{[N_J - N_j][N_J\tau^2 + \sigma^2][N_j\tau^2 + \sigma^2]}}\right) \end{aligned} \tag{1.9}$$

When we consider a noninformative prior distribution ($\mu \sim \mathcal{N}(\zeta, \tau^2)$ and taking the limit as $\tau^2 \to \infty$, the posterior distribution $\lambda(\mu|\overline{X}_j)$ is normal with mean $\overline{X}_j$ and variance σ^2/N_j. we then compute the marginal conditional distribution of $\overline{X}_J$ given $\overline{X}_j$ as having survival function

$$H_j(t_{\overline{X}J}, \zeta, \tau^2 = \infty) = 1 - \Phi\left(\frac{N_J[t_{\overline{X}J} - \overline{X}_j]}{\sigma\sqrt{\frac{N_J}{N_j}[N_J - N_j]}}\right)$$

1.4 Error Spending Measures

Another set of statistics sometimes used in the group sequential setting are those related to the error spending functions. these statistics are based on the sampling distribution of the group sequential test statistic under various hypotheses. As such, the definition of these statistics makes use

of the general form of stopping rules defined in section 4 and the group sequential density defined in section 5. We use this notation at this point (in advance of its formal introduction) in order to highlight that measures based on error spending function are in fact statistics independent of unknown parameters and that there are 1:1 correspondences between each of the error spending statistics and statistics measured on each of the scales defined above. We do note, however, that a choice needs to be made to define the error spending scale in a manner such that a 1:1 correspondence exists for all possible values of the observed statistics at the interim analyses (the path followed here), or to define the error spending scale in a manner such that a 1:1 correspondence exists only for possible values of the observed statistics at the end of a study (an approach that will term the "error spending function", rather than the "error spending scale"). Further discussion of this distinction will be made in section 6.3.

Consider a setting in which for each of the $j = 1, \ldots, J$ specified sample sizes there are specified constants measured on the partial sum scale $-\infty \leq a_{sj} \leq b_{sj} \leq c_{sj} \leq d_{sj} \leq \infty, a_{sJ} = b_{sJ}$ and $c_{sJ} = d_{sJ}$. Further suppose there are four specified hypotheses μ_a, μ_b, μ_c, and μ_d, with $\mu_b \leq \mu_d \leq \mu_c$ and $\mu_b \leq \mu_a \leq \mu_c$. Let $p(j, s; \mu)$, $f(j, s; \mu)$, and $F(j, s; \mu)$ be defined by eqns (5.2), (5.3), and (5.4), with $C_j \equiv (a_{sj}, b_{sj}] \cup [c_{sj}, d_{sj})$ for $j = 1, \ldots, J$.

We define statistics on the error spending scale as

$$
\begin{aligned}
E_{aj}(\mu_a) &\equiv \frac{1}{\alpha_\ell}\left[\sum_{i=1}^{j-1}\int_{-\infty}^{a_{Si}} p(i,s;\mu_a)\,ds + \int_{-\infty}^{S_j} f(j,s;\mu_a)\,ds\right] = \frac{1}{\alpha_\ell}\left[\sum_{i=1}^{j-1} F(i,a_{Si};\mu_a) + F(j,S_j;\mu_a)\right] \\
E_{bj}(\mu_b) &\equiv \frac{1}{[1-\beta_\ell]}\left[\sum_{i=1}^{j-1}\int_{b_{Si}}^{\infty} p(i,s;\mu_b)\,ds + \int_{S_j}^{\infty} f(j,s;\mu_b)\,ds\right] \\
&= \frac{1}{[1-\beta_\ell]}\left[\sum_{i=1}^{j-1}\int_{b_{Si}}^{\infty} p(i,s;\mu_b)\,ds + F(j,\infty;\mu_b) - F(j,S_j;\mu_b)\right] \\
E_{cj}(\mu_c) &\equiv \frac{1}{[1-\beta_u]}\left[\sum_{i=1}^{j-1}\int_{-\infty}^{c_{Si}} p(i,s;\mu_c)\,ds + \int_{-\infty}^{S_j} f(j,s;\mu_c)\,ds\right] \\
&= \frac{1}{[1-\beta_u]}\left[\sum_{i=1}^{j-1}\int_{-\infty}^{c_{Si}} p(i,s;\mu_c)\,ds + F(j,S_j;\mu_c)\right] \\
E_{dj}(\mu_d) &\equiv \frac{1}{\alpha_u}\left[\sum_{i=1}^{j-1}\int_{d_{Si}}^{\infty} p(i,s;\mu_d)\,ds + \int_{S_j}^{\infty} f(j,s;\mu_d)\,ds\right] \\
&= \frac{1}{\alpha_u}\left\{\sum_{i=1}^{j-1}[F(i,\infty;\mu_d) - F(i,d_{Si};\mu_d)] + F(j,\infty;\mu_d) - F(j,S_j;\mu_d)\right\} \qquad (1.10)
\end{aligned}
$$

where constants $\alpha_\ell, \alpha_u, \beta_\ell$, and β_u are defined by

$$\begin{aligned}
\alpha_\ell &\equiv \sum_{i=1}^{J} \int_{-\infty}^{a_{Si}} p(i,s;\mu_a)\,ds = \sum_{i=1}^{J} F(i,a_{Si};\mu_a) \\
\beta_\ell &\equiv \sum_{i=1}^{J} \int_{-\infty}^{a_{Si}} p(i,s;\mu_b)\,ds = \sum_{i=1}^{J} F(i,a_{Si};\mu_b) \\
\beta_u &\equiv \sum_{i=1}^{J} \int_{d_{Si}}^{\infty} p(i,s;\mu_c)\,ds = \sum_{i=1}^{J} [F(i,\infty;\mu_c) - F(i,d_{Si};\mu_c)] \\
\alpha_u &\equiv \sum_{i=1}^{J} \int_{d_{Si}}^{\infty} p(i,s;\mu_d)\,ds = \sum_{i=1}^{J} [F(i,\infty;\mu_d) - F(i,d_{Si};\mu_d)]
\end{aligned} \tag{1.11}$$

In section 6.3, the above statistics will be related to the error spending functions for a group sequential test having continuation sets for the partial sum statistic S_j defined by the values of the a_{sj}'s, b_{sj}'s, c_{sj}'s, and d_{sj}'s. The values of $\alpha_\ell, \alpha_u, \beta_\ell$, and β_u will relate to the size and power of the group sequential design.

1.5 Transformations Among the Various Scales

As noted above, each of the statistics defined by eqns (1.3), (1.6), (1.7), (1.9), and (1.10) are equivalent in the sense that knowing one of the statistics determines the values of the other statistics precisely. This then suggests that we can express a group sequential setting in terms of any of the above statistics, thereby establishing a scale for the problem. We shall at times abbreviate the scales according to the notation used for that statistic. Hence,

S-scale	partial sum scale
X-scale	sample mean scale
Z-scale	normalized scale
P-scale	fixed sample P value scale
B-scale	Bayesian scale (a function of hypothesized mean)
C-scale	conditional probability scale (a function of threshold and hypothesized mean
H-scale	predictive probability scale (a function of threshold)
E_a-scale	lower type I error spending scale
E_b-scale	lower type II error spending scale
E_c-scale	upper type II error spending scale
E_d-scale	upper type I error spending scale

(1.12)

Critical values on a given scale can be similarly converted to other scales using transformations as follows. In defining these conversions, we shall provide formulas for converting a value on an arbitrary scale to the S-scale and for converting the S-scale to any other scale. We again make use of the definitions of $p(j,s;\mu)$, $f(j,s;\mu)$, and $F(j,s;\mu)$ as defined by eqns (5.2), (5.3), and (5.4) in section 5.1. We note that it is assumed that the threshold $t_{\overline{X}J}$ is measured on the X-scale, and that the boundaries a_{sj}, b_{sj}, c_{sj}, and d_{sj} for $j = 1, \ldots, J$ are measured on the S-scale.

Suppose at the jth analysis, y is a value measured on one of the possible scales. The following table provides conversions for y measured on each of the scales to s measured on the S-scale. Note

that in the cases of the B-scale, C-scale, and H-scale, y is a function of a mean μ_* and/or a threshold $t_{\overline{X}j}$.

$$
\begin{aligned}
S\text{-scale} \quad s &= y \\
X\text{-scale} \quad s &= N_j y \\
Z\text{-scale} \quad s &= \sqrt{N_j}\sigma y + N_j\mu_0 \\
P\text{-scale} \quad s &= \sqrt{N_j}\sigma\Phi^{-1}(1-y) + N_j\mu_0 \\
B\text{-scale} \quad s &= \frac{\mu_*[N_j\tau^2+\sigma^2] - \sigma^2\zeta - \sigma\tau\sqrt{N_j\tau^2+\sigma^2}\Phi^{-1}\left(1-y(\zeta,\tau^2,\mu_*)\right)}{\tau^2} \\
\text{noninf } B\text{-scale} \quad s &= \mu_* N_j - \sigma\sqrt{N_j}\Phi^{-1}\left(1-y(\zeta,\tau^2=\infty,\mu_*)\right) \\
C\text{-scale} \quad s &= N_J t_{\overline{X}J} - [N_J - N_j]\mu_* - \sigma\sqrt{N_J - N_j}\Phi^{-1}\left(1-y(t_{\overline{X}J},\mu_*)\right) \\
H\text{-scale} \quad s &= \frac{N_J[N_j\tau^2+\sigma^2]t_{\overline{X}J} - \sigma^2[N_J-N_j]\zeta}{N_J\tau^2+\sigma^2} \\
&\qquad - \frac{\sigma\sqrt{[N_J-N_j][N_J\tau^2+\sigma^2][N_j\tau^2+\sigma^2]}\Phi^{-1}\left(1-y(t_{\overline{X}J},\zeta,\tau^2)\right)}{N_J\tau^2+\sigma^2} \\
\text{noninf } H\text{-scale} \quad s &= N_j t_{\overline{X}J} - \sigma\sqrt{\frac{N_j}{N_J}[N_J-N_j]}\,\Phi^{-1}\left(1-y(t_{\overline{X}J},\zeta,\tau^2=\infty)\right) \\
E_a\text{-scale} \quad s &= F^{-1}\left(j,\, \alpha_\ell y(\mu_a) - \sum_{i=1}^{j-1} F(i,a_{Si};\mu_a)\,;\,\mu_a\right) \\
E_b\text{-scale} \quad s &= F^{-1}\left(j,\, F(j,\infty;\mu_b) - [1-\beta_\ell]y(\mu_b) + \sum_{i=1}^{j-1}\int_{b_{Si}}^{\infty} p(i,u;\mu_b)\,du\,;\,\mu_b\right) \\
E_c\text{-scale} \quad s &= F^{-1}\left(j,\, [1-\beta_u]y(\mu_c) - \sum_{i=1}^{j-1}\int_{-\infty}^{c_{Si}} p(i,u;\mu_c)\,du\,;\,\mu_c\right) \\
E_d\text{-scale} \quad s &= F^{-1}\left(j,\, F(j,\infty;\mu_d) - \alpha_u y(\mu_d) + \sum_{i=1}^{j-1}\int_{d_{Si}}^{\infty} p(i,u;\mu_d)\,du\,;\,\mu_d\right)
\end{aligned}
\tag{1.13}
$$

The following table provides the formulas for converting between a value s measured on the S-scale and a value y on any of the other scales. Note that in the cases of the B-scale, C-scale, and H-scale, y is a function of a mean μ_* and/or a threshold $t_{\overline{X}j}$.

$$
\begin{aligned}
S\text{-scale} \quad & y = s \\
X\text{-scale} \quad & y = \frac{s}{N_j} \\
Z\text{-scale} \quad & y = \sqrt{N_j}\frac{[\frac{s}{N_j} - \mu_0]}{\sigma} \\
P\text{-scale} \quad & y = 1 - \Phi\left(\sqrt{N_j}\frac{[\frac{s}{N_j} - \mu_0]}{\sigma}\right) \\
B\text{-scale} \quad & y(\zeta, \tau^2, \mu_*) = 1 - \Phi\left(\frac{\mu_*[N_j\tau^2 + \sigma^2] - \tau^2 s - \sigma^2\zeta}{\sigma\tau\sqrt{N_j\tau^2 + \sigma^2}}\right) \\
\text{noninf } B\text{-scale} \quad & y(\zeta, \tau^2 = \infty, \mu_*) = 1 - \Phi\left(\frac{\mu_* N_j - s}{\sigma\sqrt{N_j}}\right) \\
C\text{-scale} \quad & y(t_{\overline{X}J}, \mu_*) = 1 - \Phi\left(\frac{N_J[t_{\overline{X}J} - \mu_*] - s + N_j\mu_*)}{\sigma\sqrt{N_J - N_j}}\right) \\
H\text{-scale} \quad & y(t_{\overline{X}J}, \zeta, \tau^2) = 1 - \Phi\left(\frac{N_J[N_j\tau^2 + \sigma^2][t_{\overline{X}J} - \frac{s}{N_j}] + \sigma^2[N_J - N_j][\frac{s}{N_j} - \zeta]}{\sigma\sqrt{[N_J - N_j][N_J\tau^2 + \sigma^2][N_j\tau^2 + \sigma^2]}}\right) \\
\text{noninf } H\text{-scale} \quad & y(t_{\overline{X}J}, \zeta, \infty) = 1 - \Phi\left(\frac{N_J[t_{\overline{X}J} - \frac{s}{N_j}]}{\sigma\sqrt{\frac{N_J}{N_j}[N_J - N_j]}}\right) \\
E_a\text{-scale} \quad & y(\mu_a) = \frac{1}{\alpha_\ell}\left[\sum_{i=1}^{j-1} F(i, a_{Si}; \mu_a) + F(j, s; \mu_a)\right] \\
E_b\text{-scale} \quad & y(\mu_b) = \frac{1}{[1 - \beta_\ell]}\left[\sum_{i=1}^{j-1}\int_{b_{Si}}^{\infty} p(i, u; \mu_b)\, du + F(j, \infty; \mu_b) - F(j, s; \mu_b)\right] \\
E_c\text{-scale} \quad & y(\mu_c) = \frac{1}{[1 - \beta_u]}\left[\sum_{i=1}^{j-1}\int_{-\infty}^{c_{Si}} p(i, u; \mu_c)\, du + F(j, s; \mu_c)\right] \\
E_d\text{-scale} \quad & y(\mu_d) = \frac{1}{\alpha_u}\left\{\sum_{i=1}^{j-1}[F(i, \infty; \mu_d) - F(i, d_{Si}; \mu_d)] + F(j, \infty; \mu_d) - F(j, s; \mu_d)\right\}
\end{aligned}
\tag{1.14}
$$

1.6 Standardizing Transformation

We find it useful to introduce a standardizing transformation on two grounds:

1. In later sections we shall find that many of the calculations required for statistical inference with group sequential sampling are extremely computationally intensive. By reducing each problem down to some standardized form, we can develop computer routines to perform general functions under that standardizing transformation, and then we can transform the output to the original scale desired by the user.

2. In most study design situations, we are interested in determining the sample size which would provide adequate power to detect an alternative hypothesis of interest. We thus need to be able to compute the operating characteristics of a group sequential test in some standardized form that is independent of the sample size, and then solve for the sample size that would provide those operating characteristics for a specific alternative.

We therefore adopt the following standardizing transformation

$$X_i^* = \frac{X_i - \mu_0}{\sigma\sqrt{N_J}},\ i = 1, \ldots, N = N_J. \tag{1.15}$$

Note that when $\mu_0 = 0$ and $N_J\sigma^2 = 1$, this is just the identity transformation.

For notational convenience, we define for $j = 1, \ldots, J$ the proportion of the maximal information accrued by the jth analysis as $\Pi_j = N_j/N_J$ and the proportion accrued between the $(j-1)$th and jth analyses as $\pi_j = n_j/N_J$.

1.6.1 Frequentist Test Statistics

In the standardized setting, the various test statistics can be defined in an analogous fashion to those based on the original data. In particular, in analogy with eqn (1.3), for $j = 1, \ldots, J$, we consider test statistics

$$\begin{aligned}
\text{(partial sum) } S_j^* &= \sum_{i=1}^{N_j} X_i^* \\
\text{(sample mean) } \overline{X}_j^* &= \frac{S_j^*}{\Pi_j} \\
\text{(normalized statistic) } Z_j^* &= \overline{X}_j^*\sqrt{\Pi_j} \\
\text{(fixed sample P value) } P_j &= 1 - \Phi(Z_j^*) = \int_{-\infty}^{Z_j^*} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du
\end{aligned} \tag{1.16}$$

The parallels between eqns (1.3) and (1.16) may not be immediately apparent for the definition of the sample mean and the normalized statistic. However, the connection becomes clear if we examine the distribution of the transformed data. In the general case, $X_i^* \sim \mathcal{N}(\frac{\delta}{N_J}, \frac{1}{N_J})$, where the standardized mean δ is related to the original unknown mean μ according to

$$\delta = \sqrt{N_J}\frac{(\mu - \mu_0)}{\sigma}. \tag{1.17}$$

In particular, in this standardized setting, the null hypothesis μ_0 corresponds to $\delta_0 = 0$, and the prespecified alternatives μ_- and μ_+ correspond to $\delta_- = \sqrt{N_J}(\mu_- - \mu_0)/\sigma$ and $\delta_+ = \sqrt{N_J}(\mu_+ - \mu_0)/\sigma$, respectively.

For a fixed (nonrandom) N_j, the distribution of the standardized partial sum is given by $S_j^* \sim \mathcal{N}(\Pi_j\delta, \Pi_j)$ which depends on the sample size only as the proportion of the maximal sample size. The statistics defined in eqn (1.16) then have the sample mean estimating δ and the null distribution of the normalized statistic (in the fixed sample case) being the standard normal distribution. The formulas given in eqn (1.16) follow directly from eqn (1.3) when we choose $\sigma^2 = 1, \mu_0 = 0, N_J = 1$, and $N_j = \Pi_j$. It is apparent, then, that with the standardizing transformation, N_j is in a sense counting the accrued observations in units corresponding to a proportion of the maximal statistical information N_J/σ^2, and $\frac{1}{\sigma^2}$ is the average statistical information contributed by a single sampling unit.

Each of the above statistics based on the standardizing transformation can be easily related to the corresponding statistic on the untransformed scale.

$$\begin{aligned} S_j^* &= \frac{S_j - N_j\mu_0}{\sqrt{N_J}\sigma} \\ \overline{X}_j^* &= \sqrt{N_J}\frac{\overline{X}_j - \mu_0}{\sigma} \\ Z_j^* &= Z_j \\ P_j^* &= P_j \end{aligned} \tag{1.18}$$

1.6.2 Bayesian Statistics

We can also compute the Bayesian statistics for a given prior. We note that prior distribution $\mu \sim \mathcal{N}(\zeta, \tau^2)$ implies $\delta \sim \mathcal{N}(\zeta^*, \tau^{*2}$, where $\zeta^* = \sqrt{N_J}(\zeta - \mu_0)/\sigma$ and $\tau^{*2} = N_J\tau^2/\sigma^2$. We then find the posterior distribution of δ conditioned on the observations $X_1^*, \ldots, X_{N_j}^*$ is then

$$\delta \,|\, (X_1^*, \ldots, X_{N_j}^*) \sim \mathcal{N}\left(\frac{\Pi_j\tau^{*2}\overline{X}_j^* + \zeta^*}{\Pi_j\tau^{*2}+1}, \frac{\tau^{*2}}{\Pi_j\tau^{*2}+1}\right).$$

the posterior probabilities that the mean δ is greater than some hypothesis δ_* (which might typically be the null hypothesis $\delta_0 = 0$ or prespecified alternative hypotheses δ_+ and δ_-) are given by

$$\begin{aligned} B_j^*(\zeta^*, \tau^{*2}, \delta_*) &= Pr(\delta \geq \delta_* \,|\, (X_1^*, \ldots, X_{N_j^*})) \\ &= 1 - \Phi\left(\frac{\delta_*[\Pi_j\tau^{*2}+1] - \Pi_j\tau^{*2}\overline{X}_j^* - \zeta^*}{\tau^*\sqrt{\Pi_j\tau^{*2}+1}}\right) \end{aligned} \tag{1.19}$$

Each of the above statistics based on the standardizing transformation is exactly equal to the corresponding statistic on the untransformed scale. That is for δ_* and μ_* related by eqn (1.17)

$$B_j^*(\zeta^*, \tau^{*2}, \delta_*) = B_j(\zeta, \tau^2, \mu_*) \tag{1.20}$$

1.6.3 Measures of Futility

The statistics based on conditional probabilities or predictive probabilities can also be computed under the standardizing transformation. We consider the use of $\overline{X}_J^*$ as the test statistic and define $t_{\overline{X}J}^* = \sqrt{N_J}[t_{\overline{X}J} - \mu_)]/\sigma$ as the transformed threshold (using eqn (1.18)) for that test statistic at the Jth analysis. As noted above, these measures of futility could also be specified based on any of the statistics defined in this section, with a suitable transformation of the threshold as defined in section 1.6.5.

The conditional distribution of $\overline{X}_J^*$ given $\overline{X}_j^*$ is found to be

$$\overline{X}_J^* \,|\, \overline{X}_j^* \sim \mathcal{N}\left(\delta + \Pi_j[\overline{X}_j^* - \delta], 1 - \Pi_j\right).$$

Computing conditional probabilities based on some hypothesis δ_* (which might typically be the null hypothesis $\delta_0 = 0$, one of the alternative hypotheses δ_+ or δ_-, or the maximum likelihood estimate $\hat{\delta} = \overline{X}_j^*$, each of which correspond to the appropriate value of μ according to eqn (1.17)) yields:

$$\begin{aligned} C_j^*(t_{\overline{X}j}^*, \delta_*) &\equiv Pr(\overline{X}_J^* > t_{\overline{X}J}^* \,|\, \overline{X}_j^*; \delta_*) \\ &= 1 - \Phi\left(\frac{[t_{\overline{X}J}^* - \delta_*] - \Pi_j[\overline{X}_j^* - \delta]}{\sqrt{1-\Pi_j}}\right) \end{aligned} \tag{1.21}$$

Under the standardizing transformation, the predictive probability based on a normal prior distribution for $\delta(\delta \sim \mathcal{N}(\zeta^*, \tau^{*2})$ where $\zeta^* = \sqrt{N_J}(\zeta - \mu_0)/\sigma$ and $\tau^{*2} = N_J\tau^2/\sigma^2)$ yields a posterior distribution $\lambda^*(\delta|\overline{X}_j^*)$ that is normal as given in section 1.6.2 above. We then compute the marginal conditional distribution of $\overline{X}_J^*$ given $\overline{X}_j^*$ as a normal distribution having mean $\{[\tau^{*2}+1]\Pi_j\overline{X}_j^* + [1-\Pi_j]\zeta^*\}/[\Pi_j\tau^{*2}+1]$ and variance $[1-\Pi_j][\tau^{*2}+1]/[\Pi_j\tau^{*2}+1]$ and survival function

$$\begin{aligned} H_j^*(t_{\overline{X}J}^*, \zeta^*, \tau^{*2}) &\equiv \int Pr(\overline{X}_J^* > t_{\overline{X}J}^* \,|\, \overline{X}_j^*, \delta)\, \lambda^*(\delta \,|\, \overline{X}_j^*)\, d\delta \\ &= 1 - \Phi\left(\frac{[\Pi_j\tau^{*2}+1][t_{\overline{X}J}^* - \overline{X}_j^*] + [1-\Pi_j][\overline{X}_j^* - \zeta^*]}{\sqrt{[1-\Pi_j][\tau^{*2}+1][\Pi_j\tau^{*2}+1]}}\right) \end{aligned} \tag{1.22}$$

When we consider a noninformative prior distribution (taking the limit as $\tau^{*2} \to \infty$), we obtain

$$H_j^*(t_{\overline{X}j}^*, \zeta^*, \infty) = 1 - \Phi\left(\sqrt{\Pi_j}\frac{[t_{\overline{X}J}^* - \overline{X}_j^*]}{\sqrt{1-\Pi_j}}\right)$$

Each of the above statistics based on the standardizing transformation is exactly equal to the corresponding statistic on the untransformed scale.

$$\begin{aligned} C_j^*(t_{\overline{X}j}^*, \delta_*) &= C_j(t_{\overline{X}j}, \mu_*) \\ H_j^*(t_{\overline{X}j}^*, \zeta^*, \tau^{*2}) &= H_j(t_{\overline{X}j}, \zeta, \tau^2) \end{aligned} \tag{1.23}$$

1.6.4 Error Spending Measures

Under the standardizing transformation, the statistics on the error spending scale are computed for constants transformed using eqn (1.18)

$$\begin{aligned} a_{Sj}^* &= \frac{a_{Sj} - N_j\mu_0}{\sigma\sqrt{N_J}} \\ b_{Sj}^* &= \frac{b_{Sj} - N_j\mu_0}{\sigma\sqrt{N_J}} \\ c_{Sj}^* &= \frac{c_{Sj} - N_j\mu_0}{\sigma\sqrt{N_J}} \\ d_{Sj}^* &= \frac{d_{Sj} - N_j\mu_0}{\sigma\sqrt{N_J}} \end{aligned} \tag{1.24}$$

and specified hypotheses transformed using eqn (1.17)

$$\begin{aligned} \delta_a &= \sqrt{N_J}[\mu_a - \mu_0]/\sigma \\ \delta_b &= \sqrt{N_J}[\mu_b - \mu_0]/\sigma \\ \delta_c &= \sqrt{N_J}[\mu_c - \mu_0]/\sigma \\ \delta_d &= \sqrt{N_J}[\mu_d - \mu_0]/\sigma \end{aligned} \tag{1.25}$$

using the functions $f^*(j, s^*; \delta), F^*(j, s^*; \delta)$, and $p^*(j, s^*; \delta)$ defined by eqns (5.2), (5.3), and (5.4) for the standardizing transformation. We thus can define error spending statistics

$$
\begin{aligned}
E^*_{aj}(\delta_a) &= \frac{1}{\alpha^*_\ell}\left[\sum_{i=1}^{j-1} F^*(i, a^*_{Si}; \delta_a) + F^*(j, S^*_j; \delta_a)\right] \\
E^*_{bj}(\delta_b) &= \frac{1}{[1-\beta^*_\ell]}\left[\sum_{i=1}^{j-1} \int_{b^*_{Si}}^{\infty} p^*(i, s^*; \delta_b)\, ds^* + F^*(j, \infty; \delta_b) - F^*(j, S^*_j; \delta_b]\right) \\
E^*_{cj}(\delta_c) &= \frac{1}{[1-\beta^*_u]}\left[\sum_{i=1}^{j-1} \int_{-\infty}^{c^*_{Si}} p^*(i, s^*; \delta_c)\, ds^* + F^*(j, S^*_j; \delta_c)\right] \\
E^*_{dj}(\delta_d) &= \frac{1}{\alpha^*_u}\left\{\sum_{i=1}^{j-1} [F^*(i, \infty; \delta_d) - F^*(i, d^*_{Si}; \delta_d)] + F^*(j, \infty; \delta_d) - F^*(j, S^*_j; \delta_d)\right\} \qquad (1.26)
\end{aligned}
$$

where constants $\alpha^*_\ell, \alpha^*_\mu, \beta^*_\ell$, and β^*_μ are defined by

$$
\begin{aligned}
\alpha^*_\ell &= \sum_{i=1}^{J} F^*(i, a^*_{Si}; \delta_a) \\
\beta^*_\ell &= \sum_{i=1}^{J} F^*(i, a^*_{Si}; \delta_b) \\
\beta^*_u &= \sum_{i=1}^{J} [F^*(i, \infty; \delta_c) - F^*(i, d^*_{Si}; \delta_c)] \\
\alpha^*_u &= \sum_{i=1}^{J} [F^*(i, \infty; \delta_d) - F^*(i, d^*_{Si}; \delta_d)] \qquad (1.27)
\end{aligned}
$$

Each of the above statistics based on the standardizing transformation is exactly equal to the corresponding statistic on the untransformed scale.

$$
\begin{aligned}
E^*_{aj}(\delta_a) &= E_{aj}(\mu_a) \\
E^*_{bj}(\delta_b) &= E_{bj}(\mu_b) \\
E^*_{cj}(\delta_c) &= E_{cj}(\mu_c) \\
E^*_{dj}(\delta_d) &= E_{dj}(\mu_d) \qquad (1.28)
\end{aligned}
$$

Furthermore, we also have that the constants defined by eqn (1.24) are exactly equal to those defined by eqn (1.11). That is, $\alpha^*_\ell = \alpha_\ell, \beta^*_\ell = \beta_\ell, \alpha^*_\mu = \alpha_\mu$, and $\beta^*_\mu = \beta_\mu$.

1.6.5 Summary of Correspondences

Conversions among the various scales under the standardizing transformations are analogous to those presented in eqns (1.13) and (1.14). The following table provides conversions for y^* measured on each of the scales under the standardizing transformation to s^* measured on the S^*-scale. Note that in the cases of the B^*-scale, C^*-scale, and H^*-scale, y^* is a function of a mean δ_* and/or a threshold $t^*_{\overline{X}_j}$.

$$
\begin{aligned}
S^*\text{-scale } s^* &= y^* \\
X^*\text{-scale } s^* &= \Pi_j y^* \\
Z^*\text{-scale } s^* &= \sqrt{\Pi_j} y^* \\
P^*\text{-scale } s^* &= \sqrt{\Pi_j}\Phi^{-1}(1-y^*) \\
B^*\text{-scale } s^* &= \frac{\delta_*[\Pi_j\tau^{*2}+1]-\zeta^*-\tau^*\sqrt{\Pi_j\tau^{*2}+1}\Phi^{-1}(1-y^*(\zeta^*,\tau^{*2},\delta_*))}{\tau^{*2}} \\
\text{noninf } B^*\text{-scale } s^* &= \delta_*\Pi_j - \sqrt{\Pi_j}\Phi^{-1}(1-y^*(\zeta^*,\tau^{*2}=\infty,\delta_*)) \\
C^*\text{-scale } s^* &= t^*_{\overline{X}J} - [1-\Pi_j]\delta_* - \sqrt{[1-\Pi_j]}\Phi^{-1}(1-y^*(t^*_{\overline{X}J},\delta_*)) \\
H^*\text{-scale } s^* &= \frac{[\Pi_j\tau^{*2}+1]t^*_{\overline{X}J}-[1-\Pi_j]\zeta^*}{\tau^{*2}+1} \\
&\qquad -\frac{\sqrt{[1-\Pi_j][\tau^{*2}+1][\Pi_j\tau^{*2}+1]}\Phi^{-1}\left(1-y^*(t^*_{\overline{X}J},\zeta^*,\tau^{*2})\right)}{\tau^{*2}+1} \\
\text{noninf } H^*\text{-scale } s^* &= \Pi_j t^*_{\overline{X}J} - \sqrt{\Pi_j[1-\Pi_j]}\Phi^{-1}(1-y^*(t^*_{\overline{X}J},\zeta^*,\tau^{*2}=\infty)) \\
E^*_a\text{-scale } s^* &= F^{*-1}\left(j,\ \alpha^*_\ell y^*(\delta_a) - \sum_{i=1}^{j-1} F^*(i,a^*_{Si};\delta_a)\,;\ \delta_a\right) \\
E^*_b\text{-scale } s^* &= F^{*-1}\left(j,\ F^*(j,\infty;\delta_b) - [1-\beta^*_\ell]y^*(\delta_b) + \sum_{i=1}^{j-1}\int_{b^*_{Si}}^{\infty} p^*(i,u^*;\delta_b)\,du^*\,;\ \delta_b\right) \\
E^*_c\text{-scale } s^* &= F^{*-1}\left(j,\ [1-\beta^*_u]y^*(\delta_c) - \sum_{i=1}^{j-1}\int_{-\infty}^{c^*_{Si}} p^*(i,u^*;\delta_c)\,du^*\,;\ \delta_c\right) \\
E^*_d\text{-scale } s^* &= F^{*-1}\left(j,\ F^*(j,\infty;\delta_d) - \alpha^*_u y^*(\delta_d) + \sum_{i=1}^{j-1}\int_{d^*_{Si}}^{\infty} p^*(i,u^*;\delta_d)\,du^*\,;\ \delta_d\right) \qquad (1.29)
\end{aligned}
$$

The following table provides the formulas for converting between a value s^* measured on the S^*-scale and a value y^* on any of the other scales under the standardizing transformation. Note that in the cases of the B^*-scale, C^*-scale, and H^*-scale, y^* is a function of a mean δ_* and/or a threshold $t^*_{\overline{X}j}$.

S^*-scale $\quad y^* = s^*$

X^*-scale $\quad y^* = \dfrac{s^*}{\Pi_j}$

Z^*-scale $\quad y^* = \dfrac{s^*}{\sqrt{\Pi_j}}$

P^*-scale $\quad y^* = 1 - \Phi\left(\dfrac{s^*}{\sqrt{\Pi_j}}\right)$

B^*-scale $\quad y^*(\zeta^*, \tau^{*2}, \delta_*) = 1 - \Phi\left(\dfrac{\delta_*[\Pi_j\tau^{*2}+1] - \tau^{*2}s^* - \zeta^*}{\tau^*\sqrt{\Pi_j\tau^{*2}+1}}\right)$

noninf B^*-scale $\quad y^*(\zeta^*, \tau^{*2}=\infty, \delta_*) = 1 - \Phi\left(\dfrac{\delta_*\Pi_j - s^*}{\sqrt{\Pi_j}}\right)$

C^*-scale $\quad y^*(t^*_{\overline{X}j}, \delta_*) = 1 - \Phi\left(\dfrac{t^*_{\overline{X}J} - \delta_* - s^* + \Pi_j\delta_*)}{\sqrt{1-\Pi_j}}\right)$

H^*-scale $\quad y^*(t^*_{\overline{X}j}, \zeta^*, \tau^{*2}) = 1 - \Phi\left(\dfrac{[\Pi_j\tau^{*2}+1][t^*_{\overline{X}J} - \frac{s^*}{\Pi_j}] + [1-\Pi_j][\frac{s^*}{\Pi_j} - \zeta^*]}{\sqrt{[1-\Pi_j][\tau^{*2}+1][\Pi_j\tau^{*2}+1]}}\right)$

noninf H^*-scale $\quad y^*(t^*_{\overline{X}j}, \zeta^*, \tau^{*2}=\infty) = 1 - \Phi\left(\dfrac{t^*_{\overline{X}J} - \frac{s^*}{\Pi_j}}{\sqrt{\frac{(1-\Pi_j)}{\Pi_j}}}\right)$

E^*_a-scale $\quad y^*(\delta_a) = \dfrac{1}{\alpha^*_\ell}\left[\sum_{i=1}^{j-1} F^*(i, a^*_{Si}; \delta_a) + F^*(j, s^*; \delta_a)\right]$

E^*_b-scale $\quad y^*(\delta_b) = \dfrac{1}{[1-\beta^*_\ell]}\left[\sum_{i=1}^{j-1}\int_{b^*_{Si}}^{\infty} p^*(i, u^*; \delta_b)\, du^* + F^*(j, \infty; \delta_b) - F^*(j, s^*; \delta_b)\right]$

E^*_c-scale $\quad y^*(\delta_c) = \dfrac{1}{[1-\beta^*_u]}\left[\sum_{i=1}^{j-1}\int_{-\infty}^{c^*_{Si}} p^*(i, u^*; \delta_c)\, du^* + F^*(j, s^*; \delta_c)\right]$

E^*_d-scale $\quad y^*(\delta_d) = \dfrac{1}{\alpha^*_u}\left\{\sum_{i=1}^{j-1}[F^*(i, \infty; \delta_d) - F^*(i, d^*_{Si}; \delta_d)] + F^*(j, \infty; \delta_d) - F^*(j, s^*; \delta_d)\right\}$ (1.30)

Conversions between the scales under the standardized transformation and the corresponding scales on the untransformed scales are straightforward. The following table provides conversions between a measurement y on an untransformed scale and a measurement y^* on the corresponding standardized scale. Note that in the cases of the B^*-scale, C^*-scale, and H^*-scale, y^* is a function of a mean δ_* and/or a threshold $t^*_{overlineXj}$ which correspond respectively to the mean μ_* and threshold $t_{\overline{X}_j}$ according to the transformations specified by eqns (1.17) and (1.18) as in section (1.6.3). Similarly, the error spending scales assume that the hypotheses and boundaries on the standardized scale have been transformed from the original scale as outline by eqns (1.24) and (1.25) in section (1.6.4).

$$
\begin{aligned}
S\text{-scale} \rightarrow S^*\text{-scale} \qquad & y^* = \frac{y - N_j\mu_0}{\sqrt{N_J}\sigma} \\
X\text{-scale} \rightarrow X^*\text{-scale} \qquad & y^* = \sqrt{N_J}\frac{y - \mu_0}{\sigma} \\
Z\text{-scale} \rightarrow Z^*\text{-scale} \qquad & y^* = y \\
P\text{-scale} \rightarrow P^*\text{-scale} \qquad & y^* = y \\
B\text{-scale} \rightarrow B^*\text{-scale} \qquad & y^*(\zeta^*, \tau^{*2}, \delta_*) = y(\zeta, \tau, \mu_*) \\
C\text{-scale} \rightarrow C^*\text{-scale} \qquad & y^*(t^*_{\overline{X}j}, \delta_*) = y(t_{\overline{X}j}, \mu_*) \\
H\text{-scale} \rightarrow H^*\text{-scale} \qquad & y^*(t^*_{\overline{X}j}, \zeta^*, \tau^{*2}) = y(t_{\overline{X}J}, \zeta, \tau^2) \\
E_a\text{-scale} \rightarrow E^*_a\text{-scale} \qquad & y^*(\delta_a) = y(\mu_a) \\
E_b\text{-scale} \rightarrow E^*_b\text{-scale} \qquad & y^*(\delta_b) = y(\mu_b) \\
E_c\text{-scale} \rightarrow E^*_c\text{-scale} \qquad & y^*(\delta_c) = y(\mu_c) \\
E_d\text{-scale} \rightarrow E^*_d\text{-scale} \qquad & y^*(\delta_d) = y(\mu_d)
\end{aligned} \tag{1.31}
$$

The following table provides conversions between a measurement y^* on a standardized scale and a measurement y on the corresponding untransformed scale.

$$
\begin{aligned}
S^*\text{-scale} \rightarrow S\text{-scale} \qquad & y = N_j\mu_0 + \sqrt{N_J}\sigma y^* \\
X^*\text{-scale} \rightarrow X\text{-scale} \qquad & y = \mu_0 + \frac{\sigma}{\sqrt{N_J}} y^* \\
Z^*\text{-scale} \rightarrow Z\text{-scale} \qquad & y = y^* \\
P^*\text{-scale} \rightarrow P\text{-scale} \qquad & y = y^* \\
B^*\text{-scale} \rightarrow B\text{-scale} \qquad & y(\zeta, \tau, \mu_*) = y^*(\zeta^*, \tau^{*2}, \delta_*) \\
C^*\text{-scale} \rightarrow C\text{-scale} \qquad & y(t_{\overline{X}j}, \mu_*) = y^*(t^*_{\overline{X}j}, \delta_*) \\
H^*\text{-scale} \rightarrow H\text{-scale} \qquad & y(t_{\overline{X}j}, \zeta, \tau^2) = y^*(t^*_{\overline{X}j}, \zeta^*, \tau^{*2}) \\
E^*_a\text{-scale} \rightarrow E_a\text{-scale} \qquad & y(\mu_a) = y^*(\delta_a) \\
E^*_b\text{-scale} \rightarrow E_b\text{-scale} \qquad & y(\mu_b) = y^*(\delta_b) \\
E^*_c\text{-scale} \rightarrow E_c\text{-scale} \qquad & y(\mu_c) = y^*(\delta_c) \\
E^*_d\text{-scale} \rightarrow E_d\text{-scale} \qquad & y(\mu_d) = y^*(\delta_d)
\end{aligned} \tag{1.32}
$$

1.7 Parameter Scales

Heretofore in this section, we have considered various scales to be used in describing the sample space for the test statistic in the setting of a one sample test for the mean of a normally distributed random variable. In sections (1.1)–(1.5), we considered the scales for the sample space of the test statistics for the unstandardized problem, and in section (1.6) we considered the analogous scales for the sample space of the test statistic for the standardized problem.

It is now useful to formally define scales for the parameter space. That is, in the unstandardized problem, we were primarily interested in making inference about μ, the mean of the normally distributed random variable, and we can refer to this parameter scale as the unstandardized parameter scale and denote it as the μ-scale. In the standardized problem, we can make equivalent inference about the standardized parameter δ, and we can refer to the parameter scale in the standardized problem as the standardized parameter scale and denote it as the δ-scale. From eqn (1.17) we can derive the formulas for converting between the μ and δ scales

$$\begin{aligned} &\mu\text{-scale} \rightarrow \delta\text{-scale} && \delta = \sqrt{N_J}\frac{(\mu - \mu_0)}{\sigma} \\ &\delta\text{-scale} \rightarrow \mu\text{-scale} && \mu = \mu_0 + \frac{\sigma}{\sqrt{N_J}}\delta \end{aligned} \tag{1.33}$$

As discussed in greater detail in section 3, the basic probability model that considers the mean of a normally distributed random variable can serve as the foundation for a wide variety of probability models frequently used in the analysis of clinical trial data. In these various probability models, a parameter θ measuring the treatment effect can be related to the parameter μ of our basic model. Typically θ is a parameter that is more generally understood by non-statisticians than the μ parameter that is perhaps used in a statistical model. For instance, θ representing the odds ratio (in logistic regression models) or the hazard ratio (in proportional hazards regression models) might be measures of treatment effect more readily understood by a clinician than the log odds ratio or the log hazard ratio (which are the interpretations of the regression parameters in the corresponding statistical models). We will therefore refer to θ as the "natural" parameter. We note that this terminology is not at all restrictive, as in section 3 we do allow for (but do not particularly recommend using) probability models in which θ represents, say, the log odds ratio or the log hazard ratio.

In the general case, we consider a transformation

$$\mu = \psi g(\theta) \tag{1.34}$$

where ψ is some constant and $g(\cdot)$ is a link function used in the statistical model. We can thus also define the probability model parameter scale (denoted as the θ-scale) for the parameter space based on the relationship specified by eqn (1.34), and derive conversions between the μ-scale and the θ-scale as

$$\begin{aligned} &\theta\text{-scale} \rightarrow \mu\text{-scale} && \mu = g(\theta)\psi \\ &\mu\text{-scale} \rightarrow \theta\text{-scale} && \theta = g^{-1}\left(\frac{\mu}{\psi}\right) \end{aligned} \tag{1.35}$$

Under this parameterization, we note that it will generally be of more interest to consider at the jth analysis a test statistic $\hat{\theta}_j$ corresponding to the maximum likelihood estimate of θ, rather than focusing on the maximum likelihood estimate $\overline{X}_j$ of μ. We thus define

$$\hat{\theta}_j = g^{-1}\left(\frac{\overline{X}_j}{\psi}\right). \tag{1.36}$$

Based on the above, we can also define an estimate scale (denoted as the $\hat{\theta}$-scale) for statistics at the jth analysis. We note that at the jth analysis, conversions between a measurement x on the X-scale and a measurement $\hat{\theta}$ on the $\hat{\theta}$-scale are easily derived from eqn (1.36) as

$$\begin{aligned} &\hat{\theta}\text{-scale} \rightarrow X\text{-scale} && x = g(\hat{\theta})\psi \\ &X\text{-scale} \rightarrow \hat{\theta}\text{-scale} && \hat{\theta} = g^{-1}\left(\frac{\overline{X}_j}{\psi}\right) \end{aligned} \tag{1.37}$$

We note that the $\hat{\theta}$-scale is primarily of interest when providing an intuitive interface for one of the statistical models presented in section 3. As is described in later sections, computations related to group sequential inference will generally consider the standardized partial sum scale (S^*-scale)

for statistics and the standardized parameter scale (δ-scale) for parameters. We assume that the user will typically want to use only the "natural" scale (the θ-scale as defined by the probability model) for input and output related to the parameter space. Hence, while the user will typically be allowed to choose any of the unstandardized scales (as summarized in eqn (1.12)) for input and output related to the sample space for statistics, we shall assume that the $\hat{\theta}$-scale (estimating the natural parameter) will be of far greater interest than the X-scale (potentially estimating some transformation of the natural parameter). For this reason, we will generally suppress the true X-scale and use the $\hat{\theta}$-scale in its place. In fact, in the S+SeqTrial functions, it is the $\hat{\theta}$-scale that will be referred to as the "sample mean" scale. In this document, however, we will use the term "sample mean" scale to refer to the X-scale and the term "estimate" scale to refer to the $\hat{\theta}$-scale. In this setting it is unnecessary to define a separate standardized $\hat{\theta}$-scale, because we can consider the X^*-scale as a standardized form of the $\hat{\theta}$-scale.

1.8 Impact of Boundary Scales on User Interface for S+SeqTrial

The major instances in which the user will need to consider the scale for expressing test statistics include

1. Specification of parameters for the Bayesian, conditional futility, predictive futility scales, and error spending scales

   These scales for test statistics require input of additional parameters, some of which must in turn be specified on a particular scale.

2. Specification of design family for the stopping boundaries.

   The various families of group sequential designs described in section 8 are in turn based on specific choices of scales for test statistics outlined above, or particular combinations of those scales as described briefly below. The design parameters A_*, P_*, R_*, and G_* will refer to boundary relationships on the scale(s) corresponding to the design family. Simple relationships between stopping boundaries at successive relationships will tend to exist only on the boundary scale corresponding to the group sequential design family.

3. Specification of exact, minimum, or maximum constraints for stopping boundaries.

   The constrained boundaries are based on one or more of the design families. Hence any user specified stopping boundary at a particular analysis time will be interpreted according to the test statistic scale used in defining the constraint. (See section 11 for a more detailed description of constraints on group sequential design family.)

4. Specification of display scale for the stopping boundaries.

   The output stopping boundaries will be expressed on the scale specified by the display scale. The spectrum of test statistic scales used for display is somewhat richer than the spectrum for which design families have been designed.

5. Specification of test statistics for input to module routines for integration of the sampling density, monitoring of the study, or reporting results of a final analysis.

6. Conversion of test statistics between individual scales.

Specifications of the scales is through definition of a `seqScale` object using the S+SeqTrial function `seqScale()`. The required argument to that function is `scaleType`, which accepts a character valued scalar which is one of 'S' (for partial sum statistic family), 'X' (for unified or

sample mean statistic family), 'Z' (for normalized statistic family), 'P' (for fixed sample P value statistic family), 'B' (for Bayesian family), 'C' (for conditional futility family), 'H' (for predictive futility family), or 'E' (for error spending family). Some of the scales corresponding to these design families require input of additional parameters, and this is effected through specification of the argument `scaleParameters` as a numeric vector which has interpretation specific to the scale family selected. An alternative (and generally easier) specification of the parameters makes use of an appropriate selection of the additional arguments `threshold`, `hypTheta`, `priorTheta`, `priorVariation`, `pessimism`, and `boundaryNumber`.

the additional parameters required for certain boundary scales reflect the parameters needed for computation of the test statistics on those scales. The include

- the B-scale statistic's definition is based on the mean ζ and variance τ^2 of the prior distribution for μ, as well as a threshold μ_* for the computation of the posterior probability,
- the C-scale statistic's definition is based on the threshold $t_{\overline{X}J}$ and the hypothesized value of the mean μ_* to use in the computation of the conditional probability,
- the H-scale statistic's definition is based on the threshold $t_{\overline{X}J}$ and the mean ζ and variance of τ^2 of the prior distribution for μ, and
- the E_a-, E_b-, E_c-, and E_d-scales are combined into a single error spending scale which then requires specification of which of the four subscales is desired as well as the value of the hypothesized means μ_a, μ_b, μ_c, or μ_d.

Furthermore, within each of those four families of scales, combinations which use different parameters for each of the four boundaries are useful in evaluation of the operating characteristics of group sequential stopping rules and in the definition of families of group sequential designs. Each of those scales will therefore need an additional indicator of the particular combination of parameters to be used. Hence, the possible choices for the specification of scale parameters for each of the statistics scales are as given below. When discussing the scales used for presentation of boundaries (as opposed to statistics representing arbitrary possible outcomes), we sue the notation a_j, b_j, c_j, and d_j to represent the stopping boundaries at the jth analysis as introduced in section 4 when the scale is either unimportant or clear. When it is necessary to distinguish the scale on which the stopping boundaries are represented, we denote that with an additional subscript, e.g. $a_{\overline{X}j}$, $a_{\hat{\theta}j}$, and a_{Bj} will represent the "a" boundary at the jth analysis on the sample mean, estimate, and Bayesian scales, respectively. We use $\mu_a = \mu_{0-}, \mu_b = \mu_-, \mu_c = \mu_+$, and $\mu_d = \mu_{0+}$ to represent the hypotheses being rejected by the corresponding boundaries as discussed in sections 2 and 8. We note that input of statistics to the S+SeqTrial `seqScale` function will generally be made on the estimate ($\hat{\theta}$) scale and input of parameters will generally be made on the probability model (θ) scale. These two scales are described in section 1.7.

S-scale `seqScale` ("S") (no parameters needed)

X-scale `seqScale` ("X") (no parameters needed)

Z-scale `seqScale` ("Z") (no parameters needed)

P-scale `seqScale` ("P") (no parameters needed)

B-scale Three families of subscales are possible where the interpretation of additional parameters to `seqScale`("B",...) are as follows

0. `scaleParameters` $= c(0, g^{-1}(\zeta/\psi), \tau^2, g^{-1}(\mu_*/\psi))$ (or `priorTheta` $= g^{-1}(\zeta/\psi)$, `priorVariation` $= \tau^2$, `threshold` $= g^{-1}(\mu_*/\psi)$) when used for boundaries or statistics will correspond to $B_j(\zeta, \tau^2, \mu_*)$

1. `scaleParameters` $= c(1, g^{-1}(\zeta/\psi), \tau^2)$ (or `priorTheta` $= g^{-1}(\zeta/\psi)$, `priorVariation` $= \tau^2$) when used for boundaries will correspond to comparing the similarly transformed test statistics to boundaries as follows

$$\begin{aligned} a_{Bj} &= 1 - B_j(\zeta, \tau^2, \mu_a = \mu_{0-}) \\ b_{Bj} &= B_j(\zeta, \tau^2, \mu_b = \mu_{-}) \\ c_{Bj} &= 1 - B_j(\zeta, \tau^2, \mu_c = \mu_{+}) \\ d_{Bj} &= B_j(\zeta, \tau^2, \mu_d = \mu_{0+}) \end{aligned}$$

where $\mu_- \le \mu_{0+} \le \mu_{0-} \le \mu_+$ are hypotheses being tested in the group sequential clinical test as described in section 2.

2. `scaleParameters` $= c(2, \omega, \tau^2)$ (or `priorVariation` $= \tau^2$, `pessimism` $= \omega$) when used for boundaries will correspond to comparing the following statistics to boundaries as follows

$$\begin{aligned} a_{Bj} &= 1 - B_j(\zeta = \mu_a + \omega\tau, \tau^2, \mu_a = \mu_{0-}) \\ b_{Bj} &= B_j(\zeta = \mu_b - \omega\tau, \tau^2, \mu_b = \mu_{-}) \\ c_{Bj} &= 1 - B_j(\zeta = \mu_c + \omega\tau, \tau^2, \mu_c = \mu_{+}) \\ d_{Bj} &= B_j(\zeta = \mu_d - \omega\tau, \tau^2, \mu_d = \mu_{0+}) \end{aligned}$$

where ω is a measure of the pessimism which is to be sued in determining the prior distribution when rejecting particular hypotheses, and $u_- \le \mu_{0+} \le \mu_{0-} \le \mu_+$ are hypotheses being tested in the group sequential clinical test as described in section 2.

C-scale Three families of subscales are possible where the interpretation of additional parameters to `seqScale`("C",...) are as follows

0. `scaleParameters` $= c(0, g^{-1}(t_{\overline{X}J}/\psi), g^{-1}(\mu_*/\psi))$ (or `hypTheta` $= g^{-1}(\mu_*/\psi)$, `threshold` $= g^{-1}(t_{\overline{X}J}/\psi)$) when used for boundaries or statistics will correspond to $C_j(t_{\overline{X}J}, \mu_*)$
1. `scaleParameters` $= 1$ (or `hypTheta` = "design") when used for boundaries will correspond to comparing the following statistics to boundaries as follows for $j < J$

$$\begin{aligned} a_{Cj} &= C_j(a_{\overline{X}J}, \mu_a = \mu_{0-}) \\ b_{Cj} &= 1 - C_j(b_{\overline{X}J}, \mu_b = \mu_{-}) \\ c_{Cj} &= C_j(c_{\overline{X}J}, \mu_c = \mu_{+}) \\ d_{Cj} &= 1 - C_j(d_{\overline{X}J}, \mu_d = \mu_{0+}) \end{aligned}$$

where $\mu_- \le \mu_{0+} \le \mu_{0-} \le \mu_+$ are hypotheses being tested in the group sequential clinical test as described in section 2. For $j = J$, we define $a_{cJ} = b_{cJ} = c_{cJ} = d_{cJ} = 0.5$.

2. `scaleParameters` $= 2$ (or `hypTheta` = "estimate") when used for boundaries will correspond to comparing the following statistics to boundaries as follows for $j < J$

$$\begin{aligned} a_{Cj} &= C_j(a_{\overline{X}J}, \mu_a = a_{\overline{X}J}) \\ b_{Cj} &= 1 - C_j(b_{\overline{X}J}, \mu_b = b_{\overline{X}J}) \\ c_{Cj} &= C_j(c_{\overline{X}J}, \mu_c = c_{\overline{X}J}) \\ d_{Cj} &= 1 - C_j(d_{\overline{X}J}, \mu_d = d_{\overline{X}J}) \end{aligned}$$

For $j = J$, we define $a_{cJ} = b_{cJ} = c_{cJ} = d_{cJ} = 0.5$.

H-scale Three families of subscales are possible where the interpretation of additional parameters to `seqScale`("H",...) are as follows

0. `scaleParameters` $= c(0, g^{-1}(\zeta/\psi), \tau^2, g^{-1}(t_{\overline{X}J}/\psi))$ (or `priorTheta` $= g^{-1}(\zeta/\psi)$, `priorVariation` $= \tau^2$, `threshold` $= g^{-1}(t_{\overline{X}J}/\psi)$) when used for boundaries or statistics will correspond to $H_j(t_{\overline{X}J}, \zeta, \tau^2)$

1. `scaleParameters` $= c(1, g^{-1}(\zeta/\psi), \tau^2)$ (or `priorTheta` $= g^{-1}(\zeta/\psi)$, `priorVariation` $= \tau^2$) when used for boundaries will correspond to comparing the following statistics to boundaries as follows for $j < J$

$$\begin{aligned} a_{Hj} &= H_j(a_{\overline{X}J}, \zeta, \tau^2) \\ b_{Hj} &= 1 - H_j(b_{\overline{X}J}, \zeta, \tau^2) \\ c_{Hj} &= H_j(c_{\overline{X}J}, \zeta, \tau^2) \\ d_{Hj} &= 1 - H_j(d_{\overline{X}J}, \zeta, \tau^2) \end{aligned}$$

   For $j = J$, we define $a_{HJ} = b_{HJ} = c_{HJ} = d_{HJ} = 0.5$.

2. `scaleParameters` $= c(2, \omega, \tau^2)$ (or `pessimism` $= \omega$, `priorVariation` $= \tau^2$ when used for boundaries will correspond to comparing the following statistics to boundaries as follows

$$\begin{aligned} a_{Hj} &= H_j(a_{\overline{X}J}, \zeta = \mu_a + \omega\tau, \tau^2) \\ b_{Hj} &= 1 - H_j(b_{\overline{X}J}, \zeta = \mu_b - \omega\tau, \tau^2) \\ c_{Hj} &= H_j(c_{\overline{X}J}, \zeta = \mu_c + \omega\tau, \tau^2) \\ d_{Hj} &= 1 - H_j(d_{\overline{X}J}, \zeta = \mu_d - \omega\tau, \tau^2) \end{aligned}$$

   where ω is a measure of the pessimism which is to be used in determining the prior distribution when rejecting particular hypotheses. For $j = J$, we define $a_{HJ} = b_{HJ} = c_{HJ} = d_{HJ} = 0.5$.

E-scale Two families of subscales are possible where the interpretation of additional parameters to `seqScale`("E",...) are as follows

0. `scaleParameters` $= c(0, b, g^{-1}(\mu/\psi))$ (or `boundaryNumber` = c("a", "b", "c", "d",)[$b+1$], `hypTheta` $= g^{-1}(\mu/\psi)$) when used for boundaries of statistics will correspond to $E_{aj}(\mu), E_{bj}(\mu), E_{cj}(\mu)$, or $E_{dj}(\mu)$ according to whether $b = 0, 1, 2$, or 3, respectively.

1. `scaleParameters` = 1 (or no additional parameters specified) when used for boundaries will correspond to the error spending functions for a stopping rule (see section 6.3) when the `seqScale` object is supplied to S+SeqTrial functions `seqDesign()` or seqBoundary() and will correspond to comparing the following statistics to boundaries as follows when the `seqScale` object is supplied to S+SeqTrial function `changeSeqScale()`

$$\begin{aligned} a_{Ej} &= E_{aj}(\mu_a = \mu_{0-}) \\ b_{Ej} &= E_{bj}(\mu_b = \mu_{-}) \\ c_{Ej} &= E_{cj}(\mu_c = \mu_{+}) \\ d_{Ej} &= E_{dj}(\mu_d = \mu_{0+}) \end{aligned}$$

   The `seqScale` objects defined in the manner described above are then used in the following ways

1. The group sequential design family to be used with the design parameters A_*, P_*, R_*, and G_* is specified through

   ```
   scale = seqScale(...)
   ```

`seqDesign(...,design.family=scale,...)`

where scale is a `seqScale` object corresponding to one of the "X", "S", "Z", or "E" scale families for stopping boundaries. (Group sequential design families defined for other scales have not yet been implemented in S+SeqTrial.)

2. User specification of exact values for stopping boundaries at specific analyses is effected through

   `scale = seqScale(...)`

   `bounds = seqBoundry(`*`bndrymtx`*`, scale)`

   `seqDesign(...,exact.constraint=bounds,`*`ldots`*`)`

   where *bndrymtx* is a numeric matrix containing the desired values specified on the boundary scale specified by `scale`, which is a `seqScale` object corresponding to one of the scale families for stopping boundaries. The boundary scale used for exact constraints must be compatible with the boundary scale used for the design family: Only constraints expressed on the error spending scale are valid with the error spending design family, and all scales except the error spending scale are valid for constraints used with other design families.

3. User specifications of minimum values for stopping boundaries at specific analyses is effected through

   `scale = seqScale(...)`

   `bounds = seqBoundry(`*`bndrymtx`*`, scale)`

   `seqDesign(...,minimum.constraint=bounds,`*`ldots`*`)`

   where *bndrymtx* is a numeric matrix containing the desired values specified on the boundary scale specified by `scale`, which is a `seqScale` object corresponding to one of the scale families for stopping boundaries. The boundary scale used for minimum constraints must be compatible with the boundary scale used for the design family: Only constraints expressed on the error spending scale are valid with the error spending design family, and all scales except the error spending scale are valid for constraints used with other design families.

4. User specification of maximum values for stopping boundaries at specific analyses is effected through

   `scale = seqScale(...)`

   `bounds = seqBoundry(`*`bndrymtx`*`, scale)`

   `seqDesign(...,maximum.constraint=bounds,...)`

   where *bndrymtx* is a numeric matrix containing the desired values specified on the boundary scale specified by `scale`, which is a `seqScale` object corresponding to one of the scale families for stopping boundaries. The boundary scale used for maximum constraints must be compatible with the boundary scale used for the design family: Only constraints expressed on the error spending scale are valid with the error spending design family, and all scales except the error spending scale are valid for constraints used with other design families.

5. User specification of the boundary scale for display of boundaries is effected through

   `scale = seqScale(...)`

   `seqDesign(...,display.scale=scale,...)`

   where `scale` is any valid `seqScale` object. (Note that the valid parameters for use with input and output of test statistics or output of stopping boundaries are more varied than those which are valid for design families.)

6. User specification of the boundary scale for input of test statistics is effected through, for instance,

   ```
   scale = seqScale(...)

   seqInference(...,inScale=scale,...)
   ```

   where `scale` is any valid `seqScale` object and corresponds to the scale that the `seqInference()` argument `observed` is measured on. (Note that the valid parameters for use with input and output of test statistics or output of stopping boundaries are more varied than those which are valid for design families.)

2 Statistical Decision Rules

In conducting a clinical trial, we are most often interested in deciding how some new treatment affects a clinical outcome. If the parameter μ is a measure of that treatment effect, then the goal of the clinical trial is often phrased in terms of making a decision for one of several hypotheses by constructing a decision rule that defines for which outcomes a particular decision is made. Typically, the statistical decision rule is constructed according to frequentist methods which quantify the probability of observing particular data when some null hypothesis is true. Alternative approaches can be bas on Bayesian methods which use the data along with some prior probability distribution to quantify the probability that some hypothesis is true.

2.1 A Frequentist Approach: Hypothesis Testing

In classical hypothesis testing, we generally wish to discriminate among at most three hypotheses: that the unknown mean is greater than the null hypothesis ($H_+ : \mu > \mu_0$), that the unknown mean is less than the null hypothesis ($H_- : \mu < \mu_0$), or that the data are consistent with the null hypothesis ($H_0 : \mu = \mu_0$). We note that in one-sided hypothesis testing, we may not try to distinguish two of the hypotheses. For instance, when testing H_0 against a higher alternative, we may not distinguish between H_0 and H_-.

Although it is not uncommon for researchers to speak of deciding in favor of one of the above hypotheses, we must recall that our frequentist inference is actually based on rejecting one or more hypotheses. Thus, we speak of deciding for H_+ only if we have rejected H_0 and H_-, and we speak of deciding for H_- only if we have rejected H_0 and H_+. In the classical frequentist hypothesis testing, we never decide for H_0. This is because for any finite sample size, there are, for instance, samples that are typical both of H_0 and H_+. For a finite sample size, it is always possible to find some small $\epsilon > 0$ such that the distributions of the data are statistically indistinguishable when $\mu = \mu_0$ or when $\mu = \mu_0 + \epsilon$. However, when we are using the results of a hypothesis test to decide whether to adopt a new treatment, if we do not reject H_0, we usually take an action that is in essence rejecting H_+ and H_-. Thus, we desire to develop a decision theoretic model under which we can quantify the interpretation of a failure to reject H_0.

Such a model will demand a reformulation of our alternative hypotheses, because, as noted above, with a finite sample size we can never reject the possibility that μ is marginally smaller than or greater than μ_0. Thus, we now formulate our alternative hypotheses as $H_+ : \mu \geq \mu_+$ and H_-, where $\mu_+ > \mu_0$ and $\mu_- < \mu$). The values of the alternatives μ_- and μ_+ can be chosen in one of two ways. In the first scenario, the alternatives are chosen to correspond to differences in outcome which it is clinically important to distinguish. Study sample sizes are then chosen to allow sufficient statistical power to reject H_0 when $\mu = \mu_+$ or sufficient statistical power to reject H_0 when $\mu = \mu_-$ (note that it is not always possible to satisfy arbitrarily chosen power constraints with arbitrarily chosen values of μ_+ and μ_-). In the second approach, the available sample size is constrained, and we instead compute the alternatives μ_- and μ_+ for which the test design has sufficient statistical power.

In this framework, we can regard a two-sided hypothesis test as a combination of two one-sided tests: an upper test of $H_{0+} : \mu \leq \mu_{0+}$ versus $H_+ : \mu \geq \mu_+$ and a lower test of $H_{0-} : \mu \geq \mu_{0-}$ versus $H_- : \mu \leq \mu_-$. In the classical two-sided hypothesis test, we choose $\mu_{0+} = \mu_{0-} = \mu_0$. However, we introduce the more general setting in which $\mu_- \leq \mu_{0+} \leq \mu_{0-} \leq \mu_+$ in order to accommodate more flexible designs in the group sequential setting (see sections 8 and 9).

In order to maintain the same level of evidence for rejection of any hypothesis, we can choose a study design for which the type I and type II errors are equal. Thus, if we conduct a one-sided level α test of H_{0+} against H_+ (respectively, H_{0-} against H_-), we choose μ_+ (respectively, μ_-) such that we reject H_{0+} (respectively, H_{0-} with probability $1 - \alpha$ when $\mu = \mu_+$ (respectively, $\mu = \mu_-$). In a two-sided level 2α test of H_0 (so $\mu_{0+} = \mu_{0-} = \mu_0$ and we falsely reject H_{0+} in favor of H_+

with probability α and falsely reject H_{0-} in favor of H_- with probability α), we choose μ_+, μ_-, and our sample size such that we reject H_{0+} with probability $1-\alpha$ when $\mu = \mu_-$ and we reject H_{0-} with probability $1-\alpha$ when $\mu = \mu_+$. Such a strategy guarantees that at the end of the study the $100(1-2\alpha)\%$ confidence interval will with probability 1 not contain both μ_+ and μ_{0+}, nor would it contain both μ_- and μ_{0-}. In this way, the study will with $100(1-2\alpha)\%$ confidence discriminate between the null and alternative hypotheses for each of the overlaid one-sided hypothesis tests.

While the above formulation using a common criterion for statistical evidence is our preferred approach, many users will choose power constraints at some level less than $1-\alpha$. Hence, for generality, we shall introduce the notation

$$\begin{aligned} Pr(\text{reject } H_{0+} \text{ for } H_+;\, \mu_{0+}) &= \alpha_u \\ Pr(\text{reject } H_{0+} \text{ for } H_+;\, \mu_+) &= \beta_u \\ Pr(\text{reject } H_{0-} \text{ for } H_-;\, \mu_{0-}) &= \alpha_\ell \\ Pr(\text{reject } H_{0-} \text{ for } H_-;\, \mu_-) &= \beta_\ell \end{aligned} \tag{2.1}$$

The two-sided hypothesis test constructed under this notation is then level $\alpha_u + \alpha_\ell$. As noted above, we shall most often recommend the choice $\alpha_u = \alpha_\ell = \alpha$ and $\beta_u = \beta_\ell = 1-\alpha$, which is symmetric in the type I and type II statistical errors. As a rule, we shall adopt this latter convention in this document, although the more general notation will be used when it is desirable to have the widest application. in any case, in order to preserve the natural ordering of $\mu_+ \geq \mu_0 \geq \mu_{0+} \geq \mu_-$, we will demand that the following constraints be satisfied

$$\begin{aligned} \alpha_u &\leq \beta_u \\ \alpha_\ell &\leq \beta_\ell \\ \alpha_u + \alpha_\ell &\leq 1 \end{aligned} \tag{2.2}$$

In choosing our hypotheses in the above symmetric fashion, we note that some values of μ do not belong strictly to any single hypothesis. For example, for a one-sided level α test of H_{0+} versus H_+ having statistical power $1-\alpha$ to detect H_+, if $\mu \geq \mu_+$, we can with $100(1-\alpha)\%$ confidence state that our study will result in rejection of H_{0+} in favor of H_{0+}, and if $\mu \leq \mu_{0+}$, we can with $100(1-\alpha)\%$ confidence state that our study will result in rejection of H_+ in favor of H_{0+}. However, if $\mu_{0+} < \mu < \mu_+$, we can not be $100(1-\alpha)\%$ confident of either rejecting H_{0+} or H_+. This would suggest that our rejection of H_{0+} can only be interpreted *a priori* as being consistent with the decision that $\mu > \mu_{0+}$, and that rejection of H_+ can only be interpreted *a priori* as being consistent with the decision that $\mu < \mu_+$ (we note that at the completion of the study, computation of confidence intervals will provide more precise quantification of our inference). The interval (μ_{0+}, μ_+) constitutes an equivocal region of our parameter space for the unknown mean μ, because it is not inconsistent (at the $100(1-\alpha)\%$ level of confidence) with either decision.

To formalize this idea, we define an equivocal region EQ for a hypothesis test by

$$EQ = \{\mu : Pr(\text{reject } H_{0+} \text{ for } H_+; \mu) < 1-\alpha \text{ and } Pr(\text{reject } H_{0-} \text{ for } H_-; \mu) < 1-\alpha\}. \tag{2.3}$$

The interpretation of the equivocal region will depend upon the particular application. For instance, in one-sided tests of a new treatment against a placebo, the equivocal region should carry the interpretation of levels of improvement that are not of sufficient clinical importance to warrant increasing the sample size in order to be confident that they will be detected. In two-sided tests comparing two treatments, the equivocal region should be interpreted as levels of difference between the treatments that are so small as to allow decisions of equivalence. In any case, if the equivocal region includes values of the mean that it is clinically important to distinguish from the null hypothesis, the sample size should be increased to make the equivocal region smaller.

2.2 A Bayesian Approach

We now consider a Bayesian approach to distinguishing among the hypotheses considered above. In a Bayesian analysis, decisions are based on the posterior probability of a specific hypothesis. There are a variety of equally valid ways of defining a Bayesian testing procedure. For instance, one can make decisions for a null hypothesis when the posterior probability that the mean is in some close neighborhood of the null hypothesis is sufficiently high, or one can make a decision for a null hypothesis when the posterior probability that the mean is in either of the alternative hypotheses is low. In our development here, we adopt the latter strategy. That is, in order to maintain the greatest parallel with the frequentist approach, we choose to describe our decisions in terms of rejecting hypotheses rather than acceptance of hypotheses.

We again consider the most general case of the superposition of two decision problems, which we will continue to describe in terms of hypotheses. Hence we consider an upper pair of hypotheses $H_{0+} : \mu \leq \mu_{0+}$ versus $H_+ : \mu \geq \mu_+$ and a lower pair of hypotheses $H_{0-} : \mu \geq \mu_{0-}$ versus $H_- : \mu \leq \mu_-$. We note that in one-sided decision problems, we again may not try to distinguish two of the hypotheses.

For generality, we introduce the following notation for our decision rules

$$\begin{array}{lll} \text{reject } H_{0+} \text{ for } H_+ \text{ when} & Pr & (\mu > \mu_{0+} \,|\, X_1, \ldots) \geq \beta_u \\ \text{reject } H_{0-} \text{ for } H_- \text{ when} & Pr & (\mu < \mu_{0-} \,|\, X_1, \ldots) \geq \beta_\ell \\ \text{reject } H_+ \text{ for } H_{0+} \text{ when} & Pr & (\mu > \mu_+ \,|\, X_1, \ldots) \leq \alpha_u \\ \text{reject } H_- \text{ for } H_{0-} \text{ when} & Pr & (\mu < \mu_- \,|\, X_1, \ldots) \leq \alpha_\ell \end{array} \tag{2.4}$$

Under the above, we note that a decision is made for the null hypothesis only if both of the alternatives have been rejected. Due to the natural ordering of the hypotheses, whenever the null hypothesis has been rejected in favor H_+, the alternative H_- has also been rejected by at least the same criteria. Similar arguments hold for decisions in favor of H_-. In order to preserve the natural ordering of $\mu_+ \geq \mu_{0-} \geq \mu_{0+} \geq \mu_-$, we will again demand that the constraints specified by eqn (2.2) be satisfied.

In choosing the values of the alternatives μ_- and μ_+, parallels can be drawn to the frequentist approaches to sample size determination. That is, we assume that the outcome of the study must correspond to a decision for exactly one of the above hypotheses. In one approach, we choose μ_+ (respectively, μ_-) to correspond to differences in outcome which it is clinically important to distinguish. We then find the sample size which results in contiguous, nonoverlapping decision sets for H_{0+} and H_+ (respectively, H_{0-} and H_-). Because it is not always possible to have contiguous decision sets for all three hypotheses for arbitrary choices of α_u, α_ℓ,β_u, and β_ℓ, the value of μ_- (respectively, μ_+) must then be chosen to satisfy the probability constraints in eqn (2.4).

In a second approach, the available sample size is constrained, and we find the value of μ_+ and μ_- that satisfy the probability constraints in eqn (2.4).

3 Examples of Applications

The formulation of the fundamental model in section 1 applies directly to the case of a one sample test for the mean of a normal distribution estimated from independent, identically distributed observations. In fact, however, we can use the designs derived for this simple situation in a variety of other useful clinical trial settings. In particular, we can consider the following departures from the assumptions of the previous section.

1. Each random variable X_i, $i = 1, \ldots, N$ can represent a summary measure from a sampling unit. For instance, in a two sample study, we might choose to describe our probability model in terms of sampling units consisting of 1 subject sampled from population 1 and r subjects sampled from population 2.

2. Each random variable X_i, $i = 1, \ldots, N$ can have a distinct mean μ_i and a distinct variance σ_i^2. For instance, we may design a clinical trial in which the best measure of treatment effect is the slope of a linear dose-response relationship. The test statistic may be based on the efficient scores, in which case each observation corresponding to the efficient score would potentially have a different mean and variance (although presumably the mean of each observation would depend in some way on a common parameter).

3. The distribution of the random variables X_i, $i = 1, \ldots, N$ need not be normal. Because we are analyzing the data after groups of observations have been accrued, it will often be the case that a central limit theorem will guarantee that the increments of information $S_j - S_{j-1}$ are approximately normally distributed. We will find in section 5 that the sampling density for the group sequential statistic depends only upon the normal distribution for those increments, and thus our methods are valid whenever those increments are approximately normally distributed.

In the following we consider a variety of statistical models for which the group sequential methods described herein are valid. In their most general form, we consider original observations Y_i, $i = 1, \ldots, M$ with distributions depending on parameter of interest θ and potentially on nuisance parameters $(\gamma, \nu, \ldots)$ and covariates $\vec{W}_i$, $i = 1, \ldots, M$.

(It should be noted that the use of the notation M in this section refers to a sample size in an untransformed setting. In later sections, we will use M to denote a random variable measuring the analysis at which a clinical trial stopped. While such overlap of notation is undesirable, there should not truly be any ambiguity, as it is only in this section that M will denote the sample size.)

(It should be further noted that this document describes the definition of the sample size N used by the C code, which uses N to refer to count sampling units. S+SeqTrial defines the sample size according to the total sample requirements across all arms, and thus the value of M as used in this section corresponds to the output of S+SeqTrial.)

We assume that it is of interest to test a null hypothesis $H_0 : \theta = \theta_0$ using test statistic $T(\vec{Y})$, a function of the observations $\vec{Y} = (Y_1, \ldots, Y_M)$. We then relate this original model to observations X_i, $i = 1, \ldots, N$ made on N independent sampling units. The ith observation X_i has moments $E(X_i) = \mu_i$ and $Var(X_i) = \sigma_i^2$, with

$$\begin{aligned} \mu &= \frac{1}{N}\sum_{i=1}^{N} \mu_i \\ \sigma^2 &= \frac{1}{N}\sum_{i=1}^{N} \sigma_i^2 \end{aligned} \tag{3.1}$$

representing the average tendencies for the moments across the population sampled. We further consider a parameterization in which the most direct measure of treatment effect is measured by

parameter θ, and that μ is a function of that parameter

$$\mu = \psi g(\theta) \tag{3.2}$$

where ψ is some constant and $g(\cdot)$ is a link function used in the statistical model. Under this parameterization, we note that it will generally be of more interest to consider at the jth analysis a test statistic $\hat{\theta}_j$ corresponding to the maximum likelihood estimate of θ, rather than focusing on the maximum likelihood estimate $\overline{X}_j$ of μ. In section 1.7 we defined transformations between the estimate ($\hat{\theta}$) scale and the sample mean (X) scale, as well as the δ-, μ-, and θ-scales for the parameter space.

In each of the following models, we define the sampling unit, the observation X_i, the moments μ_i and σ_i^2, the averages μ and σ^2, the link function $g(\cdot)$, and the constant ψ. We also describe the most typical ways in which the hypotheses θ_0, θ_+, and θ_- might be specified. In addition, we describe the correspondences between the commonly used statistics in the original statistical model (as might be obtained from computer output) and the statistics $\overline{X}_j$, $\hat{\theta}_j$, Z_j, and P_j.

3.1 Normally Distributed Responses

When the treatment response is measured on a continuous scale, it is common to base statistical inference on the assumption that the underlying observations are normally distributed. It should be noted that these same models are valid for nonnormal data provided the group sizes are sufficiently large as to allow the central limit theorem to provide a good approximation. This may not be a good assumption when the distribution is markedly skewed, however in those cases the methods described below for log normal responses may work.

It should be noted that in all of the models presented in this section, it is typically the case that the variance σ^2 is unknown. Hence, rather than using statistics which have the normal distribution, one typically assumes a t distribution. In what follows, however, we will take the approach of using the usual estimate of the variance, but continuing to use the normal based methods. We note that such an approach is valid in large sample sizes. We also note that there is some evidence [11] to suggest that if the statistics P_j are taken from the t distribution, the small sample behavior of the group sequential methodology parallels that of the small sample behavior of the t test in that same data (which t test may also not be exact due to nonnormality of the underlying observations).

3.1.1 One Sample test of a Normal Mean

Suppose we sample independently from a population with $Y_i \sim \mathcal{N}(\theta, \nu^2)$ for $i = 1, \ldots, M$. We wish to test a null hypothesis about the population mean θ, $H_0 : \theta = \theta_0$, and in a fixed sample test we perform a one sample Z test using test statistic

$$T(\vec{Y}) = \sqrt{M}\frac{[\overline{Y}_M - \theta_0]}{\nu}.$$

Such a trial corresponds exactly to our fundamental model with a sampling unit corresponding to a single observation. Hence, $N = M$, and the observations $X_i = Y_i$ on those sampling units have moments $\mu_i = \theta$ and $\sigma_i^2 = \nu^2$, with averages $\mu = \theta$ and $\sigma^2 = \nu^2$. Under this parameterization, the link function $g(\cdot)$ is merely the identity function $g(\theta) = \theta$, and the constant $\psi = 1$. The null and alternative hypotheses are typically specified directly, with $\theta_0 = 0$ being the usual choice for the null hypothesis.

At the jth analysis, the statistic $\overline{X}_j$ is the sample mean of the first N_j observations, and $\hat{\theta}_j = \overline{X}_j$. In a typical situation, the variance ν^2 is not known, and one would typically use the sample variance as an estimate. For our purposes, we can usually assume sufficient sample sizes such that a reasonable approximate test is obtained by using as Z_j the one-sample t statistic. The

statistic P_j is the one-sided p value from such a test used to detect the alternative $H_+ : \theta > \theta_0$. If only a two-sided P value is provided by statistical software (as is quite often the case), then P_j is half the two-sided P value when $Z_j > 0$, and P_j is 1 minus half the two-sided P value when $Z_j < 0$. As discussed above, it might be more robust to use the t distribution rather than the standard normal distribution when the variance is unknown, hence it is probably easiest to use the P-scale when using this statistical model.

3.1.2 Two Sample Test of Normal Means

Suppose we sample independently from two populations: a treatment group with $Y_{1i} \sim \mathcal{N}(\gamma_1, \nu_1^2)$ for $i = 1, \ldots, M_1$ and a comparison group with $Y_{2i} \sim \mathcal{N}(\gamma_2, \nu_2^2)$ for $i = 1, \ldots, M_2$. We wish to test a null hypothesis about the difference in population means $\theta = \gamma_1 - \gamma_2$, $H_0 : \theta = \theta_0$, and in a fixed sample test we perform a two sample Z test using test statistic

$$T(\vec{Y}) = \frac{[\overline{Y}_{1M_1} - \overline{Y}_{2M_2}] - \theta_0}{\sqrt{\frac{\nu_1^2}{M_1} + \frac{\nu_2^2}{M_2}}}.$$

For notational convenience, we define $r = M_1/M_2$. Such a trial corresponds to our fundamental model with a sampling unit corresponding to a single observation from the comparison group, and r observations from the treatment group. Hence, $N = M_2$, and the observations $X_i = \sum_{k=ri-r+1}^{ri} Y_{1k}/r - Y_{2i}$ on those sampling units have moments $\mu_i = \theta$ and $\sigma_i^2 = \nu_1^2/r + \nu_2^2$, with averages $\mu = \theta$ and $\sigma^2 = \nu_1^2/r + \nu_2^2$. Under this parameterization, the link function $g(\cdot)$ is merely the identity function $g(\theta) = \theta$, and the constant $\psi = 1$. The total sample size required in the study (across both arms) is $[r + 1]N$.

In specifying the hypotheses, most often θ_0 and θ_+ or θ_- are specified directly. Typically, the null hypothesis is $\theta_0 = 0$. Alternative methods of specifying the hypotheses include: 1) specifying the values of γ_1 and γ_2 under the null hypothesis, and also the values of γ_1 and γ_2 under the alternative hypothesis, and 2) specifying γ_1 under each of the null and alternative hypotheses, and assuming that γ_2 under both the null and alternative is equal to what γ_1 is under the null. In allowing for alternative specifications of the hypotheses, it should be noted that it is easy to distinguish between the usual specification of θ_0 and θ_+ or θ_- and the first alternative based on the number of values given in the specification. Distinguishing between the usual specification and the second alternative is not possible by such means, but it is truly unimportant. Treating those two methods of specification the same will result in the exact same sample size calculation, because in each case $\theta_+ - \theta_0$ is the same value. When a computer interface reports values back to a user, those values may be $\theta = \gamma_1 - \gamma_2$ (in the case of the usual specification or γ_1 (in the case of the second alternative), and only the user need know which is which. In the case of the first alternative, it probably is most straightforward to convert the input to the corresponding values of θ and to report those values.

At the jth analysis (and assuming the ratio between the number of measurements from the first population and the number of measurements from the second population is r:1 for all j), the statistic $\overline{X}_j$ is the sample mean of the first N_j observations

$$\overline{X}_j = \frac{1}{rN_j} \sum_{i=1}^{rN_j} Y_{1i} - \frac{1}{N_j} \sum_{i=1}^{N_j} Y_{2i},$$

and $\hat{\theta}_j = \overline{X}_j$. In a typical situation, the variances ν_1^2 and ν_2^2 are not known, and one would typically use the sample variances as estimates. For our purposes, we can usually assume sufficient sample sizes such that a reasonable approximate test is obtained by using as Z_j the two-sample t statistic assuming unequal variances. The statistic P_j is the one-sided P value from such a test used to detect the alternative $H_+ : \theta > \theta_0$. If only a two-sided P value is provided by statistical software

(as is quite often the case), then P_j is half the two-sided P value when $Z_j > 0$, and P_j is 1 minus half the two-sided P value when $Z_j < 0$. As discussed above, it might be more robust to use the t distribution rather than the standard normal distribution when the variance is unknown, hence it is probably easiest to use the P-scale when using this statistical model.

It should be noted that although the above derivation assumes that the ratio between the number of measurements from the first population and the number of observations from the second population is r:1 at each analysis, the statistical behavior of the group sequential test is not substantially affected by slight deviations from that ratio across the different interim analyses. Hence at the design stage, it is sufficient to assume a constant value for r, and then when actually monitoring the study to use the observed r at each analysis. This is equivalent to just ignoring any variation in r across analysis times and using the value of P_j as defined above at each analysis.

Of course, major deviations in the distribution of sample sizes from the two populations across analysis times will affect the statistical behavior of the group sequential test when boundaries are determined solely on the basis of the number of sampling units accrued to date. This problem is alleviated for the most part when stopping boundaries are determined on the basis of the proportion of the planned maximal statistical information accrued to date. This aspect is discussed further in section 12.

3.1.3 Test of Linear Regression Slope

Suppose we sample paired observations (Y_i, W_i) for $i = 1, \ldots, M$, with response variable $Y_i \sim \mathcal{N}(\gamma_i, \nu^2)$ and where we assume a regression model based on covariates W_i as $\gamma_i = \alpha + \theta W_i$. We wish to test a null hypothesis about the linear slope θ, $H_0 : \theta = 0$, and in a fixed sample test we use as test statistic the t test based on the estimate of the slope $\hat{\theta} = \sum_{i=1}^{M} Y_i[W_i - \overline{W}] / \sum_{i=1}^{M} [W_i - \overline{W}]^2$ and its standard error $\hat{se}(\hat{\theta}) = \hat{\nu} / \sqrt{\sum_{i=1}^{M} [W_i - \overline{W}]^2}$, where $\hat{\nu}^2 = \sum_{i=1}^{M} [Y_i - \hat{\alpha} - \hat{\theta} W_i]^2 / [M - 2]$ is the estimated residual mean squared error. Thus we use test statistic

$$T(\vec{Y}) = \frac{\sum_{i=1}^{M} Y_i [W_i - \overline{W}]}{\hat{\nu} \sqrt{\sum_{i=1}^{M} [W_i - \overline{W}]^2}},$$

which has a t distribution with $M - 2$ degrees of freedom. For large M, this distribution is well approximated by a standard normal distribution.

Such a trial corresponds to our fundamental model with a sampling unit corresponding to a single observation. Hence, $N = M$, and the observations $X_i = Y_i[W_i - \overline{W}]$ on those sampling units have moments $\mu_i = [W_i - \overline{X}]\gamma_i$ and $\sigma_i^2 = [W_i - \overline{W}]^2 \nu^2$, with averages $\mu = \theta V_W$ and $\sigma^2 = \nu^2 V_W$, where $V_W = \sum_{i=1}^{M} [W_i - \overline{W}]^2 / M$ is the variance of the covariates. Under this parameterization, the link function $g(\cdot)$ is merely the identity function $g(\theta) = \theta$, and the constant $\psi = V_W$.

An alternative correspondence to our fundamental model again has a sampling unit corresponding to a single observation with $N = M$, but the observations on those sampling units are taken to be $X_i = Y_i[W_i - \overline{W}]/V_W$ having moments $\mu_i = [W_i - \overline{X}]\gamma_i / V_W$ and $\sigma_i^2 = [W_i - \overline{W}]^2 \nu^2 / V_W^2$, leading to averages $\mu = \theta$ and $\sigma^2 = \nu^2 / V_W$, where again $V_W = \sum_{i=1}^{M} [W_i - \overline{W}]^2 / M$ is the variance of the covariates. Under this parameterization, the link function $g(\cdot)$ is merely the identity function $g(\theta) = \theta$, and the constant $\psi = 1$.

In either correspondence to the fundamental model, the null and alternative hypotheses are typically specified directly, with $\theta_0 = 0$ being the usual choice for the null hypothesis.

At the jth analysis (and assuming the mean of the covariates is $\overline{W}$ and the variance of the covariates is V_W for all j), the statistic $\overline{X}_j$ is the sample mean of the first N_j observations. In the first correspondence, we then have

$$\overline{X}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} Y_i[W_i - \overline{W}],$$

and in the second correspondence, we have

$$\overline{X}_j = \frac{1}{N_j V_W} \sum_{i=1}^{N_j} Y_i[W_i - \overline{W}].$$

In either case we have that $\hat{\theta}_j$ is just the least squares estimate of the slope based on the first N_j observations. For our purposes, we can usually assume sufficient sample sizes such that a reasonable approximate test is obtained by using as Z_j the t statistic for the test of the slope. The statistic P_j is the one-sided P value from such a test used to detect the alternative $H_+ : \theta > 0$. If only a two-sided P value is provided by statistical software (as is generally the case), then P_j is half the two-sided P value when $Z_j > 0$, and P_j is 1 minus half the two-sided P value when $Z_j < 0$. As discussed above, it might be more robust to use the t distribution rather than the standard normal distribution when the variance is unknown, hence it is probably easiest to use the P-scale when using this statistical model.

It should be noted that although the above derivation assumes that the mean $\overline{W}$ and variance V_W for the covariates is constant at each analysis, the statistical behavior of the group sequential test is not substantially affected by slight deviations from those values across the different interim analyses. Hence at the design stage, it is sufficient to assume constant value for $\overline{W}$ and V_W, and then when actually monitoring the study to use the observed values at each analysis. This is equivalent to just ignoring any variation in $\overline{W}$ and V_W across analysis times and using the value of P_j as defined above at each analysis.

Of course, major deviations in the distribution of covariates across analysis times will affect the statistical behavior of the group sequential test when boundaries are determined solely on the basis of the number of sampling units accrued to date. This problem is alleviated for the most part when stopping boundaries are determined on the basis of the proportion of the planned maximal statistical information accrued to date. This aspect is discussed further in section 12.

We note that a two sample test of equality between means assuming equal variances is equivalent to a test of linear slope in a regression model when the covariates W_i are dichotomous. The same inference is obtained under either the regression model described here or the two sample model described in section 3.1.2 with $\nu_1^2 = \nu_2^2$.

3.1.4 Test of Equality of Means Among K Groups (ANOVA)

Suppose we sample independently from K populations with $Y_{ki} \sim \mathcal{N}(\gamma_k, \nu^2)$ for $i = 1, \ldots, M_k$ and $k = 1, \ldots, K$. For notational convenience, we define $M = \sum_{k=1}^{K} M_k$ and $\vec{r} = (r_1, \ldots, r_K)$, where $r_k = M_k/M$ is the proportion of the total sample size that is apportioned to the kth group. We wish to test a null hypothesis about the equality of population means $\gamma_1 = \gamma_2 = \cdots = \gamma_K$, and in a fixed sample test we perform a F test from a one-way analysis of variance (ANOVA).

Such a trial cannot be couched in our fundamental model. The F statistic is asymptotically distributed according to a chi square distribution as the sample sizes within every group gets large, and that distribution would be approximately normal only as the number of groups K approaches infinity. Hence, the methods we derive for group sequential tests will not apply directly to this statistical model. We do present here sample size formula that can apply to this setting for fixed sample trials. The null hypothesis is understood to be exact equality of the means in this model, and alternative hypothesis can be specified either by listing the K values γ_1, γ_2, $\ldots$, γ_K, or by providing the variance of the values for the γ's.

The sample size formula for the K-sample problem can be derived by considering the regression model in which dummy variables are fit for groups 2 through K, and then testing that those $K-1$ regression parameters are simultaneously equal to 0. If we assume that ν^2 is known, the test statistic has a noncentral $\chi^2_{K-1}(\delta^2)$ distribution with $K-1$ degrees of freedom and noncentrality parameter $\delta_2 = MV_\gamma/\nu^2$ where

$$V_\gamma = \left\{ \sum_{k=1}^{K} r_k \gamma_k^2 - \left[\sum_{k=1}^{K} r_k \gamma_k \right]^2 \right\} \tag{37}$$

is a weighted variance of the population means. Note that $\delta^2 = 0$ under H_0, and in that case the test statistic has a central χ^2 distribution.

In experimental design, we often desire to find a sample size which would under some specified alternative hypothesis supply prespecified power β to reject the null hypothesis when performing a level α ANOVA. To obtain a level α test, we compare the ANOVA test statistic to the critical value $\chi^2_{K-1}(0, 1-\alpha)$ which is the upper αth quantile of the central chi square distribution with $k-1$ degrees of freedom. The distribution of the test statistic depends on the alternative only through the value V_γ. Thus we need to find the value of M such that a random variable U having a noncentral chi square distribution with $K-1$ degrees of freedom and noncentrality parameter MV_γ/ν^2 satisfies

$$Pr(U > \chi^2_{K-1}(0, 1-\alpha)) = \beta.$$

In S-Plus this can be effected using the following code (where `alpha` $= \alpha$ is the size of the test, `beta` $= \beta$ is the desired power, `gamma` $= \vec{\gamma} = (\gamma_1, \ldots, \gamma_K)$ is the vector of population means under the alternative, `vrnc` $= \nu^2$ is the within group variance, and `r` $= \vec{r}$ is the vector of sample size proportions to be accrued to each group)

```
K <- length (gamma)
crit.value <- qchisq ( 1-alpha, K-1)
noncent <- (sum (r * gamma^2) - sum (r * gamma)^2 ) / vrnc
pwr <- 1 - pchisq (crit.value, K-1, (1:1000) * noncent)
group.sample.sizes <- r * (sum (pwr < 1 - beta) + 1)
```

The vector `group.sample.sizes` $= (M_1, \ldots, M_K)$.

3.2 Lognormal Responses

When dealing with continuous positive response variables that are heavily skewed, it is not uncommon to assume a lognormal distribution for those variables. Such an assumption is equivalent to assuming that the logarithm of the response variable has a normal distribution. A similar transformation of the data is also often used when the data exhibit a mean-variance relationship in which the variance of the response variable is proportional to the square of the mean, even though the response variable might not have the lognormal distribution. As noted above with the normal model, due to the central limit theorem, the inference based on assuming a normal model for log transformed response is fairly robust to departures from normality.

Analyses based on the normal model for the log transformed response can be viewed as inference based on the mean of the log response. Such does not have an easy interpretation on the original response scale. However, so long as the log response has a symmetric distribution (which is certainly satisfied by the normal model), the mean log response is the median log response. The median is easily back transformed to the original response scale. Hence, we consider these models based on log transformed response to be inference about the median and the median ratio. We note that inference could also be considered on the basis of the geometric mean and the ratio of geometric means of the distributions of the original response.

In the following, it is assumed that a user would want to work with the median of the original response variables, rather than working with summary measures of the transformed responses. This is the most natural model, however, in the event that it were desired to work on the transformed scale, the correspondences discussed in each subsection would hold exactly, with the exception that the parameter of interest would be the log median $\log(\theta)$ or the difference in log medians and the link function $g(\cdot)$ would be the identity function $g(\log\theta) = \log(\theta)$.

3.2.1 One Sample Test of a Lognormal Median

Suppose we sample independently from a population with $U_i = \log(Y_i) \sim \mathcal{N}(\log(\theta), \nu^2)$ for $i = 1, \ldots, M$. We wish to test a null hypothesis about the population median θ, $H_0 : \theta = \theta_0$, and in a fixed sample test we perform a one sample Z test using test statistic

$$T(\vec{Y}) = \sqrt{M}\frac{[\overline{U}_M - \log(\theta_0)]}{\nu}.$$

Such a trial corresponds exactly to the model described in section 2.1 for the transformed response and a transformed parameter of interest. Hence, $N = M$, and the observations $X_i = U_i = \log(Y_i)$ on those sampling units have moments $\mu_i = \log(\theta)$ and $\sigma_i^2 = \nu^2$, with averages $\mu = \log(\theta)$ and $\sigma^2 = \nu^2$. Under this parameterization, the link function $g(\cdot)$ is the logarithmic function $g(\theta) = \log(\theta)$, and the constant $\psi = 1$. The null and alternative hypotheses are typically specified directly, with $\theta_0 = 1$ being the usual choice for the null hypothesis.

At the jth analysis, the statistic $\overline{X}_j$ is the sample mean of the first N_j logarithmically transformed observations, and $\hat{\theta}_j = \exp(\overline{X}_j)$. In a typical situation, the variance ν^2 is not known, and one would typically use the sample variance as an estimate. For our purposes, we can usually assume sufficient sample sizes such that a reasonable approximate test is obtained by using as Z_j the one-sample t statistic on the log transformed observations. The statistic P_j is the one-sided p value from such a test used to detect the alternative $H_+ : \theta > \theta_0$. If only a two-sided P value is provided by statistical software (as is quite often the case), then P_j is half the two-sided P value when $Z_j > 0$, and P_j is 1 minus half the two-sided P value when $Z_j < 0$. As discussed above, it might be more robust to use the t distribution rather than the standard normal distribution when the variance is unknown, hence it is probably easiest to use the P-scale when using this statistical model.

3.2.2 Two Sample Test of Lognormal Medians

Suppose we sample independently from two populations: a treatment group with $U_{1i} = \log(Y_{1i}) \sim \mathcal{N}(\log(\gamma_1), \nu_1^2)$ for $i = 1, \ldots, M_1$ and a comparison group with $U_{2i} = \log(Y_{2i}) \sim \mathcal{N}(\log(\gamma_2), \nu_2^2)$ for $i = 1, \ldots, M_2$. We wish to test a null hypothesis about the ratio of population medians $\theta = \gamma_1/\gamma_2$, $H_0 : \theta = \theta_0$, and in a fixed sample test we perform a two sample Z test using test statistic

$$T(\vec{Y}) = \frac{[\overline{U}_{1M_1} - \overline{U}_{2M_2}] - \log(\theta_0)}{\sqrt{\frac{\nu_1^2}{M_1} + \frac{\nu_2^2}{M_2}}}.$$

For notational convenience, we define $r = M_1/M_2$. Such a trial corresponds to our fundamental model with a sampling unit corresponding to r observations from the treatment group (population 1), and a single observation from the comparison group (population 2). Hence, $N = M_2$, and the observations $X_i = \sum_{k=ri-r+1}^{ri} U_{1k}/r - U_{2i}$ on those sampling units have moments $\mu_i = \log(\theta)$ and $\sigma_i^2 = \nu_1^2/r + \nu_2^2$, with averages $\mu = \log(\theta)$ and $\sigma^2 = \nu_1^2/r + \nu_2^2$. Under this parameterization, the link function $g(\cdot)$ is the logarithmic function $g(\theta) = \log(\theta)$, and the constant $\psi = 1$. The total sample size required in the study (across both arms) is $[r + 1]N$.

In specifying the null hypothesis, most often θ_0 and θ_+ or θ_- are specified directly. Typically, the null hypothesis is $\theta_0 = 1$. Alternative methods of specifying the hypotheses include: 1) specifying the values of γ_1 and γ_2 (the medians of the respective distributions) under the null hypothesis, and also the values of γ_1 and γ_2 under the alternative hypothesis, and 2) specifying γ_1 under each of the null and alternative hypotheses, and assuming that γ_2 under both the null and alternative is equal to what γ_1 is under the null. In allowing for alternative specifications of the hypotheses, it should be noted that it is easy to distinguish between the usual specification of θ_0 and θ_+ or θ_- and the first alternative based on the number of values given in the specification. Distinguishing between the usual specification and the second alternative is not possible by such means, but it is truly unimportant. Treating those two methods of specification the same will result in the exact same sample size calculation, because in each case θ_+/θ_0 is the same value. When a computer interface reports values back to a user, those values may be $\theta = \gamma_1/\gamma_2$ (in the case of the usual specification or γ_1 (in the case of the second alternative), and only the user need know which is which. In the case of the first alternative, it probably is most straightforward to convert the input to the corresponding values of θ and to report those values.

At the jth analysis (and assuming the ratio between the number of measurements from the first population and the number of measurements from the second population is r:1 for all j), the statistic $\overline{X}_j$ is the sample mean of the first N_j logarithmically transformed observations

$$\overline{X}_j = \frac{1}{rN_j}\sum_{i=1}^{rN_j} U_{1i} - \frac{1}{N_j}\sum_{i=1}^{N_j} U_{2i},$$

and $\hat{\theta}_j = \exp(\overline{X}_j)$. In a typical situation, the variances ν_1^2 and ν_2^2 are not known, and one would typically use the sample variances as estimates. For our purposes, we can usually assume sufficient sample sizes such that a reasonable approximate test is obtained by using as Z_j the two-sample t statistic assuming unequal variances. The statistic P_j is the one-sided P value from such a test used to detect the alternative $H_+ : \theta > \theta_0$. If only a two-sided P value is provided by statistical software (as is quite often the case), then P_j is half the two-sided P value when $Z_j > 0$, and P_j is 1 minus half the two-sided P value when $Z_j < 0$. As discussed above, it might be more robust to use the t distribution rather than the standard normal distribution when the variance is unknown, hence it is probably easiest to use the P-scale when using this statistical model.

It should be noted that although the above derivation assumes that the ratio between the number of measurements from the first population and the number of observations from the second population is r:1 at each analysis, the statistical behavior of the group sequential test is not substantially affected by slight deviations from that ratio across the different interim analyses. Hence at the design stage, it is sufficient to assume a constant value for r, and then when actually monitoring the study to use the observed r at each analysis. This is equivalent to just ignoring any variation in r across analysis times and using the value of P_j as defined above at each analysis.

Of course, major deviations in the distribution of sample sizes from the two populations across analysis times will affect the statistical behavior of the group sequential test when boundaries are determined solely on the basis of the number of sampling units accrued to date. This problem is alleviated for the most part when stopping boundaries are determined on the basis of the proportion of the planned maximal statistical information accrued to date. This aspect is discussed further in section 12.

3.2.3 Test of Log Median Regression Slope

Suppose we sample paired observations (Y_i, W_i) for $i = 1, \ldots, M$, with transformed response variable $U_i = \log(Y_i) \sim \mathcal{N}(\gamma_i, \nu^2)$ and where we assume a regression model based on covariates W_i as $\gamma_i = \alpha + \log(\theta)W_i$. We wish to test a null hypothesis about the back transformed linear slope

θ (which has interpretation as the median ratio between groups differing by one unit in their covariate values), $H_0 : \theta = 1$, and in a fixed sample test we use as test statistic the t test based on the estimate of the slope $\log(\hat{\theta}) = \sum_{i=1}^{M} U_i[W_i - \overline{W}] / \sum_{i=1}^{M} [W_i - \overline{W}]^2$ and its standard error $\hat{se}(\log(\hat{\theta})) = \hat{\nu} / \sqrt{\sum_{i=1}^{M} [W_i - \overline{W}]^2}$, where $\hat{\nu}^2 = \sum_{i=1}^{M} [U_i - \hat{\alpha} - \log(\hat{\theta}) W_i]^2 / [M - 2]$ is the estimated residual mean squared error. Thus we use test statistic

$$T(\vec{Y}) = \frac{\sum_{i=1}^{M} U_i[W_i - \overline{W}]}{\hat{\nu} \sqrt{\sum_{i=1}^{M} [W_i - \overline{W}]^2}},$$

which has a t distribution with $M - 2$ degrees of freedom. For large M, this distribution is well approximated by a standard normal distribution.

Such a trial corresponds to our fundamental model with a sampling unit corresponding to a single observation. Hence, $N = M$, and the observations $X_i = U_i[W_i - \overline{W}]$ on those sampling units have moments $\mu_i = [W_i - \overline{X}]\gamma_i$ and $\sigma_i^2 = [W_i - \overline{W}]^2 \nu^2$, with averages $\mu = \log(\theta) V_W$ and $\sigma^2 = \nu^2 V_W$, where $V_W = \sum_{i=1}^{M} [W_i - \overline{W}]^2 / M$ is the variance of the covariates. Under this parameterization, the link function $g(\cdot)$ is the logarithmic function $g(\theta) = \log(\theta)$, and the constant $\psi = V_W$.

An alternative correspondence to our fundamental model again has a sampling unit corresponding to a single observation with $N = M$, but the observations on those sampling units are taken to be $X_i = U_i[W_i - \overline{W}]/V_W$ having moments $\mu_i = [W_i - \overline{X}]\gamma_i / V_W$ and $\sigma_i^2 = [W_i - \overline{W}]^2 \nu^2 / V_W^2$, leading to averages $\mu = \log(\theta)$ and $\sigma^2 = \nu^2 / V_W$, where again $V_W = \sum_{i=1}^{M} [W_i - \overline{W}]^2 / M$ is the variance of the covariates. Under this parameterization, the link function $g(\cdot)$ is the logarithmic function $g(\theta) = \log(\theta)$, and the constant $\psi = 1$.

In either correspondence to the fundamental model, the null and alternative hypotheses are typically specified directly, with $\theta_0 = 1$ being the usual choice for the null hypothesis.

At the jth analysis (and assuming the mean of the covariates is $\overline{W}$ and the variance of the covariates is V_W for all j), the statistic $\overline{X}_j$ is the sample mean of the first N_j observations. In the first correspondence, we then have

$$\overline{X}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} U_i[W_i - \overline{W}],$$

and in the second correspondence, we have

$$\overline{X}_j = \frac{1}{N_j V_W} \sum_{i=1}^{N_j} U_i[W_i - \overline{W}].$$

In either case we have that $\hat{\theta}_j$ is just the exponentiation of the least squares estimate of the slope based on the first N_j observations. For our purposes, we can usually assume sufficient sample sizes such that a reasonable approximate test is obtained by using as Z_j the t statistic for the test of the slope. The statistic P_j is the one-sided P value from such a test used to detect the alternative $H_+ : \theta > 0$. If only a two-sided P value is provided by statistical software (as is generally the case), then P_j is half the two-sided P value when $Z_j > 0$, and P_j is 1 minus half the two-sided P value when $Z_j < 0$. As discussed above, it might be more robust to use the t distribution rather than the standard normal distribution when the variance is unknown, hence it is probably easiest to use the P-scale when using this statistical model.

It should be noted that although the above derivation assumes that the mean $\overline{W}$ and variance V_W for the covariates is constant at each analysis, the statistical behavior of the group sequential test is not substantially affected by slight deviations from those values across the different interim analyses. Hence at the design stage, it is sufficient to assume constant value for $\overline{W}$ and V_W, and then when actually monitoring the study to use the observed values at each analysis. This is

equivalent to just ignoring any variation in $\overline{W}$ and V_W across analysis times and using the value of P_j as defined above at each analysis.

Of course, major deviations in the distribution of covariates across analysis times will affect the statistical behavior of the group sequential test when boundaries are determined solely on the basis of the number of sampling units accrued to date. This problem is alleviated for the most part when stopping boundaries are determined on the basis of the proportion of the planned maximal statistical information accrued to date. This aspect is discussed further in section 12.

We note that a two sample test of equality between medians assuming equal variances of the log transformed response (so variance proportional to the means for the original response) is equivalent to a test of linear slope in a regression model when the covariates W_i are dichotomous. The same inference is obtained under either the regression model described here or the two sample model described in section 3.2.2 with $\nu_1^2 = \nu_2^2$.

3.2.4 Test of Equality of Medians Among K Groups (ANOVA)

Suppose we sample independently from K populations with $U_{ki} = \log(Y_{ki}) \sim \mathcal{N}(\log(\gamma_k), \nu^2)$ for $i = 1, \ldots, M_k$ and $k = 1, \ldots, K$. For notational convenience, we define $M = \sum_{k=1}^{K} M_k$ and $\vec{r} = (r_1, \ldots, r_K)$, where $r_k = M_k/M$ is the proportion of the total sample size that is apportioned to the kth group. We wish to test a null hypothesis about the equality of population medians $\gamma_1 = \gamma_2 = \cdots = \gamma_K$, and in a fixed sample test we perform a F test from a one-way analysis of variance (ANOVA) on the log transformed responses.

Such a trial cannot be couched in our fundamental model. The F statistic is asymptotically distributed according to a chi square distribution as the sample sizes within every group gets large, and that distribution would be approximately normal only as the number of groups K approaches infinity. Hence, the methods we derive for group sequential tests will not apply directly to this statistical model. We do present here sample size formula that can apply to this setting for fixed sample trials. The null hypothesis is understood to be exact equality of the medians in this model, and the alternative hypothesis can be specified either by listing the K values γ_1, γ_2, ..., γ_K, or by providing the variance of the values for the $\log(\gamma)$'s.

The sample size formula for the K-sample problem can be derived by considering the regression model in which dummy variables are fit for groups 2 through K, and then testing that those $K-1$ regression parameters are simultaneously equal to 0. If we assume that ν^2 is known, the test statistic has a noncentral $\chi^2_{K-1}(\delta^2)$ distribution with $K-1$ degrees of freedom and noncentrality parameter $\delta_2 = MV_{\log(\gamma)}/\nu^2$ where

$$V_{\log(\gamma)} = \left\{ \sum_{k=1}^{K} r_k \log^2(\gamma_k) - \left[\sum_{k=1}^{K} r_k \log(\gamma_k) \right]^2 \right\} \tag{37}$$

is a weighted variance of the population means. Note that $\delta^2 = 0$ under H_0, and in that case the test statistic has a central χ^2 distribution.

In experimental design, we often desire to find a sample size which would under some specified alternative hypothesis supply prespecified power β to reject the null hypothesis when performing a level α ANOVA. To obtain a level α test, we compare the ANOVA test statistic to the critical value $\chi^2_{K-1}(0, 1-\alpha)$ which is the upper αth quantile of the central chi square distribution with $k-1$ degrees of freedom. The distribution of the test statistic depends on the alternative only through the value V_γ. Thus we need to find the value of M such that a random variable U having a noncentral chi square distribution with $K-1$ degrees of freedom and noncentrality parameter MV_γ/ν^2 satisfies

$$Pr(U > \chi^2_{K-1}(0, 1-\alpha)) = \beta.$$

In S-Plus this can be effected using the following code (where `alpha` $= \alpha$ is the size of the test, `beta` $= \beta$ is the desired power, `gamma` $= \vec{\gamma} = (\gamma_1, \ldots, \gamma_K)$ is the vector of population means under the alternative, `vrnc` $= \nu^2$ is the within group variance, and `r` $= \vec{r}$ is the vector of sample size proportions to be accrued to each group)

```
K <- length (gamma)
crit.value <- qchisq ( 1-alpha, K-1)
noncent <- (sum (r * log(gamma)^ 2) - sum (r * log(gamma))^ 2 ) / vrnc
pwr <- 1 - pchisq (crit.value, K-1, (1:1000) * noncent)
group.sample.sizes <- r * (sum (pwr < 1 - beta) + 1)
```

The vector `group.sample.sizes` $= (M_1, \ldots, M_K)$.

3.3 Dichotomous Responses

In many clinical trials, the outcome is measured on a binary scale: success or failure. The summary measure used to describe the probability distribution for the response variable is typically either the binomial proportion (the probability of success) or the binomial odds (the odds of success). Treatment effects are, respectively, summarized as the difference in binomial proportions or the odds ratio.

In this application, we assume that sample sizes are sufficiently large to allow inference based on the normal approximation to the binomial distribution.

3.3.1 One Sample Test of a Binomial Proportion

Suppose we have a random sample of independent Bernoulli random variables with $Y_i \sim \mathcal{B}(1, \theta)$ for $i = 1, \ldots, M$. We wish to test a null hypothesis about the population mean θ, $H_0 : \theta = \theta_0$, and in a fixed sample test we perform a one sample Z test using test statistic

$$T(\vec{Y}) = \sqrt{M} \frac{[\overline{Y}_M - \theta_0]}{\sqrt{\overline{Y}_M (1 - \overline{Y}_M)}}.$$

Such a trial corresponds exactly to our fundamental model with a sampling unit corresponding to a single observation. Hence, $N = M$, and the observations $X_i = Y_i$ on those sampling units have moments $\mu_i = \theta$ and $\sigma_i^2 = \theta(1 - \theta)$, with averages $\mu = \theta$ and $\sigma^2 = \theta(1 - \theta)$. Under this parameterization, the link function $g(\cdot)$ is merely the identity function $g(\theta) = \theta$, and the constant $\psi = 1$.

At the jth analysis, the statistic $\overline{X}_j$ is the sample mean of the first N_j observations, and $\hat{\theta}_j = \overline{X}_j$. In a typical situation, the variance $\theta(1 - \theta)$ is not known, and one would typically use either the variance under the null hypothesis or, more usually, the maximum likelihood estimate $\hat{\theta}_j(1 - \hat{\theta}_j)$ as an estimate. The test statistic Z_j is just the test statistic for a one sample test of a binomial proportion as given above, and the statistic P_j is the one-sided P value from such a test used to detect the alternative $H_+ : \theta > \theta_0$. If only a two-sided P value is provided by statistical software (as is quite often the case), then P_j is half the two-sided P value when $Z_j > 0$, and P_j is 1 minus half the two-sided P value when $Z_j < 0$.

3.3.2 Two Sample Test of Binomial Proportions

Suppose we sample independently from two populations: a treatment group with $Y_{1i} \sim \mathcal{B}(1, \gamma_1)$ for $i = 1, \ldots, M_1$ and a comparison group with $Y_{2i} \sim \mathcal{B}(1, \gamma_2)$ for $i = 1, \ldots, M_2$. For notational convenience, we define $r = M_1/M_2$. We wish to test a null hypothesis about the difference in population probabilities of success $\theta = \gamma_1 - \gamma_2$, $H_0 : \theta = \theta_0$, and in a fixed sample test we perform a two sample Z test using test statistic

$$T(\vec{Y}) = \frac{[\overline{Y}_{1M_1} - \overline{Y}_{2M_2}] - \theta_0}{\sqrt{\overline{Y}_{\cdot M}[1 - \overline{Y}_{\cdot M}]\left[\frac{1}{M_1} + \frac{1}{M_2}\right]}}.$$

Such a trial corresponds to our fundamental model with a sampling unit corresponding to a single observation from the comparison group (population 2), and r observations from the treatment group (population 1). Hence, $N = M_2$, and the observations $X_i = \sum_{k=ri-r+1}^{ri} Y_{1k}/r - Y_{2i}$ on those sampling units have moments $\mu_i = \theta$ and $\sigma_i^2 = \gamma_1[1-\gamma_1]/r + \gamma_2[1-\gamma_2]$, with averages $\mu = \theta$ and $\sigma^2 = \gamma_1[1-\gamma_1]/r + \gamma_2[1-\gamma_2]$. Under this parameterization, the link function $g(\cdot)$ is merely the identity function $g(\theta) = \theta$, and the constant $\psi = 1$.

At the jth analysis (and assuming the ratio between the number of measurements from the treatment group and the number of measurements from the comparison group is r:1 for all j), the statistic $\overline{X}_j$ is the sample mean of the first N_j observations

$$\overline{X}_j = \frac{1}{rN_j}\sum_{i=1}^{rN_j} Y_{1i} - \frac{1}{N_j}\sum_{i=1}^{N_j} Y_{2i},$$

and $\hat{\theta}_j = \overline{X}_j$. In a typical situation, the variance must be estimated because γ_1^2 and γ_2^2 are not known, and one would typically use the sample means $\overline{Y}_{1j}$ and $\overline{Y}_{2j}$ as estimates for γ_1 and γ_2, respectively. The statistic Z_j is the Z test statistic comparing two binomial proportions, as given above. This is the signed square root of Pearson's chi square statistic. The statistic P_j is the one-sided P value from the Z test used to detect the alternative $H_+ : \theta > \theta_0$. If only a two-sided P value is provided by statistical software (as is quite often the case), then P_j is half the two-sided P value when $Z_j > 0$, and P_j is 1 minus half the two-sided P value when $Z_j < 0$.

It should be noted that although the above derivation assumes that the ratio between the number of measurements from the first population and the number of observations from the second population is r:1 at each analysis, the statistical behavior of the group sequential test is not substantially affected by slight deviations from that ratio across the different interim analyses. Hence at the design stage, it is sufficient to assume a constant value for r, and then when actually monitoring the study to use the observed r at each analysis. This is equivalent to just ignoring any variation in r across analysis times and using the value of P_j as defined above at each analysis.

Of course, major deviations in the distribution of sample sizes from the two populations across analysis times will affect the statistical behavior of the group sequential test when boundaries are determined solely on the basis of the number of sampling units accrued to date. This problem is alleviated for the most part when stopping boundaries are determined on the basis of the proportion of the planned maximal statistical information accrued to date. This aspect is discussed further in section 12.

3.3.3 One Sample Test of Binomial Odds

Suppose we have a random sample of independent Bernoulli random variables with $Y_i \sim \mathcal{B}(1, \gamma)$ for $i = 1, \ldots, M$. We wish to test a null hypothesis about the population odds of success $\theta = \gamma/[1-\gamma]$, $H_0 : \theta = \theta_0$. One possible approach in a fixed sample test is to use the score test for the intercept in a logistic regression model having no covariates. We thus perform a one sample Z test using test statistic

$$T(\vec{Y}) = \sqrt{M}\frac{\{\overline{Y}_M - e^{\theta_0}/[1+e^{\theta_0}]\}}{\sqrt{\overline{Y}_M(1-\overline{Y}_M)}}.$$

Such a trial corresponds exactly to our fundamental model with a sampling unit corresponding to a single observation. Hence, $N = M$, and the observations $X_i = Y_i$ on those sampling units have

moments $\mu_i = e^\theta/[1+e^\theta]$ and $\sigma_i^2 = \gamma(1-\gamma)$, with averages $\mu = e^\theta/[1+e^\theta]$ and $\sigma^2 = \gamma[1-\gamma]$. Under this parameterization, the link function $g(\cdot)$ is the function $g(\theta) = e^\theta/[1+e^\theta]$, and the constant $\psi = 1$.

At the jth analysis, the statistic $\overline{X}_j$ is the sample mean of the first N_j observations, and $\hat{\theta}_j = \overline{X}_j/[1-\overline{X}_j]$. In a typical situation, the variance $\gamma(1-\gamma)$ is not known, and one would typically use either the variance under the null hypothesis or, more usually, the maximum likelihood estimate $\overline{X}_j[1-\overline{X}_j]$ as an estimate. The test statistic Z_j is just the test statistic for a one sample test of a binomial proportion as given above, and the statistic P_j is the one-sided P value from such a test used to detect the alternative $H_+ : \theta > \theta_0$. If only a two-sided P value is provided by statistical software (as is quite often the case), then P_j is half the two-sided P value when $Z_j > 0$, and P_j is 1 minus half the two-sided P value when $Z_j < 0$.

An alternative test statistic in the fixed sample case can be based on the Wald test of the intercept from a logistic regression model having no covariates. This corresponds to a test based on a logarithmic transformation of the maximum likelihood estimate of the odds, with a standard error derived using the delta method. Hence, in a fixed sample setting we might use test statistic

$$T(\vec{Y}) = \sqrt{M}\frac{\log(\overline{Y}_M/[1-\overline{Y}_M]) - \log(\theta_0)}{\sqrt{\frac{1}{\overline{Y}_M(1-\overline{Y}_M)}}}.$$

Such a trial corresponds exactly to our fundamental model with a sampling unit corresponding to a single observation. Hence, $N = M$, but the observations X_i are not easily characterized. We can, however, define averages $\mu = \log(\theta)$ and $\sigma^2 = 1/\{\gamma[1-\gamma]\}$. Under this parameterization, the link function $g(\cdot)$ is the function $g(\theta) = \log(\theta)$, and the constant $\psi = 1$.

At the jth analysis, the statistic $\overline{X}_j$ is the estimate of the intercept in a logistic regression model having no covariates based on the first N_j observations, and $\hat{\theta}_j = e^{\overline{X}_j}$. In a typical situation, the variance $1/\{\gamma[1-\gamma]\}$ is not known, and one would typically use either the variance under the null hypothesis or, more usually, the maximum likelihood estimate $1/\{\overline{X}_j[1-\overline{X}_j]\}$ as an estimate. The test statistic Z_j is just the test statistic for the intercept in the logistic regression model described above, and the statistic P_j is the one-sided p value from such a test used to detect the alternative $H_+ : \theta > \theta_0$. If only a two-sided P value is provided by statistical software (as is quite often the case), then P_j is half the two-sided P value when $Z_j > 0$, and P_j is 1 minus half the two-sided P value when $Z_j < 0$.

3.3.4 Two Sample Test of Binomial Ratio

Suppose we sample independently from two populations: a treatment group with $Y_{1i} \sim \mathcal{B}(1, \gamma_1)$ for $i = 1, \ldots, M_1$ and a comparison group with $Y_{2i} \sim \mathcal{B}(1, \gamma_2)$ for $i = 1, \ldots, M_2$. For notational convenience, we define $r = M_1/M_2$. We wish to test a null hypothesis about the ratio of odds of success in the two populations $\theta = \gamma_1[1-\gamma_2]/\{\gamma_2[1-\gamma_1]\}$, $H_0 : \theta = 1$, and in a fixed sample test we perform the score test from a logistic regression model with a dichotomous covariate. Such a test is equivalent to the signed square root of Pearson's chi square test and the two sample Z test of binomial proportions, with test statistic

$$T(\vec{Y}) = \frac{[\overline{Y}_{1M_1} - \overline{Y}_{2M_2}] - \theta_0}{\sqrt{\overline{Y}_{\cdot M}[1-\overline{Y}_{\cdot M}]\left[\frac{1}{M_1} + \frac{2}{M_2}\right]}}.$$

Such a trial corresponds exactly to the model specified in section 3.3.2 above.

An alternative test could be based on the asymptotic distribution of the log odds ratio $\log(\hat{\theta}) = \log(\hat{\gamma}_1[1-\hat{\gamma}_2]/\{\hat{\gamma}_2[1-\hat{\gamma}_1]\})$, where $\hat{\gamma}_1 = \overline{Y}_{1M_1}$ and $\hat{\gamma}_2 = \overline{Y}_{2M_2}$. The statistic is thus the Wald test of the slope parameter in a logistic regression with binary predictor. This corresponds to our

fundamental model with a sampling unit corresponding to a single observation from the comparison group, and r observations from the treatment group. Hence, $N = M_2$, and the average moments of our observations are $X_i = Y_{1i} - \sum_{k=ri-r+1}^{ri} Y_{2k}/r$ on those sampling units $\mu = \log(\theta)$ and $\sigma^2 = 1/\{r\gamma_1[1-\gamma_1]\} + 1/\{\gamma_2[1-\gamma_2]\}$. Under this parameterization, the link function $g(\cdot)$ is the log function $g(\theta) = \log(\theta)$, and the constant $\psi = 1$.

At the jth analysis (and assuming the ratio between the number of measurements from the first population and the number of measurements from the second population is r:1 for all j), the statistic $\overline{X}_j = \log(\hat{\theta}_j)$ is the log odds ratio estimate based on the first N_j observations. In a typical situation, the variance must be estimated because γ_1 and γ_2 are not known, and one would typically use the sample means $\overline{Y}_{1j}$ and $\overline{Y}_{2j}$ as estimates for γ_1 and γ_2, respectively. The statistic Z_j is the Z test statistic from a logistic regression analysis. The statistic P_j is the one-sided P value from the Z test used to detect the alternative $H_+ : \theta > \theta_0$. If only a two-sided P value is provided by statistical software (as is quite often the case), then P_j is half the two-sided P value when $Z_j > 0$, and P_j is 1 minus half the two-sided P value when $Z_j < 0$.

It should be noted that although the above derivation assumes that the ratio between the number of measurements from the first population and the number of observations from the second population is r:1 at each analysis, the statistical behavior of the group sequential test is not substantially affected by slight deviations from that ratio across the different interim analyses. Hence at the design stage, it is sufficient to assume a constant value for r, and then when actually monitoring the study to use the observed r at each analysis. This is equivalent to just ignoring any variation in r across analysis times and using the value of P_j as defined above at each analysis.

Of course, major deviations in the distribution of sample sizes from the two populations across analysis times will affect the statistical behavior of the group sequential test when boundaries are determined solely on the basis of the number of sampling units accrued to date. This problem is alleviated for the most part when stopping boundaries are determined on the basis of the proportion of the planned maximal statistical information accrued to date. This aspect is discussed further in section 12.

3.3.5 Test of Logistic Regression Slope

Suppose we sample paired observations (Y_i, W_i) for $i = 1, \ldots, M$, with binary response variable $Y_i \sim \mathcal{B}(1, \gamma_i)$ and where we assume a regression model based on covariates W_i as $\log(\gamma_i/[1-\gamma_i]) = \alpha + \beta W_i$, with $\theta = e^{\beta}$. We wish to test a null hypothesis about the odds ratio θ comparing two populations which differ by one unit in their value of W, $H_0 : \theta = 1$, and in a fixed sample test we use as test statistic the Wald test of the slope parameter in a logistic regression, which statistic is asymptotically normally distributed.

Such a trial corresponds to our fundamental model with a sampling unit corresponding to a single observation, hence, $N = M$. The specification of the individual observations can be based on weighted versions of the efficient scores from the logistic regression model. The average moments from this model will be

$$\mu = \beta = \log(\theta) \quad \text{and} \quad \sigma^2 = \frac{\sum_{i=1}^{M} \gamma_i[1-\gamma_i]}{\sum_{i=1}^{M} \gamma_i[1-\gamma_i] \sum_{i=1}^{M} \gamma_i[1-\gamma_i]W_i^2 - \{\sum_{i=1}^{M} \gamma_i[1-\gamma_i]W_i\}^2}.$$

Under this parameterization, the link function $g(\cdot)$ is the log function $g(\theta) = \log(\theta)$, and the constant $\psi = 1$.

The null and alternative hypotheses are typically specified directly, with $\theta_0 = 1$ being the usual choice for the null hypothesis.

At the jth analysis (and assuming the mean of the covariates is $\overline{W}$ and the weighted variance of the covariates is constant for all j), the statistic $\overline{X}_j = \hat{\beta}$ is the estimate of the slope from logistic

regression in an analysis of the first N_j observations. For our purposes, we can usually assume sufficient sample sizes such that a reasonable approximate test is obtained by using as Z_j the Z statistic for the test of the slope. The statistic P_j is the one-sided P value from such a test used to detect the alternative $H_+ : \theta > 0$. If only a two-sided P value is provided by statistical software (as is generally the case), then P_j is half the two-sided P value when $Z_j > 0$, and P_j is 1 minus half the two-sided P value when $Z_j < 0$. As discussed above, it might be more robust to use the t distribution rather than the standard normal distribution when the variance is unknown, hence it is probably easiest to use the P-scale when using this statistical model.

It should be noted that although the above derivation assumes that the mean $\overline{W}$ and weighted variance or the covariates is constant at each analysis, the statistical behavior of the group sequential test is not substantially affected by slight deviations from those values across the different interim analyses. Hence at the design stage, it is sufficient to assume constant values, and then when actually monitoring the study to use the observed values at each analysis. This is equivalent to just ignoring any variation in $\overline{W}$ and the weighted variance of the covariance across analysis times and using the value of P_j as defined above at each analysis.

Of course, major deviations in the distribution of covariates across analysis times will affect the statistical behavior of the group sequential test when boundaries are determined solely on the basis of the number of sampling units accrued to date. This problem is alleviated for the most part when stopping boundaries are determined on the basis of the proportion of the planned maximal statistical information accrued to date. This aspect is discussed further in section 12.

3.4 Poisson Response

In some clinical trials, the outcome counts the number of events occurring over some period of time, some prescribed space, or a combination of the two. In such a setting, a common probability model is to assume those counts are distributed according to the Poisson distribution, with a summary measure based on the event rate. Comparisons across groups can be based on differences in the event rates or based on ratios of the event rates. Only the models based on the multiplicative measures (rate ratios) are implemented in S+SeqTrial.

In this application, we assume that sample sizes are sufficiently large to allow inference based on the normal approximation to the Poisson distribution.

3.4.1 One Sample Test of a Poisson Event Rate (Additive Model)

Suppose we have a random sample of independent Poisson random variables with $Y_i \sim \mathcal{P}(\theta t_i)$ for $i = 1, \ldots, M$. We wish to test a null hypothesis about the population mean event rate θ, $H_0 : \theta = \theta_0$, and in a fixed sample test we perform a one sample Z test using test statistic

$$T(\vec{Y}) = \sqrt{\sum_{i=1}^{M} t_i} \frac{[\hat{\theta}_M - \theta_0]}{\sqrt{\theta_0}}$$

where $\hat{\theta}_M = \sum_{i=1}^{M} Y_i / \sum_{i=1}^{M} t_i$.

Such a trial corresponds exactly to our fundamental model with a sampling unit corresponding to the observation of a single unit of time. Hence, $N = \sum_{i=1}^{M} t_i$ counts the subject time accrued to the study. The average moments of the observations are thus $\mu = \theta$ and $\sigma^2 = \theta$. Under this parameterization, the link function $g(\cdot)$ is merely the identity function $g(\theta) = \theta$, and the constant $\psi = 1$. In such a study, the actual number of subjects to be accrued would be computed as the value of N divided by the average time of observation $\overline{t}$, that is $M = N/\overline{t}$.

At the jth analysis, the statistic $\overline{X}_j$ is the estimated mean event rate $\hat{\theta}_j$ computed based on the observations during the first N_j of study time. In a typical situation, the variance θ is not known, and one would typically use either the variance under the null hypothesis or, more usually, the

maximum likelihood estimate $\hat{\theta}_j$ as an estimate. The test statistic Z_j is just the test statistic for a one sample test of a Poisson rate as given above, and the statistic P_j is the one-sided P value from such a test used to detect the alternative $H_+ : \theta > \theta_0$. If only a two-sided P value is provided by statistical software (as is quite often the case), then P_j is half the two-sided P value when $Z_j > 0$, and P_j is 1 minus half the two-sided P value when $Z_j < 0$.

S+SeqTrial does not implement this probability model directly.

3.4.2 Two Sample Test of Difference in Poisson Event Rates (Additive Model)

Suppose we sample independently from two populations: a treatment group with $Y_{1i} \sim \mathcal{P}(\gamma_1 t_{1i})$ for $i = 1, \ldots, M_1$ and a comparison group with $Y_{2i} \sim \mathcal{P}(\gamma_2 t_{2i})$ for $i = 1, \ldots, M_2$. For notational convenience, we define $r = \sum_{i=1}^{M_1} t_{1i} / \sum_{i=1}^{M_2} t_{2i}$. We wish to test a null hypothesis about the difference in population event rates $\theta = \gamma_1 - \gamma_2$, $H_0 : \theta = \theta_0$, and in a fixed sample test we perform a two sample Z test using test statistic

$$T(\vec{Y}) = \frac{[\hat{\gamma}_{1M_1} - \hat{\gamma}_{2M_2}] - \theta_0}{\sqrt{\hat{\gamma}_{\cdot M}\left[\frac{1}{\sum_{i=1}^{M_1} t_{1i}} + \frac{1}{\sum_{i=1}^{M_1} t_{1i}}\right]}},$$

where $\hat{\gamma}_{\ell M_\ell} = \sum_{i=1}^{M_\ell} Y_{\ell i} / \sum_{i=1}^{M_\ell} t_{\ell i}$ for $\ell = 1, 2$ and $\hat{\gamma}_{\cdot M} = [\sum_{i=1}^{M_1} Y_{1i} + \sum_{i=1}^{M_2} Y_{2i} / [\sum_{i=1}^{M_1} t_{1i} + \sum_{i=1}^{M_2} t_{2i}]$.

Such a trial corresponds exactly to our fundamental model with a sampling unit corresponding to the observation of a single unit of time on the comparison arm and r units of time on the treatment arm. Hence, $N = \sum_{i=1}^{M_2} t_{2i}$ counts the subject time accrued to the study. The average moments of the observations are thus $\mu = \theta$ and $\sigma^2 = \gamma_1 / r + \gamma_2$. Under this parameterization, the link function $g(\cdot)$ is merely the identity function $g(\theta) = \theta$, and the constant $\psi = 1$. In such a study, the total observation time for both arms is $[r + 1]N$, and the actual number of subjects to be accrued would be computed as the total observation time divided by the average time of observation $\overline{t} = \sum_{\ell=1}^{2} \sum_{i=1}^{M_\ell} [t_{\ell i} / M_\ell]$, that is $M = [r + 1]N / \overline{t}$.

At the jth analysis (and assuming the ratio between the observation time from the treatment group and from the comparison group is r:1 for all j), the statistic $\overline{X}_j$ is the difference in the estimated mean event rates $\hat{\theta}_j = \hat{\gamma}_{1j} - \hat{\gamma}_{2j}$ computed based on the first N_j of study time on the comparison arm and the first rN_j of study time on the treatment arm. In a typical situation, the variance σ^2 is not known, and one would typically use the maximum likelihood estimates $\hat{\gamma}_{1j}$ and $\hat{\gamma}_{2j}$ in estimating $\hat{\sigma}^2$. The statistic Z_j is the Z test statistic comparing two Poisson rates, as given above. The statistic P_j is the one-sided P value from the Z test used to detect the alternative $H_+ : \theta > \theta_0$. If only a two-sided P value is provided by statistical software (as is quite often the case), then P_j is half the two-sided P value when $Z_j > 0$, and P_j is 1 minus half the two-sided P value when $Z_j < 0$.

It should be noted that although the above derivation assumes that the ratio between the observation time from the first population and the observation time from the second population is r:1 at each analysis, the statistical behavior of the group sequential test is not substantially affected by slight deviations from that ratio across the different interim analyses. Hence at the design stage, it is sufficient to assume a constant value for r, and then when actually monitoring the study to use the observed r at each analysis. This is equivalent to just ignoring any variation in r across analysis times and using the value of P_j as defined above at each analysis.

Of course, major deviations in the distribution of sample sizes from the two populations across analysis times will affect the statistical behavior of the group sequential test when boundaries are determined solely on the basis of the number of sampling units accrued to date. This problem is alleviated for the most part when stopping boundaries are determined on the basis of the proportion of the planned maximal statistical information accrued to date. This aspect is discussed further in section 12.

S+SeqTrial does not implement this probability model directly.

3.4.3 One Sample Test of Poisson Event Rates (Multiplicative Model)

Suppose we have a random sample of independent Poisson random variables with $Y_i \sim \mathcal{P}(\theta t_i)$ for $i = 1, \ldots, M$. We wish to test a null hypothesis about the population mean event rate θ, $H_0 : \theta = \theta_0$. In a fixed sample test we might use the score test for the intercept in a Poisson regression model having no covariates, in which case the test statistic is the same as described in section 3.4.1.

An alternative test statistic in the fixed sample case can be based on the Wald test of the intercept from a Poisson regression model having no covariates. This corresponds to a test based on a logarithmic transformation of the maximum likelihood estimate of the event rate, with a standard error derived using the delta method. Hence, in a fixed sample setting we might use test statistic

$$T(\vec{Y}) = \sqrt{\sum_{i=1}^{M} t_i} \frac{[\log(\hat{\theta}_M) - \log(\theta_0)]}{\sqrt{\frac{1}{\hat{\theta}_M}}},$$

where $\hat{\theta}_M = \sum_{i=1}^{M} Y_i / \sum_{i=1}^{M} t_i$.

Such a trial corresponds exactly to our fundamental model with a sampling unit corresponding to the observation of a single unit of time. Hence, $N = \sum_{i=1}^{M} t_i$ counts the subject time accrued to the study. The average moments of the observations are thus $\mu = \log(\theta)$ and $\sigma^2 = 1/\theta$. Under this parameterization, the link function $g(\cdot)$ is the log function $g(\theta) = \log(\theta)$, and the constant $\psi = 1$. In such a study, the actual number of subjects to be accrued would be computed as the value of N divided by the average time of observation $\overline{t}$, that is $M = N/\overline{t}$.

At the jth analysis, the statistic $\overline{X}_j$ is the estimate of the intercept in a Poisson regression model having no covariates based on the first N_j observation time, and $\hat{\theta}_j = e^{\overline{X}_j}$. In a typical situation, the variance $1/\theta$ is not known, and one would typically use either the variance under the null hypothesis or, more usually, the maximum likelihood estimate $1/\hat{\theta}_j$ as an estimate. The test statistic Z_j is just the test statistic for the intercept in the Poisson regression model described above, and the statistic P_j is the one-sided p value from such a test used to detect the alternative $H_+ : \theta > \theta_0$. If only a two-sided P value is provided by statistical software (as is quite often the case), then P_j is half the two-sided P value when $Z_j > 0$, and P_j is 1 minus half the two-sided P value when $Z_j < 0$.

3.4.4 Two Sample Test of Poisson Event Rate Ratio (Multiplicative Model)

Suppose we sample independently from two populations: a treatment group with $Y_{1i} \sim \mathcal{P}(\gamma_1 t_{1i})$ for $i = 1, \ldots, M_1$ and a comparison group with $Y_{2i} \sim \mathcal{P}(\gamma_2 t_{2i})$ for $i = 1, \ldots, M_2$. For notational convenience, we define $r = \sum_{i=1}^{M_1} t_{1i} / \sum_{i=1}^{M_2} t_{2i}$. We wish to test a null hypothesis about the ratio in population event rates $\theta = \gamma_1/\gamma_2$, $H_0 : \theta = \theta_0$. In a fixed sample test we might use the score test for the slope in a Poisson regression model having a binary covariate, in which case the test statistic is the same as described in section 3.4.2.

An alternative test could be based on the asymptotic distribution of the log event rate ratio $\log(\hat{\theta}) = \log(\hat{\gamma}_1/\hat{\gamma}_2$, where $\hat{\gamma}_1 = \overline{Y}_{1M_1}$ and $\hat{\gamma}_2 = \overline{Y}_{2M_2}$. The statistic is thus the Wald test of the slope parameter in a Poisson regression with binary predictor.

Such a trial corresponds exactly to our fundamental model with a sampling unit corresponding to the observation of a single unit of time on the comparison arm and r units of time on the treatment arm. Hence, $N = \sum_{i=1}^{M_2} t_{2i}$ counts the subject time accrued to the study. The average moments of the observations are thus $\mu = \log(\theta)$ and $\sigma^2 = 1/[r\gamma_1] + 1/\gamma_2$. Under this parameterization,

the link function $g(\cdot)$ is the log function $g(\theta) = \log(\theta)$, and the constant $\psi = 1$. In such a study, the total observation time for both arms is $[r+1]N$, and the actual number of subjects to be accrued would be computed as the total observation time divided by the average time of observation $\overline{t} = \sum_{\ell=1}^{2} \sum_{i=1}^{M_\ell} [t_{\ell i}/M_\ell]$, that is $M = [r+1]N/\overline{t}$.

At the jth analysis (and assuming the ratio between the number of measurements from the first population and the number of measurements from the second population is r:1 for all j), the statistic $\overline{X}_j = \log(\hat{\theta}_j)$ is the log event rate ratio estimate based on the first N_j observations. In a typical situation, the variance must be estimated because γ_1 and γ_2 are not known, and one would typically use the maximum likelihood estimates $\hat{\gamma}_1$ and $\hat{\gamma}_2$, respectively. The statistic Z_j is the Z test statistic from a Poisson regression analysis. The statistic P_j is the one-sided P value from the Z test used to detect the alternative $H_+ : \theta > \theta_0$. If only a two-sided P value is provided by statistical software (as is quite often the case), then P_j is half the two-sided P value when $Z_j > 0$, and P_j is 1 minus half the two-sided P value when $Z_j < 0$.

It should be noted that although the above derivation assumes that the ratio between the observation time from the first population and observation time from the second population is r:1 at each analysis, the statistical behavior of the group sequential test is not substantially affected by slight deviations from that ratio across the different interim analyses. Hence at the design stage, it is sufficient to assume a constant value for r, and then when actually monitoring the study to use the observed r at each analysis. This is equivalent to just ignoring any variation in r across analysis times and using the value of P_j as defined above at each analysis.

Of course, major deviations in the distribution of observation times from the two populations across analysis times will affect the statistical behavior of the group sequential test when boundaries are determined solely on the basis of the number of sampling units accrued to date. This problem is alleviated for the most part when stopping boundaries are determined on the basis of the proportion of the planned maximal statistical information accrued to date. This aspect is discussed further in section 12.

3.4.5 Test of Poisson Regression Slope

Suppose we sample paired observations (Y_i, W_i) for $i = 1, \ldots, M$, with count response variable $Y_i \sim \mathcal{P}(\gamma_i)$ and where we assume a regression model based on covariates W_i as $\log(\gamma_i) = \alpha + \beta W_i$, with $\theta = e^{\beta}$. We wish to test a null hypothesis about the event rate ratio θ comparing two populations which differ by one unit in their value of W, $H_0 : \theta = 1$, and in a fixed sample test we use as test statistic the Wald test of the slope parameter in a Poisson regression, which statistic is asymptotically normally distributed.

Such a trial corresponds to our fundamental model with a sampling unit corresponding to a single observation, hence, $N = M$. The specification of the individual observations can be based on weighted versions of the efficient scores from the Poisson regression model. The average moments from this model will be

$$\mu = \beta = \log(\theta) \quad \text{and} \quad \sigma^2 = \frac{\sum_{i=1}^{M} \gamma_i}{\sum_{i=1}^{M} \gamma_i \sum_{i=1}^{M} \gamma_i W_i^2 - [\sum_{i=1}^{M} \gamma_i W_i]^2}.$$

Under this parameterization, the link function $g(\cdot)$ is the log function $g(\theta) = \log(\theta)$, and the constant $\psi = 1$.

The null and alternative hypotheses are typically specified directly, with $\theta_0 = 1$ being the usual choice for the null hypothesis.

At the jth analysis (and assuming the mean of the covariates is $\overline{W}$ and the weighted variance of the covariates is constant for all j), the statistic $\overline{X}_j = \hat{\beta}$ is the estimate of the slope from Poisson regression in an analysis of the first N_j observations. For our purposes, we can usually assume sufficient sample sizes such that a reasonable approximate test is obtained by using as Z_j the Z

statistic for the test of the slope. The statistic P_j is the one-sided P value from such a test used to detect the alternative $H_+ : \theta > 0$. If only a two-sided P value is provided by statistical software (as is generally the case), then P_j is half the two-sided P value when $Z_j > 0$, and P_j is 1 minus half the two-sided P value when $Z_j < 0$. As discussed above, it might be more robust to use the t distribution rather than the standard normal distribution when the variance is unknown, hence it is probably easiest to use the P-scale when using this statistical model.

It should be noted that although the above derivation assumes that the mean $\overline{W}$ and weighted variance or the covariates is constant at each analysis, the statistical behavior of the group sequential test is not substantially affected by slight deviations from those values across the different interim analyses. Hence at the design stage, it is sufficient to assume constant values, and then when actually monitoring the study to use the observed values at each analysis. This is equivalent to just ignoring any variation in $\overline{W}$ and the weighted variance of the covariance across analysis times and using the value of P_j as defined above at each analysis.

Of course, major deviations in the distribution of covariates across analysis times will affect the statistical behavior of the group sequential test when boundaries are determined solely on the basis of the number of sampling units accrued to date. This problem is alleviated for the most part when stopping boundaries are determined on the basis of the proportion of the planned maximal statistical information accrued to date. This aspect is discussed further in section 12.

3.5 Censored Time to Event

In some clinical trials, the outcome measures the time to some event. A complicating factor of many such studies is that some of the observations are right censored. That is, some of the events have not yet been observed, and instead we only know that they have not occurred prior to some censoring time that is noninformative with respect to the true event time.

There are many probability models that have been extended to the case of such right censored data. One such model is the semiparametric proportional hazards model. In the setting of a two arm clinical trial, that model leads to the logrank test. In this application, we assume that sample sizes are sufficiently large to allow inference based on the normal approximation to the logrank test statistic, which is described in section 3.5.1

In censored time to event analyses, the statistical information is proportional to the number of observed (uncensored) events. Hence, after obtaining a sample size estimate for the number of events, some model must be used to devise a sampling scheme that would result in that number of events in a prescribed period of follow-up. This is discussed in section 3.5.2

3.5.1 Logrank Test Comparing Times to Event in Two Sample

Consider a clinical trial in which M subjects are randomly allocated to treatment or control in the ratio of r:1. Further suppose that the hazard function for the distribution of failure times in the control group is given by $\lambda_0(t)$ and in the treatment group is given by $\lambda_1(t) = \lambda_0(t)\theta$. We wish to test the null hypothesis that the hazard ratio comparing the treatment group to the control group is 1, $H_0 : \theta = 1$. In such a trial, we can use our fundamental model with S_j the partial likelihood based score function for $\beta = \log(\theta)$ in a proportional hazards regression model at the jth analysis, with moments $\mu = \log(\theta)r/(r+1)^2$, $\mu_0 = 0$, $\sigma_2 = r/(r+1)^2$, and N_j counts the number of failures observed by the jth analysis.

In this application, in our parameterization of $\mu = \psi g(\theta) = \log(\theta)r/(r+1)^2$, the hazard ratio θ is the natural parameter θ, $\psi = r/(r+1)^2$, and the transformation $g(\cdot)$ is the logarithmic transformation $g(\theta) = \log(\theta)$. Alternatively, a user may wish to make inference in the scale of the log hazard ratio β, in which case β would be treated as the natural parameter θ, $\psi = r/(r+1)^2$, and the transformation $g(\cdot)$ is the identity transformation $g(\theta) = \theta$.

3.5.2 Determining the Sampling Scheme to Obtain a Desired Number of Events

In clinical trial that has a primary endpoint measuring time to event, the statistical information is proportional to the number of observed (uncensored) events. The actual number of subjects to be accrued will have to be computed based on some assumption about the accrual rate, the time of accrual, the time of additional follow-up after accrual has stopped, the baseline hazard $\lambda_0(t)$, and some hypothesized value of the hazard ratio θ. One such method of determining sample size can be based on the assumption of accrual of subjects uniformly over the interval (0,a), the assumption that censoring of observations occurs only by continued survival at time of analyses, that the final analysis takes place at time $\tau \geq a$, and that the survival times in the control population follows an exponential distribution with hazard rate λ_0.

Under the above model, the probability of observing a failure by time t in the control sample is

$$\begin{cases} \frac{t}{a} - \frac{1}{\lambda_0 a} + \frac{\exp\{-\lambda_0 t\}}{\lambda_0 a} & \text{if } t < a \\ 1 - \frac{\exp\{-\lambda_0(t-a)\}}{\lambda_0 a} + \frac{\exp\{-\lambda_0 t\}}{\lambda_0 a} & \text{if } t \geq a \end{cases}$$

For the treatment sample, a similar formula holds in which λ_0 is replaced by $\lambda_1 = \lambda_0\theta$. Thus if the subjects are randomized r treatment : 1 control, then in order to expect to observe N failures by time $\tau \geq a$, we must randomize M subjects overall according to

$$\begin{aligned} M \quad = N \Bigg[& \left(\tfrac{1}{r+1}\right)\left(1 - \tfrac{\exp\{-\lambda_0(t-a)\}}{\lambda_0 a} + \tfrac{\exp\{-\lambda_0 t\}}{\lambda_0 a}\right) + \\ & \left(\tfrac{r}{r+1}\right)\left(1 - \tfrac{\exp\{-\lambda_1(t-a)\}}{\lambda_1 a} + \tfrac{\exp\{-\lambda_1 t\}}{\lambda_1 a}\right) \Bigg]^{-1} \end{aligned}$$

3.6 Statistics Based on Efficient Scores

The above settings can all be shown to be special cases of tests based on regression parameters using statistics derived from efficient likelihood theory. From asymptotic likelihood theory, we have that the efficient score function $U(\theta)$ evaluated at $\theta = \theta_0$ has asymptotic distribution

$$U(\theta_0) \dot{\sim} \mathcal{N}([\theta - \theta_0]I(\theta), I(\theta)),$$

where $I(\theta)$ is Fisher's information at the true value of θ.

Thus we can apply our fundamental model to this setting by considering the normal probability model for a one arm study. Furthermore, if we use $\mu = [\theta - \theta_0]$, $\sigma^2 = 1$ and $N = I(\theta)$, the "sample size" N is just measuring the accrual of statistical information for observations $X_i = \log(f(Y_i, \theta_0))$.

This formulation can be used to implement what is often referred to as "information based" monitoring in the group sequential literature. The test statistic Z_j would just be the score statistic calculated at the jth analysis, with boundaries chosen according to the magnitude of the Fisher's information $I_j(\theta)$ at the jth analysis relative to the planned maximal information $I_J(\theta)$.

4 Group Sequential Stopping Rules

Our goal is to decide between hypotheses $H_0 : \mu = \mu_0$, $H_+ : \mu \geq \mu_+$, and $H_- : \mu \leq \mu_-$ by repeatedly analyzing the data (up to J times) as it accrues. That is, the sample sizes $N_1, \ldots, N_J = N$ correspond to the sample sizes at which an analysis of the data is performed. The group sequential stopping rule is specified by defining the conditions under which a study is stopped and the conditions under which a study is continued to the next analysis. In this section we define first a general structure for stopping rules based on the partial sum statistic. We then illustrate the way that this general framework can be used to construct some of the more common group sequential stopping rules. Finally, we discuss the transformation of stopping rules from the partial sum scale to other scales, and vice versa.

4.1 Stopping Rules on the Partial Sum Scale

A stopping rule for the partial sum statistic shall be defined by specifying *continuation sets*, $\mathcal{C}_{Sj} \subset (-\infty, \infty)$, for $j = 1, \ldots, J$, where we use the subscript 'S' to explicitly denote a continuation set defined for the S-scale. The complements of the continuation sets will be termed the stopping sets.

We will use these continuation sets to define a stopping rule in the following manner. Starting with $j = 1$, we compute S_j and compare that value to $\mathcal{C}_{Sj}$. If $S_j \notin \mathcal{C}_{Sj}$, we stop the study (later we shall discuss the decisions that we shall make at the time of stopping the study). Otherwise, we continue the study by incrementing j and again computing the value of the statistic and comparing that value to the continuation set. We shall let M be the analysis at which the study is terminated, and we shall define S (without a subscript) as the value of the partial sum statistic when the study terminates. That is,

$$\begin{aligned} M &= \min\{j : S_j \notin \mathcal{C}_{Sj}\} \\ S &= S_M \end{aligned} \tag{4.1}$$

In order to guarantee that there are at most J analyses performed, we require that the Jth continuation set be empty. In order to guarantee a unique specification for each stopping rule based on a particular statistic, we shall also adopt the convention that all continuation sets before the Jth are neither empty nor exhaustive. Thus we have constraints

$$\begin{aligned} \mathcal{C}_{Sj} &\neq \emptyset \qquad j = 1, \ldots, J-1 \\ \mathcal{C}_{Sj} &\neq (-\infty, \infty) \qquad j = 1, \ldots, J-1 \\ \mathcal{C}_{SJ} &= \emptyset \end{aligned} \tag{4.2}$$

The basic goal of a stopping rule is to stop a study as soon as we can be sufficiently confident of the decision we would have made if we had continued the study long enough to observe the entire sample. In choosing a stopping rule that is appropriate for the three hypotheses H_+, H_0, and H_-, we need only consider the possibilities that we might want to stop when the data tend to be so high as to suggest that we would want to decide for H_+, that we might want to stop when the data tend to be so low as to suggest that we would want to decide for H_-, or that we might want to stop when the data tend to be so close to μ_0 as to suggest that we would want to decide for H_0. We note that in some situations, we may not want to decide early in favor of certain of these three hypotheses. Our purpose now is to provide enough flexibility that we can construct stopping rules that would allow such early decisions if desired.

At each of the J analyses, the partial sum statistic S_j is stochastically ordered in the sense that the statistic tends to be larger when the value of μ is larger. It is intuitively reasonable that our most general stopping rule would allow us to stop the study when the statistic is extremely

large, extremely small, or tending closely to the middle. Thus our stopping rule needs to consider continuation sets that allow us to continue the study when we can not yet distinguish between H_+ and H_0 or when we can not yet distinguish between H_0 and H_-. Thus, we find it adequate to consider continuation sets that can be specified as the union of two disjoint intervals. Thus, the continuation sets for the partial sum statistic can be defined for $j = 1, \ldots, J$ as

$$\mathcal{C}_{Sj} = (a_{Sj}, b_{Sj}] \cup [c_{Sj}, d_{Sj}) \tag{4.3}$$

We shall adopt the convention that the boundaries of the continuation sets satisfy

$$a_{sj} \leq b_{sj} \leq c_{sj} \leq d_{sj}. \tag{4.4}$$

Due to the constraints imposed by eqn (4.2), we must have that

$$\begin{aligned} a_{Sj} \neq b_{Sj} \qquad \text{OR} \qquad c_{Sj} \neq d_{Sj} \qquad j = 1, \ldots, J-1 \\ a_{SJ} &= b_{SJ} \\ c_{SJ} &= d_{Sj} \end{aligned} \tag{4.5}$$

The continuation sets specified in eqn (4.3) were specified as unions of half open intervals. These continuation sets were motivated by the desire to continue whenever we had not yet distinguished between two adjacent hypotheses (i.e, either H_+ and H_0 or H_0 and H_-). This plan allows the greatest flexibility when considering decisions involving three hypotheses (e.g., a two-sided hypothesis test). However, when testing a two-sided hypothesis in which we do not desire to stop early if the data are consistent with H_0, we may want a continuation set that can be represented as a single interval. In such a case, however, we need only choose the continuation intervals to be contiguous, i.e., choose $b_j = c_j$. For notational convenience, we adopted the convention that

$$a_{Sj} < b_{Sj} = c_{Sj} < d_{Sj} \quad \Rightarrow \quad b_{Sj} \in \mathcal{C}_{Sj}. \tag{4.6}$$

That is, in this case we shall assume that the two intervals comprising the continuation set are half open intervals, rather than open intervals as denoted in eqn (4.3).

4.2 Classes of Commonly Used Group Sequential Stopping Rules

The usual way in which stopping rules are used to implement a group sequential stopping test is to divide the stopping sets (the complements of the continuation sets) into subsets corresponding to the decisions to be made regarding the null and alternative hypotheses. This concept was used as the motivation for our choice of the general structure of a continuation set as the union of two disjoint intervals as described in eqn (4.3). Thus, while the distribution of our statistics is determined solely by the continuation sets at each analysis (see section 5), it seems intuitively clear that greater statistical efficiency will be obtained if the boundaries of the continuation sets also demarcate the boundaries between decisions for H_+, H_0, and H_-, as appropriate for the application.

The exact correspondence between our stopping boundaries and the decision we make regarding the hypotheses will depend somewhat upon the goals of the clinical trial. That is, the decision that will correspond to early stopping will depend upon whether we are trying to distinguish between all three hypotheses H_+, H_0, and H_- (e.g., a two sided test), or whether we want to combine two of the hypotheses (e.g., a one sided test). For instance, there have been several basic structures proposed in the statistical literature for group sequential hypothesis tests. The following describe such stopping rules in our notation. In this description, all boundaries left unspecified can be chosen arbitrarily (subject to the constraints imposed by eqns (4.4) and (4.5)) in order to meet the desired operating characteristics for the test.

A. A single upper early stopping boundary.

$$\begin{aligned} d_{Sj} &= \text{ arbitrary for } j = 1, \ldots, J \\ c_{Sj} &= b_{Sj} j = 1, \ldots, J \\ a_{Sj} &= -\infty, j = 1, \ldots, J-1 \\ a_{SJ} &= d_{SJ} \end{aligned}$$

Situations for which such a group sequential design might be appropriate include:

1. Testing H_0 against H_+ in a situation where the study should be stopped early only in the case of evidence against the null hypothesis. That is, we do not desire to distinguish between H_- and H_0, and it is only deemed important to terminate the study early in situations where the data look to be inconsistent with H_- and H_0. Our decision rule might be to decide H_+ if $S \geq d_{SM}$, and to decide H_0 if $S < a_{SM}$. We note that the latter situation, which could also be written $S < d_{SM}$, can only occur if $M = J$.
2. Testing the one-sided hypotheses H_0 against H_- when early stopping is only desired in the case of data which is so consistent with H_0 (or H_+) as to preclude our further consideration of H_-. In such a setting, our decision rule might be to decide H_0 if $S \geq d_{SM}$, and to decide H_- if $S < a_{SM}$. We note that the latter situation can only occur if $M = J$.

B. A single lower early stopping boundary.

$$\begin{aligned} a_{Sj} &= \text{ arbitrary for } j = 1, \ldots, J \\ b_{Sj} &= c_{Sj}, j = 1, \ldots, J \\ d_{Sj} &= \infty, j = 1, \ldots, J-1 \\ d_{SJ} &= a_{SJ} \end{aligned}$$

Situations for which such a group sequential design might be appropriate include:

1. Testing H_0 against H_- in a situation where the study should be stopped early only in the case of evidence against the null hypothesis. That is, we do not desire to distinguish between H_+ and H_0, and it is only deemed important to terminate the study early in situations where the data look to be inconsistent with H_+ and H_0. Our decision rule might be to decide H_- if $S \leq a_{SM}$, and to decide H_0 if $S > d_{SM}$. We note that the latter situation can only occur if $M = J$.
2. Testing the one-sided hypotheses H_0 against H_+ when early stopping is only desired in the case of data which is so consistent with H_0 (or H_-) as to preclude our further consideration of H_+. In such a setting, our decision rule might be to decide H_0 if $S \leq a_{SM}$, and to decide H_+ if $S > d_{SM}$. We note that the latter situation can only occur if $M = J$.

C. Lower and upper early stopping boundaries which meet at the final analysis:

$$\begin{aligned} d_{Sj} &= \text{ arbitrary for } j = 1, \ldots, J \\ c_{Sj} &= b_{Sj}, j = 1, \ldots, J \\ a_{Sj} &= \text{ arbitrary for } j = 1, \ldots, J-1 \\ a_{SJ} &= d_{SJ} \end{aligned}$$

Situations for which such a group sequential design might be appropriate include:

1. Testing the one-sided hypotheses H_0 against H_+ when early stopping might be desired in the case of data which is so consistent with H_0 (or H_-) as to preclude our further consideration of H_+, or when the data suggests that H_0 (and H_-) should be rejected. In such a setting, our decision rule might be to decide H_0 if $S \leq a_{SM}$, and to decide H_+ if $S \geq d_{SM}$.
2. Testing the one-sided hypotheses H_0 against H_- when early stopping might be desired in the case of data which is so consistent with H_0 (or H_+) as to preclude our further consideration of H_-, or when the data suggests that H_0 (and H_+) should be rejected. In such a setting, our decision rule might be to decide H_0 if $S \geq d_{SM}$, and to decide H_- if $S \leq a_{SM}$.

D. Lower and upper early stopping boundaries which do not meet at the final analysis:

$$\begin{aligned} d_{Sj} &= \text{ arbitrary for } j = 1, \ldots, J \\ c_{SJ} &= d_{SJ} \\ c_{Sj} &= b_{Sj}, j = 1, \ldots, J-1 \\ b_{SJ} &= a_{SJ} \\ a_{Sj} &= \text{ arbitrary for } j = 1, \ldots, J \end{aligned}$$

Situations for which such a group sequential design might be appropriate include:

1. Testing two-sided hypotheses H_- against H_0 against H_+ when early stopping might be desired only in the case of evidence against the null hypothesis. In such a setting, our decision rule might be to decide H_+ if $S \geq d_{SM}$, to decide H_- if $S \leq a_{SM}$, and to decide H_0 if $b_{SM} < S < aSM$. We note that the latter decision can be made only if $M = J$.

E. Four early stopping boundaries:

$$\begin{aligned} d_{Sj} &= \text{ arbitrary for } j = 1, \ldots, J \\ c_{Sj} &= \text{ arbitrary for } j = 1, \ldots, J-1 \\ c_{SJ} &= d_{SJ} \\ b_{Sj} &= \text{ arbitrary for } j = 1, \ldots, J-1 \\ b_{SJ} &= a_{SJ} \\ a_{Sj} &= \text{ arbitrary for } j = 1, \ldots, J \end{aligned}$$

Situations for which such a group sequential design might be appropriate include:

1. Testing two-sided hypotheses H_- against H_0 against H_+ when early stopping might be desired in the case of evidence against the null hypothesis or when the data are so consistent with the null hypothesis as to preclude further consideration of H_+ or H_-. In such a setting, our decision rule might be to decide H_+ if $S \geq d_{SM}$, to decide H_- if $S \leq a_{SM}$, and to decide H_0 if $b_{SM} < S < c_{SM}$. We note that it frequently happens that $b_{Sj} = c_{Sj}$ at some of the earliest analyses, in which case stopping with a decision for the null hypothesis is impossible at those analysis times.

It should be clear that the above list is not exhaustive: There are many other patterns of group sequential tests that are possible within this framework. It is rare, however, that any other patterns of designs will be used in practice.

It should also be clear that the applications described for each of the types of designs is not exhaustive. In particular, we shall discuss the application of these designs to equivalence testing later in this document.

4.3 Transformations of Stopping Rules to Other Scales

In section **??**, we described stopping rules in terms of the partial sum scale. However, because of the one to one relationship between the statistics defined in eqn (1.12), we can also specify a particular stopping rule based on any of the statistics. This is because the specification of the continuation sets for any one of the statistics given in eqn (1.12) automatically induces a corresponding continuation set for the others. That is, given a stopping rule specified by particular choices of a_{Sj}, b_{Sj}, c_{Sj}, and d_{Sj} for $j = 1, \ldots, J$, the stopping rules for other choices of test statistic are easily found by applying the transformations in eqn (1.14) to each of the boundaries. For example, the stopping rule for the sample mean statistic can be found as

$$a_{\overline{X}j} = a_{Sj}/N_j \quad b_{\overline{X}j} = b_{Sj}/N_j \quad c_{\overline{X}j} = c_{Sj}/N_j \quad d_{\overline{X}j} = d_{Sj}/N_j$$

Note that if the boundaries on the partial sum scale satisfy the constraints given by eqns (4.4) and (4.5), then the boundaries on the sample mean scale satisfy similar constraints.

In order to explicitly denote the stopping rule for a specific test statistic, we shall subscript the boundary with the letter denoting the scale. Hence, for instance, a_{Sj}, $a_{\overline{X}j}$, a_{Zj}, a_{Pj}, a_{Bj}, a_{Cj}, a_{Hj}, a_{E_aj}, a_{E_bj}, a_{E_cj}, and a_{E_dj} shall denote the lower boundary for the partial sum statistic, the sample mean statistic, the normalized Z statistic, the fixed sample P value, the Bayesian posterior probability, the conditional futility statistic, the predictive futility statistic, the lower type I error spending statistic, the lower type II error spending statistic, the upper type II error spending statistic, and the upper type I error spending statistic, respectively. It should be noted that in the case of the stopping rules on the B-, C-, H-, and E-scales, the transformations depend upon some particular choice of hypothesized mean, testing threshold, or both. We shall thus have to make clear the choices of those parameters when using stopping boundaries on those scales.

In later sections, we shall define families of group sequential stopping rules based on the various scales. In fact, in some families, different transformations will be used for the a, b, c, and d boundaries in that different choices of hypothesized means and/or testing thresholds will be used. In such cases, it is not always immediately clear by inspection that the constraints in eqn (4.2) are satisfied. Nonetheless, we shall require that a group sequential stopping rule defined on other scales satisfy eqns (4.4) and (4.5) when the stopping rule is transformed to the partial sum scale.

5 Sampling Density

When choosing a group sequential stopping rule for use in frequentist hypothesis testing, we generally desire to find stopping boundaries to guarantee a level α test for some specified value of α. As discussed in the next section, there are a number of other operating characteristics that one might typically examine in the process of selecting an appropriate stopping rule for a clinical trial. In order to compute many of these operating characteristics, we need to know the sampling density for the test statistic.

5.1 Sampling Density for Partial Sum Statistic

In the previous section, we defined stopping rules for the partial sum statistic in detail, and then we discussed the ways in which stopping rules could be derived for other test statistics by using the transformations given in section 1.5. When deriving the sampling density for the test statistics, it is also easiest to derive the density for the partial sum statistic or the sample mean statistic, and to use one of those forms when making probability statements about other test statistics.

Hence, we consider a group sequential stopping rule having continuation sets for the partial sum statistic given by

$$\mathcal{C}Sj = (a_{Sj}, bSj] \cup [c_{Sj}, dSj). \tag{5.1}$$

For a particular value of μ, we desire to find the sampling density $p(m, s; \mu)$ for the test statistic $(M = m, S = s)$, $m = 1, \ldots, J$, $s \in (-\infty, \infty)$, as defined by eqn (4.1). This can be shown to be (Armitage, McPherson, and Rowe, 1969)

$$p(m, s; \mu) = \begin{cases} f(m, s; \mu) & s \notin \mathcal{C}_{Sm}, \text{and} \\ 0 & \text{else} \end{cases} \tag{5.2}$$

where the function $f(j, s; \mu)$ is recursively defined as

$$\begin{aligned} f(1, s; \mu) &= \frac{1}{\sqrt{n_1}\sigma}\phi\left(\frac{s - n_1\mu}{\sqrt{n_1}\sigma}\right) \\ f(j, s; \mu) &= \int_{\mathcal{C}_{S(j-1)}} \frac{1}{\sqrt{n_j}\sigma}\phi\left(\frac{s - u - n_j\mu}{\sqrt{n_j}\sigma}\right) f(j-1, u; \mu)\, du, \quad j = 2, \ldots, m \end{aligned} \tag{5.3}$$

where $\phi(x) = e^{-x^2/2}/\sqrt{2\pi}$ is the density for the standard normal distribution and $n_1 = N_1$ and $n_j = N_j - N_{j-1}$ for $j = 2, \ldots, J$ denote the size of the groups accrued between successive analyses.

The function $f(j, s; \mu)$ is the subdensity for S_j. For notational convenience, it is useful to define the cumulative function

$$F(j, s; \mu) = \int_{-\infty}^{s} f(j, u; \mu)\, du. \tag{5.4}$$

It should be noted that

$$F(j, \infty; \mu) = 1 - \sum_{k=1}^{j-1} Pr(M = k; \mu), \tag{5.5}$$

which is strictly less than 1 for $j > 1$. We define the inverse function $F^{-1}(j, y; \mu)$ by

$$F^{-1}(j, y; \mu) = s_y \Leftrightarrow F(j, s_y; \mu) = y. \tag{5.6}$$

The function $f(j, s; \mu)$ can not be integrated in closed form, thus numerical integration routines are necessary. The use of such routines for testing/estimating an unknown mean μ is made easier by the relation

$$f(j, s; \mu) = f(j, s; 0) \exp\left(\frac{s\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}N_j\right). \tag{5.7}$$

The most common uses of the sequential density $p(m,s;\mu)$ involve integrations of the form $\int_x^y g(m,s)p(m,s;\mu)\,ds$, where $g(m,s)$ is some known function such as $g(m,s)=1$ (for the computation of probabilities), $g(m,s)=s$ (for the computation of expectations), or $g(m,s)=s^2$ (for the computation of variances). Some computational efficiency is obtained by the following derivation. Suppose that $p(m,s;\mu)>0$ for all $s\in(x,y)$ (that is, interval (x,y) does not overlap with the continuation set at the mth analysis) and that $g(m,s)=s^t$ for $t=0,1,2$. Then by interchanging the order of integration, we can write

$$\begin{aligned}\int_x^y g(m,s)p(m,s;\mu)\,ds &= \int_x^y \int_{\mathcal{C}_{S(m-1)}} s^t \frac{1}{\sqrt{n_m}\sigma}\phi\left(\frac{s-u-n_m\mu}{\sqrt{n_m}\sigma}\right) f(m-1,u;\mu)\,du\,ds \\ &= \int_{\mathcal{C}_{S(m-1)}} f(m-1,u;\mu)\left[\int_x^y s^t \frac{1}{\sqrt{n_m}\sigma}\phi\left(\frac{s-u-n_m\mu}{\sqrt{n_m}\sigma}\right) ds\right] du \\ &= \int_{\mathcal{C}_{S(m-1)}} f(m-1,u;\mu)h_t(u,x,y,m;\mu)\,du \qquad (5.8)\end{aligned}$$

where the functions $h_0()$, $h_1()$, and $h_2()$ are given by

$$\begin{aligned}h_0(u,x,y,m;\mu) &= \Phi\left(\frac{y-u-n_m\mu}{\sqrt{n_m}\sigma}\right)-\Phi\left(\frac{x-u-n_m\mu}{\sqrt{n_m}\sigma}\right) \\ h_1(u,x,y,m;\mu) &= (u+n_m\mu)h_0(u,x,y,m;\mu)- \\ &\quad \sqrt{n_m}\sigma\left[\phi\left(\frac{y-u-n_m\mu}{\sqrt{n_m}\sigma}\right)-\phi\left(\frac{x-u-n_m\mu}{\sqrt{n_m}\sigma}\right)\right] \qquad (5.9) \\ h_2(u,x,y,m;\mu) &= ((u+n_m\mu)^2+n_m\sigma^2)h_0(u,x,y,m;\mu)- \\ &\quad \sqrt{n_m}\sigma\left[(y+u+n_m\mu)\phi\left(\frac{y-u-n_m\mu}{\sqrt{n_m}\sigma}\right)-(x+u+n_m)\phi\left(\frac{x-u-n_m\mu}{\sqrt{n_m}\sigma}\right)\right]\end{aligned}$$

The advantage afforded by the above formulas is that good approximations to the standard normal cumulative distribution function $\Phi(x)$ exist in closed form. Thus, the integral $\int g(m,s)p(m,s;\mu)ds$ is computed for approximately the same cost as computing $p(m,s;\mu)$.

5.2 Sampling Density for Sample Mean Statistic

As noted above, computation of probabilities is most easily performed on the partial sum scale. We include the sampling density for the sample mean scale here for informational purposes only.

We define the density for a group sequential stopping rule having continuation sets for the sample mean statistic given by $\mathcal{C}_{\overline{X}j}=(a_{\overline{X}j},b_{\overline{X}j})\cup(c_{\overline{X}j},d_{\overline{X}j})$, which can be derived from the stopping rule for the partial sum statistic given in eqn (5.1) by applying the transformation eqn (1.14) to each of the continuation set boundaries. The sampling distribution for $(M,\overline{X})$, denoted by $p_{\overline{X}}(m,x;\mu)$, can be written in the recursive form of Armitage, McPherson, and Rowe (1969):

$$p_{\overline{X}}(m,x;\mu)=\begin{cases} f_{\overline{X}}(m,x;\mu) & x\notin\mathcal{C}_{\overline{X}m}, \text{and} \\ 0 & \text{else}\end{cases} \qquad (5.10)$$

where the function $f_{\overline{X}}(j,s;\mu)$ is recursively defined as

$$\begin{aligned}f_{\overline{X}}(1,x;\mu) &= \frac{n_1}{\sqrt{n_1}\sigma}\phi\left(\frac{xn_1-n_1\delta}{\sqrt{n_1}\sigma}\right) \\ f_{\overline{X}}(j,x;\delta) &= \int_{\mathcal{C}_{x(j-1)}} \frac{N_j}{\sqrt{n_j}\sigma}\phi\left(\frac{xN_j-uN_{j-1}-n_j\mu}{\sqrt{n_j}\sigma}\right) f_{\overline{X}}(j-1,u;\mu)\,du, \qquad (5.11)\end{aligned}$$

for $j=2,\ldots,J$.

5.3 Sampling Density Under the Standardizing Transformation

As noted in section 1.6, in most study design situations, we are interested in determining the sample size which would provide adequate power to detect an alternative hypothesis of interest. We thus need to be able to compute the operating characteristics of a group sequential test in some standardized form, and then solve for the sample size that would provide those operating characteristics for a specific alternative. In this section, we present the sampling density for the partial sum statistic under the standardizing transformation given in eqn (1.15).

Under the standardizing transformation, the group sequential stopping rule for the standardized partial sum statistic S_j^* has continuation sets

$$\mathcal{C}_{Sj}^* = (a_{Sj}^*, b^*Sj] \cup [c_{Sj}^*, d^*Sj),$$

where the continuation set boundaries are found by applying the transformation of eqn (1.31) to the stopping boundaries $\mathcal{C}_{Sj} = (a_{Sj}, bSj) \cup (c_{Sj}, dSj)$ to obtain

$$\begin{aligned} a^*Sj &= [a_{Sj} - N_j\mu_0]/[\sigma\sqrt{N_J}] \\ b^*Sj &= [b_{Sj} - N_j\mu_0]/[\sigma\sqrt{N_J}] \\ c^*Sj &= [c_{Sj} - N_j\mu_0]/[\sigma\sqrt{N_J}] \\ d^*Sj &= [d_{Sj} - N_j\mu_0]/[\sigma\sqrt{N_J}] \end{aligned} \tag{5.12}$$

The partial sum statistic in the untransformed problem is (M^*, S^*), where $M^* = M$ and $S^* = [S - N_j\mu_0]/[\sigma\sqrt{N_J}]$ as specified in eqn (1.31). For a particular value of $\delta = \sqrt{N_J}[\mu = \mu_0]/\sigma$, we desire to find the sampling density $p^*(m^*, s^*; \delta)$ for the test statistic $(M^* = m^*, S^* = s^*)$, $m^* = 1, \ldots, J$, $s^* \in (-\infty, \infty)$. This can be shown to be (Armitage, McPherson, and Rowe, 1969)

$$p^*(m^*, s^*; \delta) = \begin{cases} f^*(m^*, s^*; \delta) & s^* \notin \mathcal{C}_{\S^* m^*}^*, \text{and} \\ 0 & \text{else} \end{cases} \tag{5.13}$$

where the function $f^*(j, s^*; \delta)$ is recursively defined as

$$\begin{aligned} f(1, s^*; \delta) &= \frac{1}{\sqrt{\pi_1}}\phi\left(\frac{s^* - \pi_1\delta}{\sqrt{\pi_1}}\right) \\ f(j, s^*; \delta) &= \int_{\mathcal{C}_{S(j-1)}^*} \frac{1}{\sqrt{\pi_j}\sigma}\phi\left(\frac{s^* - u - \pi_j\delta}{\sqrt{\pi_j}}\right) f^*(j-1, u; \mu)\, du, \quad j = 2, \ldots, m \end{aligned} \tag{5.14}$$

where $\phi(x) = e^{-x^2/2}/\sqrt{2\pi}$ is the density for the standard normal distribution and $\pi_1 = N_1/N_J$ and $\pi_j = \Pi_j - \Pi_{j-1} = [N_j - N_{j-1}]/N_J$ for $j = 2, \ldots, J$ denote the proportion of the maximal sample size which is accrued between successive analyses. We again have the computationally useful form

$$f^*(j, s^*; \delta) = f(j, s^*; 0) \exp\left(s^*\delta - \frac{\delta^2\Pi_k}{2}\right). \tag{5.15}$$

Given the 1:1 transformation of the stopping boundaries for the original data and the stopping boundaries for the standardized problem, it should be clear that testing or estimating δ in the standardized setting is equivalent to testing or estimating μ with the original data. That is, a hypothesis test of $H_0 : \delta = 0$ is equivalent to testing $H_0 : \mu = \mu_0$. Furthermore, for specified sample sizes $N_1, \ldots, N_J = N$, the operating characteristics of the group sequential test for the standardized problem for a given value of $\delta = \delta_*$ will be equivalent to the suitably transformed operating characteristics for the corresponding group sequential test defined for the original data (using the relationships given in eqns (1.32)) when

$$\mu_* = \mu_0 + \frac{\sigma}{\sqrt{N_J}}\delta_*. \tag{5.16}$$

Similarly, when we obtain an estimate $\hat{\delta}$ of δ from the standardized problem, using (1.33) we can easily obtain an estimate $\hat{\mu}$ for μ in the fundamental model and $\hat{\theta}$ for the natural parameter in an application of the fundamental model (see section 3) according to

$$\begin{aligned} \hat{\mu} &= \mu_0 + \frac{\sigma}{\sqrt{N_K}}\hat{\delta} \\ \hat{\theta} &= g^{-1}\left(\frac{\hat{\mu}}{\psi}\right). \end{aligned} \tag{5.17}$$

6 Operating Characteristics

In a fixed sample study in which all data are accrued prior to any analysis, reference to the operating characteristics of the test is usually taken to mean the size (type I error) and power curve (one minus the type II error). In the presence of a stopping rule, however, there are more features of the study design that might need to be examined. For instance, the sample size accrued during the study is now a random variable, and hence summary statistics for that distribution might be of interest. In this section we describe some of the measures that might be used to evaluate whether a particular stopping rule is appropriate in a given clinical trial situation.

6.1 Power Functions

In Neyman-Pearson hypothesis testing, we generally choose critical values for rejection of the null hypothesis such that the probability of falsely rejecting the null (referred to as the type I statistical error) is acceptably low. This agreed upon value for the type I error is called the level of significance, or just the level or size of the test.

It is often more convenient, however, to consider the probability of rejecting the null hypothesis under various hypothesized treatment effects. Thus we consider the power function of the test $\beta(\mu)$, the probability of rejecting the null hypothesis as a function of the true value of the unknown mean. The type I error is then the value of the power function when the null hypothesis is true (e.g., $\alpha = \beta(\mu_0)$, and the type II error (the probability of falsely failing to reject the null) for some given value of the unknown mean μ is $1 - \beta(\mu)$.

In the group sequential tests described in the previous section (as well as in the usual fixed sample hypothesis tests), the stopping sets consistent with rejection of the null hypothesis vary in structure according to whether we are testing one-sided or two-sided hypotheses. In defining the operating characteristics of a group sequential test, we shall therefore find it more useful to define three functions

$$\begin{aligned} \beta_+(\mu) &= Pr(S \geq d_{SM};\ \mu) \\ \beta_0(\mu) &= Pr(b_{SM} < S < c_{SM};\ \mu) \\ \beta_-(\mu) &= Pr(S \leq a_{SM};\ \mu) \end{aligned} \tag{6.1}$$

We note that these functions must satisfy $\beta_+(\mu)+\beta_0(\mu)+\beta_-(\mu) = 1$ for all μ, thus any two of these functions is actually sufficient to specify the operating characteristics of the test. We shall typically restrict attention to the 'upper power function' $\beta_+(\mu)$ and the 'lower power function' $\beta_-(\mu)$.

The functions specified in eqn (6.1) are easily related to the classical way of characterizing a hypothesis test. For instance, for the group sequential tests defined in the previous section, the classic power function would be

$$\beta(\mu) = \begin{cases} \beta_+(\mu) + \beta_-(\mu) & \text{for two-sided tests of type D or E,} \\ \beta_+(\mu) & \text{for one-sided tests of type A1, B2, or C1, and} \\ \beta_-(\mu) & \text{for one-sided tests of type A2, B1, or C2.} \end{cases} \tag{6.2}$$

A level α test of the null hypothesis would have $\beta(\mu_0) = \alpha$.

We can also define the operating characteristics of a group sequential test in the standardized setting for appropriate transformations of the hypotheses (eqn (1.33)), boundaries (eqn (1.31)), and statistics (eqn (1.31)) by

$$\begin{aligned} \beta^*_+(\delta) &= Pr(S^* \geq d^*_{SM};\ \delta) \\ \beta^*_0(\delta) &= Pr(b^*_{SM} < S < c^*_{SM};\ \delta) \\ \beta^*_-(\delta) &= Pr(S^* \leq a^*_{SM};\ \delta) \end{aligned} \tag{6.3}$$

The classic power function would be

$$\beta^*(\delta) = \begin{cases} \beta^*_+(\delta) + \beta^*_-(\delta) & \text{for two-sided tests of type D or E,} \\ \beta^*_+(\delta) & \text{for one-sided tests of type A1, B2, or C1, and} \\ \beta^*_-(\delta) & \text{for one-sided tests of type A2, B1, or C2.} \end{cases} \tag{6.4}$$

A level α test of the null hypothesis would have $\beta^*(0) = \alpha$.

6.2 Stopping Probabilities

The power function described in the previous subsection applies equally well to both the fixed sample ($J = 1$) and group sequential ($J > 1$) settings. In the group sequential setting, however, it is often of interest to consider the probability of making a given decision at each of the analysis times. Hence we define the stopping probabilities at the jth analysis time as

$$\begin{aligned} \beta_{+j}(\mu) &= Pr(S \geq d_{SM} \& M = j;\ \mu) = F(j, \infty; \mu) - F(j, d_{Sj}; \mu) \\ \beta_{0j}(\mu) &= Pr(b_{SM} < S < c_{SM} \& M = j;\ \mu) = F(j, c_{Sj}; \mu) - F(j, b_{Sj}; \mu) \\ \beta_{-j}(\mu) &= Pr(S \leq a_{SM} \& M = j;\ \mu) = F(j, a_{Sj}; \mu) \end{aligned} \tag{6.5}$$

where $F(j, s; \mu)$ is defined by eqn (5.4). We note that these stopping probabilities satisfy

$$\begin{aligned} \beta_+(\mu) &= \sum_{j=1}^{J} \beta_{+j}(\mu) \\ \beta_0(\mu) &= \sum_{j=1}^{J} \beta_{0j}(\mu) \\ \beta_-(\mu) &= \sum_{j=1}^{J} \beta_{-j}(\mu) \end{aligned}$$

It is also at times convenient to consider for $1 \leq k < j \leq J$ the probability of stopping at the jth analysis conditional upon not having stopped at or prior to the kth analysis. We thus define conditional stopping probabilities

$$\begin{aligned} \beta_{+j|k}(\mu) &= \beta_{+j}(\mu)/Pr(M > k) \\ \beta_{0j|k}(\mu) &= \beta_{0j}(\mu)/Pr(M > k) \\ \beta_{-j|k}(\mu) &= \beta_{-j}(\mu)/Pr(M > k) \end{aligned} \tag{6.6}$$

6.3 Error Spending Functions

In some group sequential design families or implementations of monitoring strategies, it is of interest to consider the rate at which a type I or type II error is allocated across analysis times. The statistical errors associated with a particular set of hypotheses and stopping rule are type I errors α_ℓ and α_u and type II errors $1-\beta_\ell$ and $1-\beta_u$. We thus can define the following four error spending

functions for an analysis when the proportion of the maximum sample size accrued is Π as

$$\begin{aligned} E_a(\Pi) &= \frac{1}{\alpha_\ell} \sum_{j:\Pi_j \le \Pi} \beta_{-j}(\mu_{0-}) \\ E_b(\Pi) &= \frac{1}{1-\beta_\ell} \sum_{j:\Pi_j \le \Pi} [\beta_{0j}(\mu_-) + \beta_{+j}(\mu_-)] \\ E_c(\Pi) &= \frac{1}{1-\beta_u} \sum_{j:\Pi_j \le \Pi} [\beta_{0j}(\mu_+) + \beta_{-j}(\mu_+)] \\ E_d(\Pi) &= \frac{1}{\alpha_u} \sum_{j:\Pi_j \le \Pi} \beta_{+j}(\mu_{0+}) \end{aligned} \tag{6.7}$$

It should be noted that the type I and II errors can be defined by

$$\begin{aligned} \alpha_\ell &= \beta_-(\mu_{0-}) = \sum_{j=1}^{J} \beta_{-j}(\mu_{0-}) \\ 1-\beta_\ell &= \beta_0(\mu_-) + \beta_+(\mu_-) = \sum_{j=1}^{J} [\beta_{0j}(\mu_-) + \beta_{+j}(\mu_-)] \\ &= 1 - \beta_-(\mu_-) = 1 - \sum_{j=1}^{J} \beta_{-j}(\mu_-) \\ 1-\beta_u &= \beta_0(\mu_+) + \beta_-(\mu_+) = \sum_{j=1}^{J} [\beta_{0j}(\mu_+) + \beta_{-j}(\mu_+)] \\ &= 1 - \beta_+(\mu_+) = 1 - \sum_{j=1}^{J} \beta_{+j}(\mu_+) \\ \alpha_u &= \beta_+(\mu_{0+}) = \sum_{j=1}^{J} \beta_{+j}(\mu_{0+}) \end{aligned} \tag{6.8}$$

Note also that the error spending functions defined above are related to, but not in all cases exactly equivalent to, the error spending scales for the group sequential test statistics defined in eqns (1.10) and (1.11). For $\mu_a = \mu_{0-}$ and $S_j = a_{Sj}$, $E_{aj} = E_a(\Pi_j)$. For $\mu_d = \mu_{0+}$ and $S_j = d_{Sj}$, $E_{dj} = E_d(\Pi_j)$. The error spending scale at the 'b' and 'c' boundaries, however, differ slightly from the type II error spending functions. The difference arises because the error spending scale was defined for every possible value of the S_j's, and thus considered the probability mass within the continuation regions at the jth analysis. The error spending function only considers values at the boundaries and does not include the probability mass at each analysis that occurs within continuation sets. It should be noted that the S+SeqTrial functions `seqDesign()` and `seqBoundary()` return the error spending functions when `display.scale="E"`, while the S+SeqTrial function `changeSeqScale()` returns the statistics on the error spending scale.

6.4 Sample Size Distribution

In group sequential testing, we are also often interested in characterizing the operating characteristics of a test with respect to the distribution of sample sizes at the time of study termination. Often this distribution is characterized by the expected number of subjects accrued prior to study termination, the average sample number (ASN), although other summary measures of the sample size distribution (median, 75th percentile, 90th percentile) might be more appropriate in specific

situations. Two tests with the same level of significance and the same statistical power to detect a particular alternative may have very different probability distributions for the sample size at the time the study is terminated. In general, the sample size distribution is a function of the stopping boundaries and the value of the true mean μ. The distribution function $F_N(n;\mu)$, the average sample size function $ASN(\mu)$, and the sample size quantile function $Q_N(p;\mu)$ are defined by

$$\begin{aligned} F_N(n;\mu) &= \sum_{j:N_j \le n} Pr(M=j;\mu) \\ ASN(\mu) &= \sum_{j=1}^{J} N_j Pr(M=j;\mu) \\ Q_N(p;\mu) &= N_j \quad \text{such that } Pr(M \le j;\mu) \ge p \text{ and } Pr(M \ge j;\mu \ge 1-p \end{aligned} \tag{6.9}$$

Computation of the probability functions is possible through the sampling density defined in section 5.

6.5 Measures of Futility

When evaluating a group sequential stopping rule, it is often of interest to evaluate a stopping rule with respect to the probability that the decision made when stopping at some interim analysis might be different than the decision which might have been reached at the final analysis had the study not been terminated prematurely. Evaluations of the stopping rule with respect to these criteria are based on the distribution of some test statistic at the final analysis conditional upon the test statistic being equal to the stopping boundary at an interim analysis. Because each stopping boundary is associated with rejection of a particular hypothesis, it may be of interest to consider the conditional probabilities under the corresponding hypotheses as determined by the group sequential design. This then leads to the following definitions for

$$\begin{aligned} C_{Daj} &= Pr(\overline{X}_J > a_{\overline{X}J} \,|\, \overline{X}_j = a_{\overline{X}j}\,;\, \mu = \mu_{0-}) \\ &= 1 - \Phi\left(\frac{N_J[a_{\overline{X}J} - \mu_{0-}] - N_j[a_{\overline{X}j} - \mu_{0-}]}{\sigma\sqrt{N_J - N_j}}\right) \\ C_{Dbj} &= Pr(\overline{X}_J < b_{\overline{X}J} \,|\, \overline{X}_j = b_{\overline{X}j}\,;\, \mu = \mu_{-}) \\ &= \Phi\left(\frac{N_J[b_{\overline{X}J} - \mu_{-}] - N_j[b_{\overline{X}j} - \mu_{-}]}{\sigma\sqrt{N_J - N_j}}\right) \\ C_{Dcj} &= Pr(\overline{X}_J > c_{\overline{X}J} \,|\, \overline{X}_j = c_{\overline{X}j}\,;\, \mu = \mu_{+}) \\ &= 1 - \Phi\left(\frac{N_J[c_{\overline{X}J} - \mu_{+}] - N_j[c_{\overline{X}j} - \mu_{+}]}{\sigma\sqrt{N_J - N_j}}\right) \\ C_{Ddj} &= Pr(\overline{X}_J < d_{\overline{X}J} \,|\, \overline{X}_j = d_{\overline{X}j}\,;\, \mu = \mu_{0+}) \\ &= \Phi\left(\frac{N_J[d_{\overline{X}J} - \mu_{0+}] - N_j[d_{\overline{X}j} - \mu_{0+}]}{\sigma\sqrt{N_J - N_j}}\right) \end{aligned} \tag{6.10}$$

It can be seen that these functions are closely related to the test statistic on the conditional probability scale defined by eqn (1.7). That is, $C_{Daj} = C_j(a_{\overline{X}J}, \mu_{0-})$ when $\overline{X}_j = a_{\overline{X}j}$, $C_{Dbj} = 1 - C_j(b_{\overline{X}J}, \mu_{-})$ when $\overline{X}_j = b_{\overline{X}j}$, $C_{Dcj} = C_j(c_{\overline{X}J}, \mu_{+})$ when $\overline{X}_j = c_{\overline{X}j}$, and $C_{Ddj} = 1 - C_j(d_{\overline{X}J}, \mu_{0+})$ when $\overline{X}_j = d_{\overline{X}j}$.

An alternative evaluation can be based on the conditional probabilities under the current best

estimate of μ. This then leads to the following definitions

$$\begin{aligned}
C_{Eaj} &= Pr(\overline{X}_J > a_{\overline{X}J} \,|\, \overline{X}_j = a_{\overline{X}j}\,;\, \mu = \overline{X}_j) \\
&= 1 - \Phi\left(\frac{N_J[a_{\overline{X}J} - a_{\overline{X}j}]}{\sigma\sqrt{N_J - N_j}}\right) \\
C_{Ebj} &= Pr(\overline{X}_J < b_{\overline{X}J} \,|\, \overline{X}_j = b_{\overline{X}j}\,;\, \mu = \overline{X}_j) \\
&= \Phi\left(\frac{N_J[b_{\overline{X}J} - b_{\overline{X}j}]}{\sigma\sqrt{N_J - N_j}}\right) \\
C_{Ecj} &= Pr(\overline{X}_J > c_{\overline{X}J} \,|\, \overline{X}_j = c_{\overline{X}j}\,;\, \mu = \overline{X}_j) \\
&= 1 - \Phi\left(\frac{N_J[c_{\overline{X}J} - c_{\overline{X}j}]}{\sigma\sqrt{N_J - N_j}}\right) \\
C_{Edj} &= Pr(\overline{X}_J < d_{\overline{X}J} \,|\, \overline{X}_j = d_{\overline{X}j}\,;\, \mu = \overline{X}_j) \\
&= \Phi\left(\frac{N_J[d_{\overline{X}J} - d_{\overline{X}j}]}{\sigma\sqrt{N_J - N_j}}\right)
\end{aligned} \tag{6.11}$$

Again it can be seen that these functions are closely related to the test statistic on the conditional probability scale defined by eqn (1.8). That is, $C_{Eaj} = C_j(a_{\overline{X}J}, \mu = \overline{X}_j)$ when $\overline{X}_j = a_{\overline{X}j}$, $C_{Ebj} = 1 - C_j(b_{\overline{X}J}, \mu = \overline{X}_j)$ when $\overline{X}_j = b_{\overline{X}j}$, $C_{Ecj} = C_j(c_{\overline{X}J}, \mu = \overline{X}_j)$ when $\overline{X}_j = c_{\overline{X}j}$, and $C_{Edj} = 1 - C_j(d_{\overline{X}J}, \mu = \overline{X}_j)$ when $\overline{X}_j = d_{\overline{X}j}$.

We can also evaluate the stopping boundaries with respect to the predictive probability that an opposite decision might be made at the final analysis, where the predictive probability is computed by conditioning on the value of the test statistic at the boundary and averaging over the posterior distribution $\lambda(\mu \,|\, \overline{X}_j)$. For instance, based on a noninformative prior distribution for μ ($\mu \sim \mathcal{N}(\zeta, \tau^2)$ and taking the limit as $\tau^2 \to \infty$) this then yields

$$\begin{aligned}
H_{aj} &= \int Pr(\overline{X}_J > a_{\overline{X}J} \,|\, \overline{X}_j = a_{\overline{X}j}, \mu)\, \lambda(\mu \,|\, \overline{X}_j = a_{\overline{X}j}) \\
&= 1 - \Phi\left(\frac{N_J[a_{\overline{X}J} - a_{\overline{X}j}]}{\sigma\sqrt{\frac{N_J}{N_j}N_J - N_j}}\right) \\
H_{bj} &= \int Pr(\overline{X}_J < b_{\overline{X}J} \,|\, \overline{X}_j = b_{\overline{X}j}, \mu)\, \lambda(\mu \,|\, \overline{X}_j = b_{\overline{X}j}) \\
&= \Phi\left(\frac{N_J[b_{\overline{X}J} - b_{\overline{X}j}]}{\sigma\sqrt{\frac{N_J}{N_j}N_J - N_j}}\right) \\
H_{cj} &= \int Pr(\overline{X}_J > c_{\overline{X}J} \,|\, \overline{X}_j = c_{\overline{X}j}, \mu)\, \lambda(\mu \,|\, \overline{X}_j = c_{\overline{X}j}) \\
&= 1 - \Phi\left(\frac{N_J[c_{\overline{X}J} - c_{\overline{X}j}]}{\sigma\sqrt{\frac{N_J}{N_j}N_J - N_j}}\right) \\
H_{dj} &= \int Pr(\overline{X}_J < d_{\overline{X}J} \,|\, \overline{X}_j = d_{\overline{X}j}, \mu)\, \lambda(\mu \,|\, \overline{X}_j = d_{\overline{X}j}) \\
&= \Phi\left(\frac{N_J[d_{\overline{X}J} - d_{\overline{X}j}]}{\sigma\sqrt{\frac{N_J}{N_j}N_J - N_j}}\right)
\end{aligned} \tag{6.12}$$

It can be seen that these functions are closely related to the test statistic on the predictive probability scale defined by eqn (1.9) with $\tau^2 = \infty$. That is, $H_{aj} = H_j(a_{\overline{X}J}, \zeta, \infty)$ when $\overline{X}_j = a_{\overline{X}j}$, $H_{bj} = 1 - H_j(b_{\overline{X}J}, \zeta, \infty)$ when $\overline{X}_j = b_{\overline{X}j}$, $H_{cj} = H_j(c_{\overline{X}J}, \zeta, \infty)$ when $\overline{X}_j = c_{\overline{X}j}$, and $H_{dj} = 1 - H_j(d_{\overline{X}J}, \zeta, \infty)$ when $\overline{X}_j = d_{\overline{X}j}$.

6.6 Bayesian Posterior Probabilities

The Bayesian properties of a particular stopping rule can be evaluated for a specified prior by considering the posterior probabilities of the various hypotheses. As discussed in section 2, we will consider posterior probabilities that are associated with rejection of the hypotheses. Hence for prior distribution $\mu \sim \mathcal{N}(\zeta, \tau^2)$ we define

$$\begin{aligned}
B_{aj} &= Pr(\mu < \mu_{0-} \,|\, \overline{X}_j = a_{\overline{X}j}) \\
&= \Phi\left(\frac{\mu_{0-}[N_j\tau^2 + \sigma^2] - N_j\tau^2 a_{\overline{X}j} - \sigma^2\zeta}{\sigma\tau\sqrt{N_j\tau^2 + \sigma^2}}\right) \\
B_{bj} &= Pr(\mu < \mu_{-} \,|\, \overline{X}_j = b_{\overline{X}j}) \\
&= \Phi\left(\frac{\mu_{-}[N_j\tau^2 + \sigma^2] - N_j\tau^2 b_{\overline{X}j} - \sigma^2\zeta}{\sigma\tau\sqrt{N_j\tau^2 + \sigma^2}}\right) \\
B_{cj} &= Pr(\mu > \mu_{+} \,|\, \overline{X}_j = c_{\overline{X}j}) \\
&= 1 - \Phi\left(\frac{\mu_{+}[N_j\tau^2 + \sigma^2] - N_j\tau^2 c_{\overline{X}j} - \sigma^2\zeta}{\sigma\tau\sqrt{N_j\tau^2 + \sigma^2}}\right) \\
B_{dj} &= Pr(\mu > \mu_{0+} \,|\, \overline{X}_j = d_{\overline{X}j}) \\
&= 1 - \Phi\left(\frac{\mu_{0+}[N_j\tau^2 + \sigma^2] - N_j\tau^2 d_{\overline{X}j} - \sigma^2\zeta}{\sigma\tau\sqrt{N_j\tau^2 + \sigma^2}}\right)
\end{aligned} \tag{6.13}$$

It can be seen that these functions are closely related to the test statistic on the Bayesian posterior probability scale defined by eqn (1.6). That is, $B_{aj} = 1 - B_j(\zeta, \tau^2, \mu_{0-})$ when $\overline{X}_j = a_{\overline{X}j}$, $B_{bj} = 1 - B_j(\zeta, \tau^2, \mu_{-})$ when $\overline{X}_j = b_{\overline{X}j}$, $B_{cj} = B_j(\zeta, \tau^2, \mu_{+})$ when $\overline{X}_j = c_{\overline{X}j}$, and $B_{dj} = B_j(\zeta, \tau^2, \mu_{0+})$ when $\overline{X}_j = d_{\overline{X}j}$.

7 Sample Size Determination

From the discussion in sections 1.6 and 5.3, it is clear that equivalent group sequential tests can be specified either for the original data or for the standardizing transformation of the data. In group sequential test design, the standardized problem allows us to determine sample sizes which provide desired operating characteristics for specific alternative hypotheses. For instance, when considering a one-sided level α hypothesis test of $H_0 : \mu = \mu_0$ we might have a specified alternative hypothesis, say, $H_1 : \mu = \mu_1$, for which we desire some level of statistical power, say $\beta(\mu_1) = \beta$. These constraints suggest that we want the group sequential test in the standardized setting to have operating characteristics $\beta^*_+(0) = \alpha$ and $\beta^*(\delta_1) = \beta$, where δ_1 and μ_1 are related by eqn (1.33).

When determining sample size, however, we have not yet determined the exact values of $N_1, \ldots, N_J$. In the standardized problem, however, we only need to know the relative sizes of the N_j's. That is, the density in eqns (5.13) - (5.15) depends on the N_k's only through the values $\Pi_j = (N_j/N_J)$. Thus, so long as we specify the values of J and $(\Pi_1 = N_1/N_J, \ldots, \Pi_J = 1)$, we can compute the density of the test statistic (M, S^*) for the standardized problem. Determination of the group sequential test design and sample size than proceeds in the following stages.

1. Search for standardized stopping boundaries having desirable operating characteristics on the standardized scale, where the characteristics defined as desirable might be any of those described in section 6.

2. Search for the standardized alternative δ_1 such that $\beta^*(\delta_1) = \beta$. Note that for the purposes of study design $\beta^*(\delta_1)$ is most typically either $\beta^*_+(\delta_1)$ or $\beta^*_-(\delta_1)$, rather than the sum of the upper and lower power functions.

3. Solve for the sample size $N = N_J$ by using the relation eqn (1.33) to obtain

$$N_J = \frac{\delta_1^2 \sigma^2}{(\mu_1 - \mu_0)^2} \tag{7.1}$$

The sample size at the kth analysis is then $N_j = \Pi_j N_J$.

It is often the case that the maximum sample size is constrained by other considerations. In this case, we would use the sampling density to determine the value of μ_+ for which the group sequential test has $\beta_+(\mu_+) = \beta_u$.

We note parenthetically that in later sections we describe families of group sequential designs which are parameterized in part by β_u and β_ℓ, which will continue to have interpretations as the power of the study under certain alternatives. Even in those settings, however, it is possible to choose sample size based on some other alternative μ_1 and a desired level of statistical power β_1. We then use the relationships $\beta_u = \beta_+(\mu_+)$ and $\beta_\ell = \beta_-(\mu_-)$ to define the values of μ_+ and μ_-.

8 General Framework for Families of Group Sequential Stopping Rules

In section 2, we discussed the desirability of framing a hypothesis test in such a way as to allow more precise interpretation of a failure to reject H_0. The deficiencies of classical hypothesis testing in this regard become even more evident when using a group sequential design. However, the application of the strategy adopted in section 2 is not always straightforward.

There is no particular problem in applying the model specified in section 2 to the case of one-sided group sequential designs (e.g., designs A - C in section 4 above). That is, providing we have chosen a stopping boundary that has the desired level of significance ($\beta_+(\mu_0) = \alpha$ for a test of H_0 versus H_+ or $\beta_-(\mu_0) = \alpha$ for a test of H_0 versus H_-) we can always find the appropriate alternative μ_+ or μ_- for which the test has statistical power β_u or β_ℓ, respectively. Similarly, for two-sided designs (e.g., designs D and E in section 4), if we have chosen a stopping boundary that has $\beta_+(\mu_0) = \alpha/2$ and $\beta_-(\mu_0) = \alpha/2$, we can find the alternatives μ_+ and μ_- such that $\beta^*_+(\mu_+) = \beta_u$ and $\beta^*_-(\mu_-) = \beta_\ell$.

There are, however, designs intermediate to the one-sided and two-sided designs. These are designs which would have, for instance, $\beta^*_+(\mu_0) = \alpha/2$ and $\beta^*_-(\mu_0) > \alpha/2$. One use of such designs would be when comparing a new treatment to a standard treatment when the goal is to show that the treatments are roughly equivalent with respect to some primary endpoint (e.g., survival), but that the new treatment is superior with respect to a secondary endpoint (e.g., quality of life). In such a situation, we may not want to use a design which treats the two therapies symmetrically. If we observed a trend for the new treatment to be worse with respect to the primary endpoint, we might be unwilling to continue the trial to show statistically that it is actually worse than the standard therapy. On the other hand, the requirements for the burden of proof may be such that in order to abandon the standard therapy, we would need to have a result that is highly statistically significant.

In the above discussion, we have implicitly parameterized these intermediate tests by the asymmetry of the upper and lower power functions under the null hypothesis. In keeping with the philosophy presented in section 2, however, we would like to maintain common standards of evidence for rejection of hypotheses. Hence we consider an alternative parameterization of the intermediate tests based on the hypotheses rejected by each of the stopping boundaries.

To implement this approach, we thus describe each of the four potential stopping boundaries ('a', 'b', 'c', or 'd') as having two fundamental determinants: the hypothesis μ_* being rejected by the boundary and a boundary shape function $v_*(\Pi_j)$ describing the relationship between the boundaries of the continuation sets at successive analyses. The ways in which the hypotheses and the boundary shape function is used shall differ according to the scale of the group sequential test statistic which will be compared to the boundaries. However, in all cases, a two-sided hypothesis test will be viewed as the superposition of two one-sided tests: an upper hypothesis test of $H_{0+} : \mu \leq \mu_{0+}$ versus $H_+ : \mu \geq \mu_+$, and a lower hypothesis test of $H_{0-} : \mu \geq \mu_{0-}$ versus $H_- : \mu \leq \mu_-$, subject to the constraints

$$\mu_- \leq \mu_{0+} \leq \mu_{0-} \leq \mu_+ \tag{8.1}$$

The size of the upper and lower tests will be denoted α_u and α_ℓ, respectively. Similarly, the power of the upper and lower tests to detect their respective alternative hypotheses will be denoted β_u and β_ℓ, respectively. The individual hypotheses of the superposed hypothesis tests are associated with the hypotheses rejected by each of the four stopping boundaries according to

$$\begin{aligned} \mu_a &= \mu_{0-} \\ \mu_b &= \mu_- \\ \mu_c &= \mu_+ \\ \mu_d &= \mu_{0+} \end{aligned} \tag{8.2}$$

In this representation of a hypothesis test, we can obtain the classic one- and two-sided hypothesis tests through appropriate choices of the four hypotheses:

1. A one-sided hypothesis test of the null hypothesis $H_0 : \mu \leq \mu_0$ versus a greater alternative $H_+ : \mu \geq \mu_1$ is obtained by choosing the null and alternative hypotheses of the upper hypothesis test to correspond to the desired one-sided hypothesis tests: $\mu_{0+} = \mu_0$ and $\mu_+ = \mu_1$. The lower hypothesis test is then chosen to be superposed exactly on top of that upper hypothesis test. Determination of the exact correspondence between the hypotheses of the upper and lower hypothesis tests will of course depend upon the values chosen for the size α_ℓ and power β_ℓ for the lower test. However, it is easy to see that if $\alpha_\ell = 1 - \beta_u$ and $\beta_\ell = 1 - \alpha_u$, the desired coincident tests are obtained by setting $\mu_{0-} = \mu_+$ and $\mu_- = \mu_{0+}$.

2. A one-sided hypothesis test of the null hypothesis $H_0 : \mu \geq \mu_0$ versus a lesser alternative $H_- : \mu \leq \mu_1$ is obtained by choosing the null and alternative hypotheses of the lower hypothesis test to correspond to the desired one-sided hypothesis tests: $\mu_{0-} = \mu_0$ and $\mu_- = \mu_1$. The upper hypothesis test is then chosen to be superposed exactly on top of that lower hypothesis test. Determination of the exact correspondence between the hypotheses of the upper and lower hypothesis tests will of course depend upon the values chosen for the size α_u and power β_u for the upper test. However, it is easy to see that if $\alpha_u = 1 - \beta_\ell$ and $\beta_u = 1 - \alpha_\ell$, the desired coincident tests are obtained by setting $\mu_{0+} = \mu_-$ and $\mu_+ = \mu_{0-}$.

3. A classical two-sided hypothesis test of the null hypothesis $H_0 : \mu = \mu_0$ versus two-sided alternative $H_1 : \mu \neq \mu_0$ with power β to reject the null hypothesis when $\mu = \mu_1$ is obtained by choosing the null hypotheses of the lower and upper hypothesis tests to each correspond to the null hypothesis: $\mu_{0+} = \mu_{0-} = \mu_0$. The alternative hypotheses of the lower and upper tests are then set according to the value of μ_1: If $\mu_1 > \mu_0$, then we choose $\mu_+ = \mu_1$, and if $\mu_1 < \mu_0$, we choose $\mu_- = \mu_1$. The alternative hypothesis that is not set equal to μ_1 is determined from the corresponding choice of statistical power.

We formalize this approach for other hypothesis tests intermediate to these classical tests by parameterizing the shifts of the upper and lower hypothesis tests. We define shift parameters $0 \leq \epsilon_u \leq 1$ and $0 \leq \epsilon_\ell \leq 1$ for the upper and lower hypothesis tests, respectively. The parameterization is such that when the shift parameter is zero, it is not of interest to discriminate between the hypotheses of the corresponding hypothesis test. That is, when $\epsilon_u = 0$, it is not of interest to discriminate between a null hypothesis $\mu = \mu_0$ and a greater alternative $\mu > \mu_0$. Similarly, when $\epsilon_\ell = 0$, it is not of interest to discriminate between a null hypothesis $\mu = \mu_0$ and a lesser alternative $\mu < \mu_0$.

The exact parameterization of the hypothesis shifts are based on the classical hypotheses of a two sided hypothesis test (μ_0, μ_+, and μ_- as described in section 2) and some maximal shift Δ_μ. The hypothesis rejected by each boundary is defined by

$$\begin{aligned} \mu_a &= \mu_0 + [1 - \epsilon_\ell]\Delta_\mu \\ \mu_b &= \mu_- + [1 - \epsilon_\ell]\Delta_\mu \\ \mu_c &= \mu_+ - (1 - \epsilon_u)\Delta_\mu \\ \mu_d &= \mu_0 - (1 - \epsilon_u)\Delta_\mu \end{aligned} \tag{8.3}$$

From eqns (8.1) - (8.3), and by considering the maximal shift of the hypotheses when $\epsilon_\ell + \epsilon_u = 1$, we find that

$$\Delta_\mu \leq \min(\mu_+ - \mu_0, \mu_0 - \mu_-). \tag{8.4}$$

From eqn (8.3), we can see that when $\epsilon_\ell = \epsilon_u = 1$, both the 'a' and 'd' boundaries reject the hypothesis that $\mu = \mu_0$, and a two-sided hypothesis test is obtained. Furthermore note that if

$\epsilon_u = 0$ and $\epsilon_\ell = 1$, then $\mu_d = \mu_b = \mu_-$ and $\mu_c = \mu_a = \mu_0$ when $\Delta_\mu = \mu_0 - \mu_- = \mu_+ - \mu_0$. Similarly, if $\epsilon_u = 1$ and $\epsilon_\ell = 0$, then $\mu_b = \mu_d = \mu_+$ and $\mu_a = \mu_c = \mu_0$ when $\Delta_\mu = \mu_0 - \mu_- = \mu_+ - \mu_0$. Thus we obtain one-sided tests with coincident hypotheses for the upper and lower hypothesis tests.

Choices of $\epsilon_u + \epsilon_\ell$ between 1 and 2 result in tests that are in some sense intermediate to one-sided tests (when $\epsilon_u + \epsilon_\ell = 1$) and two-sided tests (when $\epsilon_u + \epsilon_\ell = 2$). Moving ϵ_ℓ from 1 to 0 corresponds to deciding that it is unimportant to distinguish between H_- and H_0 as defined in section 2. Analogous interpretations hold as ϵ_u is decreased. Of special note is the choice $\epsilon_u = \epsilon_\ell = 0.5$, which corresponds to a test design which is sometimes used in one-sided equivalence (noninferiority) testing.

9 Parameterizations for Boundary Shifts for Group Sequential Families

The general framework for group sequential tests described in section 8 must be implemented in slightly different fashions depending upon the scale used for the group sequential test statistic. That is, the way in which the hypothesis and the boundary shape function are used to construct a group sequential stopping rule is different for the various test statistic scales.

9.1 Unified Family of Group Sequential Test Designs (Sample Mean Scale)

In the unified family of group sequential test designs described by [7], a family of group sequential designs is parameterized on the sample mean statistic scale in part because on this scale the boundary shapes are invariant to shifts in the value of μ_*. In this family, the stopping boundaries at the jth analysis are determined from the hypothesis μ_* being rejected by the boundary and the boundary shape function $v(\Pi_j)$ according to

$$\begin{aligned}
d_{\overline{X}j} &= \mu_d + v_d(\Pi_j) \\
c_{\overline{X}j} &= \begin{cases} \mu_c - v_c(\Pi_j) & \text{if } \mu_c - v_c(\Pi_j) > \mu_b + v_b(\Pi_j) \\ (d_{\overline{X}j} + a_{\overline{X}j})/2 & \text{else} \end{cases} \\
b_{\overline{X}j} &= \begin{cases} \mu_b + v_b(\Pi_j) & \text{if } \mu_c - v_c(\Pi_j) > \mu_b + v_b(\Pi_j) \\ (d_{\overline{X}j} + a_{\overline{X}j})/2 & \text{else} \end{cases} \\
a_{\overline{X}j} &= \mu_a - v_a(\Pi_j)
\end{aligned} \tag{9.1}$$

On the nonstandardized sample mean scale, the boundary shape functions $v_*(\Pi)$ will depend also on the maximal sample size $N = N_J$. At the time of study design it is most convenient to work on the standardized scale, in which case the stopping boundaries at the jth analysis will depend upon the δ_*'s and the standardized boundary shape functions $v_a^*(\cdot)$, $v_b^*(\cdot)$,$v_c^*(\cdot)$, and $v_d^*(\cdot)$.

$$\begin{aligned}
d^*_{\overline{X}j} &= \delta_d + v_d^*(\Pi_j) \\
c^*_{\overline{X}j} &= \begin{cases} \delta_c - v_c^*(\Pi_j) & \text{if } \delta_c - v_c^*(\Pi_j) > \delta_b + v_b^*(\Pi_j) \\ (d^*_{\overline{X}j} + a^*_{\overline{X}j})/2 & \text{else} \end{cases} \\
b^*_{\overline{X}j} &= \begin{cases} \delta_b + v_b^*(\Pi_j) & \text{if } \delta_c - v_c^*(\Pi_j) > \delta_b + v_b^*(\Pi_j) \\ (d^*_{\overline{X}j} + a^*_{\overline{X}j})/2 & \text{else} \end{cases} \\
a^*_{\overline{X}j} &= \delta_a - v_a^*(\Pi_j)
\end{aligned} \tag{9.2}$$

In the next section, we consider a specific form for the boundary shape functions $v_*(\Pi_k)$. Here we merely note that constraints (4.4) and (4.5) are satisfied if the boundary shape functions are monotonically nonincreasing in Π_k and if the hypotheses rejected by the boundaries satisfy the following constraints.

$$\begin{aligned}
\delta_c - \delta_d &= v_d^*(1) + v_c^*(1) \equiv \delta^+ \\
\delta_a - \delta_b &= v_b^*(1) + v_a^*(1) \equiv -\delta^- \\
\delta_a - \delta_d &\le v_d^*(1) + v_a^*(1) \equiv \delta^\#
\end{aligned} \tag{9.3}$$

where $\delta^\#$, a standardized form of Δ_μ represents the maximal shift of the lowest boundary (specified by the $a^*_{\overline{X}k}$'s) toward the uppermost boundary (specified by the $d^*_{\overline{X}k}$'s).

Applying the general framework of eqn (8.3) to this standardized setting thus results in

$$\begin{aligned}\delta_d &= (\epsilon_u - 1)\delta^{\#} \\ \delta_c &= (\epsilon_u - 1)\delta^{\#} + \delta^{+} \\ \delta_b &= (1 - \epsilon_\ell)\delta^{\#} + \delta^{-} \\ \delta_a &= (1 - \epsilon_\ell)\delta^{\#}\end{aligned} \tag{9.4}$$

where $2 \geq \epsilon_u + \epsilon_\ell \geq 1$.

9.2 Partial Sum Scale

The partial sum scale is a straightforward transformation of the sample mean scale, and thus the framework of section 9.1 is easily modified to apply to the partial sum statistic scale. In the standardized setting, the stopping boundaries are of the form

$$\begin{aligned}d^*_{Sj} &= \delta_d + v^*_d(\Pi_j) \\ c^*_{Sj} &= \begin{cases}\delta_c - v^*_c(\Pi_j) & \text{if } \delta_c - v^*_c(\Pi_j) > \delta_b + v^*_b(\Pi_j) \\ (d^*_{Sj} + a^*_{Sj})/2 & \text{else}\end{cases} \\ b^*_{Sj} &= \begin{cases}\delta_b + v^*_b(\Pi_j) & \text{if } \delta_c - v^*_c(\Pi_j) > \delta_b + v^*_b(\Pi_j) \\ (d^*_{Sj} + a^*_{Sj})/2 & \text{else}\end{cases} \\ a^*_{Sj} &= \delta_a - v^*_a(\Pi_j)\end{aligned} \tag{9.5}$$

It should be noted that the straightforward transformation from the standardized partial sum scale to the sample mean scale means that the partial sum scale family can be regarded as the sample mean scale with an alternative parameterization of the boundary shape function. That is, because $d^*_{Sj}/Pi_j = d^*_{\overline{X}j}$, the same stopping rule that is obtained with boundary shape function $v^*_d(\Pi_j)$ in the partial sum family would be obtained with boundary shape function $\Pi_j v^*_d(\Pi_j)$ in the sample mean family. This correspondence is the way that this family is implemented in S+SeqTrial.

9.3 Normalized Z Statistic Scale

The normalized Z statistic scale is a straightforward transformation of the sample mean scale, and thus the framework of section 9.1 is easily modified to apply to this scale. In the standardized setting, the stopping boundaries are of the form

$$\begin{aligned}d^*_{Zj} &= \delta_d + v^*_d(\Pi_j) \\ c^*_{Zj} &= \begin{cases}\delta_c - v^*_c(\Pi_j) & \text{if } \delta_c - v^*_c(\Pi_j) > \delta_b + v^*_b(\Pi_j) \\ (d^*_{Zj} + a^*_{Zj})/2 & \text{else}\end{cases} \\ b^*_{Zj} &= \begin{cases}\delta_b + v^*_b(\Pi_j) & \text{if } \delta_c - v^*_c(\Pi_j) > \delta_b + v^*_b(\Pi_j) \\ (d^*_{Zj} + a^*_{Zj})/2 & \text{else}\end{cases} \\ a^*_{Zj} &= \delta_a - v^*_a(\Pi_j)\end{aligned} \tag{9.6}$$

It should be noted that the straightforward transformation from the standardized partial sum scale to the sample mean scale means that the partial sum scale family can be regarded as the sample mean scale with an alternative parameterization of the boundary shape function. That is, because $d^*_{Zj}/\sqrt{Pi_j} = d^*_{\overline{X}j}$, the same stopping rule that is obtained with boundary shape function $v^*_d(\Pi_j)$ in the partial sum family would be obtained with boundary shape function $\sqrt{\Pi_j} v^*_d(\Pi_j)$ in the sample mean family. This correspondence is the way that this family is implemented in S+SeqTrial.

9.4 Error Spending Scale

Stopping boundaries for the error spending scale will be based on the error spending functions. In the definition of the error spending scales (eqns (1.10) and (1.11)) and the error spending functions (eqns (6.7) and (6.8)), the various hypotheses being rejected by their respective boundaries appear in the computation of the error spending statistics in a complicated fashion along with the stopping boundaries at the various analysis times. In defining stopping boundaries based on the error spending statistics, the dependence of the stopping boundaries on the hypothesis being rejected will be through which error spending function is related to the boundary shape function. Hence, the stopping boundaries will be defined by

$$\begin{aligned} E_d(\Pi_j) &= v_d(\Pi_j) \\ E_c(\Pi_j) &= v_c(\Pi_j) \\ E_b(\Pi_j) &= v_b(\Pi_j) \\ E_a(\Pi_j) &= v_a(\Pi_j) \end{aligned} \tag{9.7}$$

The way that this definition can be used to define stopping boundaries is illustrated by considering the 'd' boundary. The function $E_d(\Pi_j)$ represents a probability that the group sequential statistic S_j will exceed d_{Sj} under the hypothesis μ_d. Eqn (9.7) stipulates that d_{Sj} must be chosen such that $E_d(\Pi_j)$ is exactly equal to the value given by the boundary shape function $v_d(\Pi_j)$. Clearly this is an implicitly defined value– no closed form solution for d_{Sj} is possible.

In the formulation of the error spending statistic boundaries, the value of Δ_μ will be taken to be the upper bound on the range specified by eqn (8.4): $\Delta_\mu = \min(\mu_+ - \mu_0, \mu_0 - \mu_-)$.

10 Parameterizations for Boundary Shape Functions for Group Sequential Families

In specifying designs A - E in section 4 above, one or more of the boundaries of the continuation sets were described as arbitrary. Generally, the space of all possible choices of those arbitrary boundaries which would provide the desired operating characteristics for a group sequential test is too large to work with easily. In the last section, we described a partial parameterization of the boundaries which makes the search for stopping rules more tractable. In this section, we describe a family of boundary shape functions that impose a functional relationship on successive points along the continuation set boundaries.

For notational convenience, the boundary shape functions are described on the standardized scales. The general form of the boundary shape functions used in S+SeqTrial is given by

$$v^*(\Pi) = \{A + \Pi^{-P}[1 - \Pi]^R\}G \tag{10.1}$$

for specified parameter P, R, and A. As a general rule, the constant G is found in order to provide desired type I error and statistical power. The ranges of valid choices for the parameters P, R, and A depend upon the group sequential test statistic scale for which the boundaries are defined. Nevertheless, from the above, we can deduce the basic roles that each of the parameters play in determining a stopping boundary. All of the parameters can be thought of as relating to the conservatism of the decision to terminate the study at the earliest analyses. The way in which they affect that conservatism in terms of the shape of the stopping boundary is very different, however.

1. P, when positive, is a measure of conservatism at the earliest analyses: The higher the value of P, the more difficult it will be for a study to terminate at t he earliest analyses. When P is infinite, the stopping boundary is infinite at all interim analyses.

2. P when negative, is a measure of conservatism at the earliest analyses: The more negative the value of P, the more difficult it will be for a study to terminate at the earliest analyses. Exactly how difficult it will be to terminate at the earliest analysis relative to the final analysis will be affected by the value of the A parameter. (Note that it is difficult to compare the degree of conservatism for positive P and negative P, as the boundary shape is quite different.)

3. R, when positive, is a measure of lack of conservatism at the earliest analyses: The higher the value of R, the less difficult it will be for the study to terminate at the earliest analyses. The degree to which the value of R can affect the conservatism at the earliest analyses is greatly affected by the values of P and A. If P is also positive, the R parameter affects the curvature of the stopping boundary at the later analyses, but the P parameter has the greatest influence on the conservatism at the earliest analyses.

4. A is a measure of separation between the first and last analyses, and thus can affect the conservatism of the test overall. When the G critical value is positive (as tends to be the case when P is positive or zero), a larger value of A tends to make the design less conservative at the earlier analyses. When the G critical value is negative, then A tends to be negative, and a more negative value of A makes to design less conservative at the earlier analyses. This behavior can be deduced for some cases from the fact that when the magnitude of A is large, the difference between A and A+1 is less substantial.

10.1 Unified Family of Group Sequential Test Designs (Sample Mean Scale)

There are two general forms of boundary shape functions that have received substantial attention in the statistical literature. [16] and Whitehead (1992) consider boundary shape functions which are linear in Π_j on the partial sum statistic scale. That is, a plot of, say, the d_{Sj}'s versus Π_j

produces a straight line. The computer package PEST3 implements such boundaries for continuous monitoring (so $n_j = 1$ for all $j = 1, \ldots, J$) with an approximation used for group sequential tests. Such linear boundaries have been implemented in PEST3 for stopping rules of types A, C, D, and E (as described in section 4) for situations in which all stopping boundaries use the same boundary shape function.

[13] consider boundary shape functions which are powers of Π_j on the partial sum scale. Specifically, they examined tests of type D having $d_{Sj} = -a_{Sj} = \Pi_j^{\Delta}$ where $0 \leq \Delta \leq 0.5$. This family includes the [11] and [8] designs as special cases. These boundary shape functions were then extended to stopping rules of types C and E by [3] and [9]. They were used in stopping rules of type A and B by Emerson (1988). The software package EaSt implements these boundary shape functions for stopping rules of types A, C, D, and E for situations in which all stopping boundaries use the same boundary shape function.

We consider here a boundary shape function which unifies these two families, as well as extending them to include additional boundary shapes. In this parameterization, we use parameters A_*, P_*, and R_*, and critical value G_* to define

$$v_*(\Pi_k) = \{A_* + \Pi_k^{-P_*}[1 - \Pi_k]^{R_*}\}G_*.$$

As described in the next section, the parameters A_*, P_*, and R_* are usually specified by a user subject to constraints outlined below, and the critical value G_* is usually found in a computer search to obtain desired operating characteristics.

The above boundary shape function includes the following special cases:

1. $A_* = 0$, $P_* \geq 0$, $R_* = 0$: This corresponds to the [13] family of boundary shape functions extended to the range considered by [3], although the parameterization is different. In the current parameterization, the choice $P_* = 0.5$ corresponds to a [11] boundary shape, and the choice $P_* = 1$ corresponds to an [8] boundary shape. In general, P_* is a measure of the tendency of the stopping rule to test conservatively at the earliest analyses, with larger values of P_* corresponding to greater early conservatism.

2. $A_* = 1$, $P_* = 1$, $R_* = 0$: This corresponds to the boundary shape function used in the triangular and double triangular tests of [16].

3. A_* unconstrained, $P_* = 1$, $R_* = 0$: This corresponds to the boundary shape function used in the restricted procedures described by Whitehead (1992) and implemented in PEST3.

4. A_* unconstrained, $P_* = 1$, $R_* = 0$: This corresponds to the boundary shape function used in the restricted procedures described by Whitehead (1992) and implemented in PEST3.

5. A_* unconstrained, $P_* = 0.5$, $R_* = 0.5$: This corresponds to the sequential conditional probability ratio tests described by [17].

6. $A_* = 1$, $P_* = 1$, $R_* = 1$: This is an alternative parameterization of the [8] tests.

It is useful to examine the behavior for this family of boundary shape functions over a range of parameter choices. First we note that the boundary shape function is monotonically nonincreasing in Π_k only if $R_* \geq 0$. Furthermore, it may not be possible to find critical values which provide the desired operating characteristics for arbitrary choices of A_*. In general, however, the general behavior of the boundary shape function can be determined from the following table.

Range of P_*	Range of R_*	Value of $v_*(0)$	Value of $v_*(1)$	Concavity
(0,∞)	[1,∞]	∞	A_*G_*	upward
(0,∞)	(0,1)	∞	A_*G_*	upward then downward
(0,∞)	0	∞	$(A_*+1)G_*$	upward
0	(1,∞]	$(A_*+1)G_*$	A_*G_*	upward
0	1	$(A_*+1)G_*$	A_*G_*	none (line)
0	(0,1)	$(A_*+1)G_*$	A_*G_*	downward
0	0	$(A_*+1)G_*$	$(A_*+1)G_*$	none (line)
(-1,0)	0	A_*G_*	$(A_*+1)G_*$	downward
-1	0	A_*G_*	$(A_*+1)G_*$	none (line)
(-∞,-1)	0	A_*G_*	$(A_*+1)G_*$	upward

Of the three boundary shape function parameters, all tend to control the degree of conservatism used in stopping at the earliest analyses. The parameter P_* is perhaps the most interpretable of these, as larger values of P_* make it increasingly difficult to terminate a study at the earliest analyses. As discussed below, a value of $P_* = \infty$ will preclude early stopping for the corresponding boundary. Note that when $P_* \leq 0$, the stopping boundary is finite even when no data has been collected.

Generally, it can be seen that the boundary shape functions previously described and implemented in commercially available software packages are concave upward. Through expanding the range of the parameter P_*, as well as introducing R_*, we have included boundary shapes which are concave downward. Of particular note is the case where $P_* > 0$ and $0 < R_* < 1$, when the boundary shape is concave upward for $\Pi_k < (P_* - \sqrt{(P_*R_*/(P_* - R_* + 1))})/(P_* - R_*)$.

In the group sequential tests defined in (9.2), we allow each of the four potential boundaries to have its own boundary shape function. That is, we can choose A_*, P_*, R_* separately for each of the four boundaries specified by the $d_{\overline{X}k}$'s, the $c_{\overline{X}k}$'s, the $b_{\overline{X}k}$'s and the $a_{\overline{X}k}$'s. This is an extension of the designs described previously in the statistical literature, but one which facilitates the exploration of candidate stopping rules for a particular clinical trial. This is effected by the fact that such flexibility allows the basic types of designs described in section 4 to be joined by a continuous parameter.

To see this, note that when $P_* = \infty$, the corresponding boundary allows no early stopping. That is, we can construct each of the 5 types of designs in section 4 by considering only design type E with suitable choices of the boundary shape function parameters.

A. A single upper early stopping boundary: P_d arbitrary, $P_c = P_b = P_a = \infty$

1. Test of H_0 versus H_+: $\epsilon_u = 1$, $\epsilon_\ell = 0$
2. Test of H_0 versus H_-: $\epsilon_u = 0$, $\epsilon_\ell = 1$

B. A single lower early stopping boundary: P_a arbitrary, $P_d = P_c = P_b = \infty$.

1. Test of H_0 versus H_-: $\epsilon_u = 0$, $\epsilon_\ell = 1$
2. Test of H_0 versus H_+: $\epsilon_u = 1$, $\epsilon_\ell = 0$

C. Lower and upper early stopping boundaries which meet at the final analysis: P_d, P_c, P_b, and P_a arbitrary (however, choosing $v_c(\Pi_j) = v_a(\Pi_j)$ and $v_b(\Pi_j) = v_d(\Pi_j)$ will tend to yield the most intuitive relationships among the critical values G_d, G_c, G_b, G_a, because when $\epsilon_u + \epsilon_\ell = 1$ the boundaries within each of those pairs are coincident).

1. Test of H_0 versus H_+: $\epsilon_u = 1$, $\epsilon_\ell = 0$

2. Test of H_0 versus H_-: $\epsilon_u = 0$, $\epsilon_\ell = 1$

D. Lower and upper early stopping boundaries which do not meet at the final analysis: P_d and P_a arbitrary, $P_c = P_b = \infty$; $\epsilon_u + \epsilon_\ell > 1$

E. Four boundary design: P_d, P_c, P_b, P_a arbitrary; $\epsilon_u + \epsilon_\ell > 1$

It should be noted that depending upon the exact choices of the boundary shape function parameters and the maximal number of analyses J, early stopping may not be possible under all four boundaries. This should be clear, given our ability to define stopping rules of types A, B, C, and D using the general structure of these four boundary designs.

10.2 Partial Sum Scale

As noted in section 9.2, the family of group sequential designs implemented with the boundary shape function given by eqn (10.1) on the partial sum scale can just be regarded as a family of group sequential designs implemented on the sample mean scale with boundary shape function

$$v^*(\Pi) = \Pi\{A + \Pi^{-P}[1 - \Pi]^R\}G \tag{10.2}$$

This is the way that this family is implemented in S+SeqTrial.

10.3 Normalized Z Statistic Scale

As noted in section 9.3, the family of group sequential designs implemented with the boundary shape function given by eqn (10.1) on the normalized Z statistic scale can just be regarded as a family of group sequential designs implemented on the sample mean scale with boundary shape function

$$v^*(\Pi) = \sqrt{\Pi}\{A + \Pi^{-P}[1 - \Pi]^R\}G \tag{10.3}$$

This is the way that this family is implemented in S+SeqTrial.

10.4 Error Spending Scale

A design family based on a generalization of the error spending function approach of Lan and DeMets (1983) and [10] is defined by setting the error spending function for each of the four possible stopping boundaries (the 'a', 'b', 'c', and 'd' boundaries) independently. The boundary shape function is again based on eqn (10.1) to define the cumulative proportion of the type I error (for the 'a' and 'd' boundaries) or type II error (for the 'b' and 'c' boundaries) that is spent at the analysis in which proportion $0 < \Pi_j < 1$ of the statistical information has been accrued. At the final analysis, it is assumed that all of the type I and type II error will have been spent, and thus all boundaries at the final analysis correspond to error spending functions of 1.

Boundaries on the error spending function range from 0 to 1. Because of this restricted range, boundary shape functions are only possible for certain combinations of the boundary shape function parameters:

1. Negative values of P (with $R = 0$). In this setting, P measures the early conservatism of the stopping rule with more negative values of P corresponding to stopping rules that have lower probabilities of terminating the study at the earliest analyses. A value of $P = -3.25$ approximates the operating characteristics of an O'Brien-Fleming boundary relationship for a one-sided type I error of .025 (although this error spending function based stopping rule does not exhibit the very extreme conservatism at the earliest analyses that is common with the O'Brien-Fleming boundary relationship).

2. Positive values of R (with $P = 0$). In this setting, as R increases the stopping rule becomes less conservative at the earliest analyses.

3. The interesting special case of $P = 0$ and $R = 0$ can be used to preclude early termination of the study.

In each case, the values of A and G are uniquely determined by the choice of P and R, thus you never need specify either of these latter two parameters when using the error spending function family.

11 Constrained Boundaries for Group Sequential Families

It can often happen that the stopping rule obtained from a parametric design family is unsatisfactory at one or more analyses. For instance, many clinical trialists find the extreme conservatism of the O'Brien-Fleming boundary relationships at the earliest analyses undesirable. One common modification of O'Brien-Fleming boundary relationships is to use the least extreme of the O'Brien-Fleming boundary or a critical value corresponding to a fixed sample two-sided P value of .001. In order to facilitate this type of modification of stopping rules, constraints on the boundaries at particular analyses can be specified, with all unconstrained boundaries being determined from a parametric design family in such a way to maintain the desired operating characteristics (size and power) of the study design.

Constraints on the boundaries can be

1. Exact constraints. You enter the exact stopping boundary desired for a particular boundary ('a', 'b', 'c', or 'd') at a specific analysis.

2. Minimum constraints. You enter a value for the stopping boundary that is the minimum value that you would like desired for a particular boundary ('a', 'b', 'c', or 'd') at a specific analysis. If the parametric design family would result in a higher threshold for early termination at that analysis time, the boundary from the parametric family will be used instead of this minimum constraint.

3. Maximum constraints. You enter a value for the stopping boundary that is the maximum value that you would like desired for a particular boundary ('a', 'b', 'c', or 'd') at a specific analysis. If the parametric design family would result in a lower threshold for early termination at that analysis time, the boundary from the parametric family will be used instead of this maximum constraint.

If the group sequential design family is based on the sample mean, partial sum, or normalized Z statistic scales, the boundary constraints can be specified on any valid boundary scale EXCEPT the error spending function scale. On the other hand, if the group sequential design family is based on the error spending scale, the boundary constraints can ONLY be specified on the error spending function scale.

When specifying the minimum or maximum constraints, the concept of "minimum" and "maximum" is based on the ordering of the sample mean statistic. That is, one boundary is less than another if the boundary is lower on the sample mean scale. This distinction is important because some boundary scales have a reverse ordering. For instance, because the fixed sample P value scale is measured on the scale of a P value for a one-sided test of an upper alternative regardless of the type of hypothesis test being designed, a higher boundary on the sample mean scale actually corresponds to a lower number on the fixed sample P value scale. Thus if you want to apply a constraint to avoid having the upper efficacy boundary of an O'Brien-Fleming test more extreme than the critical value of a fixed sample two-sided P value of .001, you would create a maximum constraint on the fixed sample P value scale that has .0005 in the appropriate position in the constraint matrix.

On the sample mean scale, the search for a particular group sequential design is thus effected through the following steps:

1. The user specifies a particular probability model for the problem, including the number of arms, a probability model for the response, and a summary measure for describing the response within treatment arms. In this specification, specific values for ψ, θ_0, $g(\cdot)$, and σ^2 are determined (see section 3).

2. The user specifies the desired operating characteristics

(a) The size of the upper test α_u

(b) The size of the lower test α_ℓ

(c) The power of the upper test β_u

(d) The power of the lower test β_ℓ

In keeping with the philosophy of [3] and the discussion of section 2, typical choices might generally be $\alpha_u = \alpha_\ell = \alpha$ and $\beta_u = \beta_\ell = 1 - \alpha$, although group sequential stopping rules are also well-defined for other choices of operating characteristics.

3. The user specifies the number J and timing of the analyses $\Pi_1, \Pi_2, \ldots, \Pi_J$. Several authors have found that the general operating characteristics of a design are fairly robust to slight variations in the number and timing of analyses, so for design purposes it is adequate to have a rough idea of these parameters. At the time of actual monitoring of the study, exact methods can be used to maintain the general behavior of the stopping boundaries while controlling the type I error exactly. We also note that in our standardized test, it is sufficient to merely specify the Π_j's, but it is also possible to specify the N_j's and compute the Π_j's from them. Lastly, because a fixed sample test is a special case of a group sequential test, and the usual fixed sample critical values will be found by choosing $J = 1$.

4. For each of the four potential boundaries, the user specifies the values for the three parameters A_*, P_*, R_* of the boundary shape functions (see section 10).

5. The user specifies the parameters ϵ_u and ϵ_ℓ which correspond to gradations between one- and two-sided hypothesis tests (see sections 8 and 9).

6. The computer searches for critical values G_d, G_c, G_b, G_a. In general, each of these critical values are dependent upon all of the design parameters specified in steps 2 - 5 above. That is, changing the values of A_d, P_d, R_d will affect not only the value of G_d, but it will also have a slight effect on the values of G_c, G_b, G_a. In this search,

   (a) values for the critical values are guessed, and the stopping boundaries computed using eqn (9.2) with the appropriate boundary shape functions as specified by eqn (10.1). The values of the standardized alternatives δ_+ and δ_- are easily computed according to eqn (9.3).

   (b) Each of the potential boundaries are then compared to any specified constraints, and any necessary modifications made.

   (c) The operating characteristics of the trial design are computed.

   (d) A new guess for the critical values is made using a Newtonian search with finite difference estimates of the Jacobian matrix.

   (e) When a design with acceptable precision for the operating characteristics is found, the search terminates.

7. The sample size for the trial is determined. Typically, this is done by either of the following two methods:

   (a) The user specifies which of the two alternative hypotheses (H_+ or H_-) is to be used for satisfying the power constraint, and the value of the natural parameter θ is specified for that hypothesis. The sample size N_J can then be computed using (30), with the value of the other alternative computed by substituting that value of N_J into (7.1) and solving for the value of μ_- (if H_+ was used to determine the sample size) or μ_+ (if H_- was used to determine the sample size).

(b) The user specifies the sample size N_J which is practical, and that value is then substituted into (7.1) to determine the values of μ_+ and μ_-.

8. A candidate design should then typically be evaluated for its unconstrained operating characteristics and stopping boundaries. Such evaluation might typically include

   (a) Power curves as a function of various specified values of the natural parameter θ.

   (b) sample size at study termination as a function of various specified values of the natural parameter θ.

   (c) A description of the statistical inference possible (P values, point estimates, confidence intervals) at each of the stopping boundaries at each of the analyses (see section 13 below).

   (d) Examination of the stopping boundaries on other scales, such as the futility or Bayesian scales.

Through the use of exact constraints you may enter arbitrary stopping rules. When using the sample mean, partial sum, or normalized Z statistic design families, if the exact constraint matrix is fully specified, all group sequential design parameters are ignored except the alpha and beta parameters. The values of the alpha and beta parameters will be used to find the hypotheses rejected by each boundary.

When an exact constraint matrix is fully specified on the error spending scale, a group sequential design having the specified error spending functions is obtained. In this way, arbitrary error spending functions can be used for group sequential test design.

The search for boundaries using error spending functions is effected as described in the appendix of [7].

12 Flexible Implementation of Stopping Rules Based on Constrained Boundaries

The stopping rule chosen in the design of a clinical trial serves as a guideline to a Data Monitoring Committee as it makes the decision to recommend continuing or stopping a clinical trial. If all aspects of the conduct of the clinical trial adhered exactly to the conditions stipulated during the design, the stopping rule obtained during the design phase could be used directly. However there are usually at least two complicating factors that must be dealt with when during the conduct of the clinical trial.

First, the schedule of interim analyses does not follow that used in the design of the trial. Often, meetings of the Data Monitoring Committee are scheduled according to calendar time, and thus the sample sizes available for analysis at any given meeting is a random variable. Similarly, accrual may be slower or faster than planned, thereby resulting in a different number of interim analyses than was originally planned. Either of these eventualities will necessitate modifications of the stopping rule, because the exact stopping boundaries are dependent upon the number and timing of analyses. For instance, an [8] design appropriate for four equally spaced analyses has different stopping thresholds than an [8] design appropriate for four analyses scheduled after 50%, 70%, 85%, and 100% of the data have accrued.

Second, the estimate for response variability that was used at the design phase was incorrect. Often very crude estimates of response variability or baseline event rates are used at the design phase. As the trial progresses, more accurate estimates are to be used. Clearly the operating characteristics of particular stopping rules are heavily dependent on the variability of response measurement.

In order to address these issues, flexible methods of implementing stopping rules have been developed which allow the clinical trialist to maintain at least some of the operating characteristics of the stopping rule. Typically such flexible methods always maintain the size (type I error) at the prescribed level. A choice must then be made as to whether the maximal sample size or the power to detect the design alternative should be maintained.

The flexible methods of implementing stopping rules followed here are based on the idea of computing a stopping boundary for the current interim analysis in such a way that the desired operating characteristics are satisfied and that the stopping rule is constrained to agree with the stopping boundaries used at all previously conducted interim analyses. Thus the flexible monitoring methods are based on the concept of the constrained stopping boundaries described in section 11.

In this approach, a general parameterization of a stopping rule is defined at the design stage by choosing

1. desired operating characteristics α_ℓ, α_u, β_ℓ, and β_u;
2. hypothesis shift parameters ϵ_ℓ and ϵ_u;
3. a boundary scale for the group sequential test statistic;
4. boundary shape parameters P, R, and A for each of the four stopping boundaries; and
5. any desired exact, minimum, or maximum boundary constraints (specified according to the planned schedule of interim analyses).

At the design stage a method of implementing that stopping rule is also specified by choosing

1. a boundary scale for constraining the boundaries at previously conducted analyses; and
2. whether the maximal sample size or the power to detect the design alternative will be maintained (it is possible to decide to set an absolute limit on the maximal sample size, but to maintain statistical power otherwise).

If the error spending scale is chosen as the scale for constraints with a design originally chosen on the sample mean scale, the error spending function for the parametric design using the planned schedule of interim analyses is used as an exact constraint of a design on the error spending scale. That is, the stopping rule defined at the design stage is converted to a stopping rule specified by a fully constrained error spending function.

The monitoring of the trial then proceeds as follows:

1. At the first analysis, the stopping boundaries are derived by using the parametric family (possibly constrained) specified in the design. The exact stopping boundary is computed by considering the proportion Π_1 of statistical information available at that first analysis. The value of Π_1 depends on which operating characteristics of the stopping rule are maintained during the monitoring process:

   (a) If the maximal sample size N is to be maintained, $\Pi_1 = N_1/N$.

   (b) If the power of the test to detect the design alternative is to be maintained, an estimated schedule of future analyses is used to compute $\Pi_2, \ldots, \Pi_J$, and then a stopping rule using the design parametric family (possibly constrained) is found which has the desired power. This consists of searching for the value of N which has the correct type I error and power to detect the alternative for the parametric design family for the estimated schedule of interim analyses.

   In either case, interpolation of the exact, minimum, or maximum constraints specified at the design stage is used to derive any constraints for the interim analyses specified by the estimated schedule of future analyses (which may differ from the schedule specified at the design stage). The current best estimate of the statistical information contributed by a single sampling unit (based on the best estimate of σ^2) is used instead of the estimate supplied at the design stage.

2. At later interim analyses, the exact stopping boundaries used at previously conducted interim analyses are used as exact constraints at those analysis times, and the stopping boundaries at the current analysis and all future analyses specified by an estimated schedule of future analyses are computed using the parametric family of designs specified at the design stage. The basic approach is that described for the first analysis, in which the proportion of statistical information at the jth analysis is computed based either on the planned maximal sample size N if that operating characteristic is to be maintained, or it is computed based on a recomputation of a sample size which takes into account the new schedule of interim analyses and the current best estimate of the statistical information contributed by a single sampling unit. In either case, $\Pi_j = N_j/N$ is used as the proportion of statistical information available at the jth analysis (see comments below on the difference between this approach and that used by PEST and EaSt).

It should be noted that due to the re-estimation of σ^2 at each analysis, the stopping boundaries at previously conducted interim analyses depend upon which boundary scale is used when constraining the stopping rules at those analyses. That is, if the value of σ^2 used in computing the stopping rule is constant over the course of the study, it is irrelevant which boundary scale is used for the constraints at previously conducted analyses. If, as is usually the case, the estimate of that statistical information varies over the study, there will be some difference between the boundaries obtained. There is no clear advantage for one such scale over another.

This approach based on constrained boundaries is a generalization of the error spending approach of Lan & DeMets (1983) and [10]: That approach corresponds to boundary constraints specified on the error spending scale. It should be noted that if the maximal sample size is not constrained, the error spending function specified at the design stage is only approximately obtained.

It should be noted that the approach specified here differs somewhat from the methods implemented by PEST and the information based monitoring implemented by EaSt. In those programs, the statistical information at previously conducted interim analyses is not recomputed to reflect updated estimates of the value of σ^2. That is, at the jth analysis, an estimate $\hat{\sigma}_j^2$ was available, and the statistical information available at the jth analysis was estimated as $N_j/\hat{\sigma}_j^2$. PEST and East then estimate the proportion of statistical information available at previously conducted analyses using the estimate of statistical information that was available at that analysis. Using this kind of approach, if at the first analysis the estimated statistical information was estimated as $N_1/\hat{\sigma}_1^2$, and at the current jth analysis the estimated maximal statistical information is $N/\hat{\sigma}_j^2$, the value of Π_1 might be taken to be $[N_1/\hat{\sigma}_1^2]/[N/\hat{\sigma}_j^2]$ or the same proportion as was estimated at the first analysis. Again, this is just another way of approximating the true schedule of interim analyses, and it is not immediately clear that one method is uniformly better than another. The approach taken here is in effect trying to correct for poor estimates of σ^2 that might have been used at the earliest analyses, and thus perhaps better approximate the true sampling distribution. It is still just an approximation to the sampling distribution, however.

13 Estimation Following a Group Sequential Test

In a fixed sample study, where the final sample size is fixed in advance of collecting any data, the most attractive estimator of the mean of a normal distribution is the sample mean. We commonly compute P values and confidence intervals based on the distribution of this estimator. However, the use of a group sequential stopping rule generally alters the sampling distribution of the usual fixed sample statistics, thus special techniques must be used to compute point estimates, interval estimates and P values.

[4] and [1] discuss the various estimators that can be used in the group sequential setting. Suppose we have observed test statistic $(M, S) = (m, s)$ from a group sequential test. Estimates that we will be interested in include

1. P values:
$$P = Pr\left[(M, S) > (m, s); \delta = 0\right] \tag{13.1}$$

2. Point estimates:

   (a) Maximum likelihood estimate (MLE): The MLE in the group sequential setting is merely the sample mean. However, following the use of a group sequential stopping rule, the MLE is now biased, and its distribution is not normal. The MLE $\hat{\delta}$ is computed according to
   $$\hat{\delta} = \frac{s}{N_m} \tag{13.2}$$

   (b) Median unbiased estimate (MUE): The median unbiased estimate is that value $\tilde{\delta}$ such that
   $$Pr\left[(M, S) > (m, s); \delta = \tilde{\delta}\right] = 0.5 \tag{13.3}$$

   (c) Bias adjusted mean (BAM) [14]: The BAM is that value $\check{\delta}$ such that
   $$E\left[\frac{S}{N_M}; \delta = \check{\delta}\right] = \frac{s}{N_m} \tag{13.4}$$

   (d) Rao-Blackwell adjusted unbiased estimate (RBUE): This estimator is computed using the Rao-Blackwell improvement theorem. Within certain classes, this can be shown to be a uniform minimum variance unbiased estimator (Liu & Hall, 199?), and hence this estimator has been referred to as the UMVUE. The estimator $\ddot{\delta}$ is found as
   $$\ddot{\delta} = E\left[\frac{S_1}{N_1} \mid (M, S) = (m, s)\right] \tag{13.5}$$

3. Confidence intervals: A $100(1-2\alpha)\%$ confidence interval is $(\hat{\delta}_L, \hat{\delta}_U)$, where the endpoints are defined such that
$$\begin{aligned} Pr\left[(M, S) > (m, s); \hat{\delta}_L\right] &= \alpha \\ Pr\left[(M, S) > (m, s); \hat{\delta}_U\right] &= 1 - \alpha \end{aligned} \tag{13.6}$$

In order to compute the P value, MUE, and confidence intervals, we must define an ordering of the sample space for the bivariate statistic (M, S). There is no uniformly optimal ordering, but several orderings have been proposed. Two orderings that each have some advantages are

1. Ordering by analysis time [12]: This ordering is not defined for group sequential designs having $a_{Sk} < b_{Sk} < c_{Sk} < d_{Sk}$ at any of the analyses. That is, it is only defined for designs in which the continuation sets can always be written as a single interval. In such designs, the ordering is defined using the boundaries of the continuation sets. Under this ordering

$$(m_1, s_1) < (m_2, s_2) \quad \text{iff} \quad \begin{cases} m_1 < m_2 \text{ and } s_1 < x, \forall x \in \mathcal{C}_{Sm_1} \\ m_1 = m_2 \text{ and } s_1 < s_2 \\ m_1 > m_2 \text{ and } s_2 > x, \forall x \in \mathcal{C}_{Sm_2} \end{cases} \tag{13.7}$$

2. Ordering by the sample mean [4]: This ordering is defined for all group sequential designs, and for a wide variety of group sequential designs was found to average shorter confidence intervals than the ordering based on the analysis time. Under this ordering

$$(m_1, s_1) < (m_2, s_2) \quad \text{iff} \quad \frac{s_1}{N_{m_1}} < \frac{s_2}{N_{m_2}} \tag{13.8}$$

References

[1] Emerson SS (1993). Computation of the uniform minimum variance unbiased estimator of a normal mean following a group sequential trial. *Computers in Biomedical Research* **26**, 68-73.

[2] Emerson SS, Bruce A, Baldwin K (2000). *S+SeqTrial User's Manual* Seattle: Mathsoft, Inc.

[3] Emerson SS, Fleming TR (1989). Symmetric group sequential designs. *Biometrics* **45**, 905-23.

[4] Emerson SS, Fleming TR (1990). Parameter estimation following group sequential hypothesis testing. *Biometrika* **77**, 875-92.

[5] Emerson SS, Kittelson JM (1997). A computationally simpler algorithm for the UMVUE of a normal mean following a group sequential test. *Biometrics* **53**, 365-69.

[6] Jennison C, Turnbull B (1999). *Group Sequential Methods with Applications to Clinical Trials*, Boca Raton: Chapman and Hall/CRC.

[7] Kittelson JM, Emerson SS (1999). A unifying family of group sequential test designs. *Biometrics* **55**, 874-82.

[8] O'Brien PC, Fleming TR (1979). A multiple testing procedure for clinical trials. *Biometrics* **35**, 549-56.

[9] Pampallona SK, Tsiatis AA (1994). Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *Journal of Statistical Planning and Inference* **42**, 19-35.

[10] Pampallona SK, Tsiatis AA, Kim K. (1995). Spending functions for type I and type II error probabilities of group sequential tests. Technical Report, Dept. of Biostatistics, Harvard School of Public Health, Boston.

[11] Pocock SJ (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191-9.

[12] Tsiatis AA, Rosner GL, Mehta CR (1984). Exact confidence intervals following a group sequential test. *Biometrics* **40**, 797-803.

[13] Wang SK, Tsiatis AA (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* **43**, 193-200.

[14] Whitehead J (1986). On the bias of maximum likelihood estimation following a sequential clinical trial. *Biometrika* **73**, 573-81.

[15] Whitehead J (1997). *The Design and Analysis of Sequential Clinical Trials*, Revised 2nd ed. Chichester: Wiley.

[16] Whitehead J, Stratton I (1983). Group sequential clinical trials with triangular continuation regions. *Biometrics* **39**, 227-236.

[17] Xiong X (1995). A class of sequential conditional probability ratio tests. *Journal of the American Statistical Association* **90**, 1463-73.