Determining the critical values

In this document we discuss the calculation of critical values for group sequential designs (GSDs) with small sample sizes based on error spending. The direct calculation of the critical values for the t-test and for the one-way ANOVA is only possible for the very first analysis. For all later analyses the critical values need to be approximated through other techniques. For both tests one can use simulations of the distribution quantiles. For the t-test we also discuss a numerical and an analytical approximation technique.

To begin we need to introduce some notation. We use α and β to denote the probabilities of a type I and type II error respectively and K is the maximum number of planned analyses. At each analysis $i = 0,\ldots,K$, the test statistic T_i is calculated using all data available at that time, then T_i is compared with the significance and futility bounds, u_i and l_i respectively. If T_i is larger than the significance bound u_i, then the result is significant and the experiment is stopped. Similarly, if $T_i < l_i$, then the result is insufficiently promising and the experiment is terminated. If $l_i < T_i < u_i$ the results are inconclusive and we continue to the next analysis. For the last analysis $l_K = u_K$ thus forcing a decision.

When using error spending, the critical bounds u_i and l_i need to be determined in such a way that for certain values α_i and β_i with $i = 0,\ldots,K$

$$\mathbb{P}_{H0}(T_j > u_j \text{ and } T_k > l_k \text{ for any } j = 0,\ldots,i, \forall k < j) = \alpha_i \qquad (1)$$

$$\mathbb{P}_{HA}(T_j < l_j \text{ and } T_k < u_k \text{ for any } j = 0,\ldots,i, \forall k < j) = \beta_i \qquad (2)$$

Where we use $\mathbb{P}_{H0}$ and $\mathbb{P}_{HA}$ to denote the probabilities under the null and alternative hypothesis respectively. In other words, the total chance of committing a type I or type II error at or before analysis i is equal to α_i respectively β_i. These probabilities must of course lie in [0,1] and for $i = 1,\ldots,K-1$

$$\alpha_i \leq \alpha_{i+1} \qquad \alpha_K = \alpha \qquad \beta_i \leq \beta_{i+1}$$

Ideally, the achieved type II error β would be equal to the target type II error β_K. However, since sample size is not continuous, such a design is usually not possible and the only option is to fix one error type and limit the other. Additional notation we will use is λ_i for the non-centrality parameter under the alternative hypothesis and n_i as the total sample size of all the data collected at analysis i.

Numerical integration

The method described in Rom and McTague[1] is restricted to t-tests with one interim analysis and without futility bounds. This technique can be considered exact since the true value of u_2 can in theory be approximated to a desired accuracy level with absolute certainty. Simulation on the other hand, relies on probabilities of random variables and can only give estimates with desired confidence levels. While it is mathematically not too

Susanne Blotwijk[1,*], Sophie Hernot[2] and Kurt Barbé[1].

[1] Biostatistics and Medical Informatics Research Group, Vrije Universiteit Brussel, Brussels, Belgium. [2]Cellular and Molecular Immunology Research Group, Vrije Universiteit Brussel, Brussels, Belgium.

[*] susanne.blotwijk@vub.be

challenging to extend the formulas to multiple interim analyses, computing them numerically is at current computational capabilities generally not feasible without the use of simulation based techniques.

Rom and McTague use numerical integration to determine the type I error and type II error for a group sequential design with given $\hat{u}_1$ and $\hat{u}_2$. Since in alpha spending the exact u_1 corresponding to α_1 is already known. So u_2 can be approximated by iteratively increasing or decreasing $\hat{u}_2$ until the obtained type I error is sufficiently close to α_2. To obtain a suitable β one can iteratively evaluate different sample sizes or change other features of the design.

In this type of design, a type I error occurs if the t-statistic surpasses the significance bounds at either of the analyses under the null hypothesis, while a type II error happens if the t-statistic is lower than the significance bounds at both analyses under the alternative hypothesis. The probabilities of these events are

$$\begin{aligned}
\alpha &= \mathbb{P}_{H0}(\{T_1 > u_1\} \cup \{T_2 > u_2\}) \\
&= \mathbb{P}_{H0}(T_1 > u_1) + \mathbb{P}_{H0}(T_2 > u_2) - \mathbb{P}_{H0}(\{T_1 > u_1\} \cap \{T_2 > u_2\}) \\
&= \left(1 - t_{n_1-2,0}(u_1)\right) + \left(1 - t_{n_2-2,0}(u_2)\right) - \mathbb{P}_{H0}(\{T_1 > u_1\} \cap \{T_2 > u_2\}) \\
&= \left(1 - t_{n_1-2,0}(u_1)\right) + \left(1 - t_{n_2-2,0}(u_2)\right) - I_{RM}(-u_1, -u_2, n_1, n_2, 0, 0) \\
\beta &= \mathbb{P}_{HA}(\{T_1 < u_1\} \cup \{T_2 < u_2\}) \\
&= I_{RM}(u_1, u_2, n_1, n_2, \lambda_1, \lambda_2)
\end{aligned}$$

Where we use $t_{\nu,\lambda}$ to denote the cumulative distribution function of the non-central t-distribution with ν degrees of freedom and non-centrality parameter λ and I_{RM} is the integral that is approximated numerically in Rom and McTague[1]. What changes when adding the futility bound is that a type I error can only occur if the t-statistic did not drop below the futility bound, so

$$\begin{aligned}
\alpha &= \mathbb{P}_{H0}\left(\{T_1 > u_1\} \cup \left\{\{T_2 > u_2\} \cap \{T_1 > l_1\}\right\}\right) \\
&= \mathbb{P}_{H0}(T_1 > u_1) + \mathbb{P}_{H0}(\{T_2 > u_2\} \cap \{T_1 > l_1\}) - \mathbb{P}_{H0}(\{T_1 > u_1\} \cap \{T_2 > u_2\}) \\
&= \left(1 - t_{n_1-2,0}(u_1)\right) + I_{RM}(-l_1, -u_2, n_1, n_2, 0, 0) - I_{RM}(-u_1, -u_2, n_1, n_2, 0, 0)
\end{aligned}$$

where the second equation follows from $\{T_1 > u_1\} \subset \{T_1 > l_1\}$ since $u_1 > l_1$.

Similarly, we can obtain

$$\begin{aligned}
\beta &= \mathbb{P}_{HA}\left(\{T_1 < u_1\} \cup \left\{\{T_2 < l_2\} \cap \{T_1 < u_1\}\right\}\right) \\
&= \mathbb{P}_{HA}(T_1 < l_1) + \mathbb{P}_{H0}(\{T_2 < l_2\} \cap \{T_1 < u_1\}) - \mathbb{P}_{H0}(\{T_1 < l_1\} \cap \{T_2 < l_2\}) \\
&= t_{n_1-2,\lambda_1}(l_1) + I_{RM}(u_1, l_2, n_1, n_2, \lambda_1, \lambda_2) - I_{RM}(l_1, l_2, n_1, n_2, \lambda_1, \lambda_2)
\end{aligned}$$

Therefore the same numerical integration techniques can be used to obtain exact values for designs that include futility bounds.

Analytical approximation

Analytical approximation is the fastest of the three approximation techniques, but the drawback is that the accuracy of the approximation is fixed and cannot be improved through added calculations or simulations. The analytical approximation commonly used in clinical trials and used in the original paper by Lan and Demets[2] is an asymptotic approximation. It is assume the variance is known, rather than a variable that needs to be estimated, and therefore the test statistics have a normal distribution. For this normal distribution the exact bounds u_i^N and l_i^N can be calculated and then used as analytical approximations for u_i and l_i. The sample size required to achieve the desired power can also be calculated and is usually expressed as the product of the required sample size for the fixed sample Z-test multiplied with an inflation factor.

As Nikolakopoulos et al[3] have shown this approximation is poor at low sample sizes, so they proposed to use a transformation of the normal bounds instead, namely

$$\hat{u}_i = t_{n_i-2,0}^{-1}\left(\Phi(u_i^N)\right) \tag{3}$$

$$\hat{l}_i = t_{n_i-2,0}^{-1}\left(\Phi(l_i^N)\right) \tag{4}$$

Here Φ and $t_{\nu,\lambda}$ are the cumulative distribution function respectively of the standard normal distribution and of the non-central t-distribution with ν degrees of freedom and non-centrality parameter λ. This correction consists of transforming the critical values into their corresponding p-values under the normal distribution and then transforming those p-values into the corresponding test statistics under the t-distribution. Because of the definition of a p-value, this is the exact analytical answer for the critical value controlling type I error at the first interim analysis. For the later analyses however, the dependence of the test statistics on the former analyses comes into play. As a result, this correction becomes less accurate with more interim analyses.

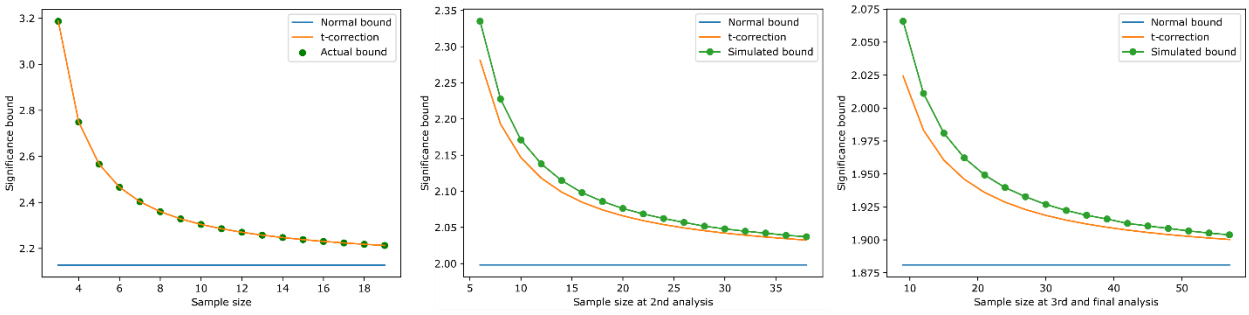

Figure 1 Example of significance bounds and their analytical approximation In this example, the design consists of three analyses with equal sample sizes and alpha spending function αt. It features the significance bounds of the normal approximation and the one proposed by Nikolakopoulos et al.[3]

This approximation works well for the significance bounds if only alpha spending is used, as is illustrated in Figure 1. However, it does not show any improvement for the futility bounds. The lower bounds are controlling the type II error under the alternative hypothesis. Since equation 4 uses p-values based on the null hypothesis,

applying this transformation on the lower bounds makes little sense and indeed can even make the approximation worse. This can be improved by basing the transformation based on the distribution of the alternative hypothesis rather than that of the null hypothesis, specifically

$$\hat{l}_i^* = t_{n_i-2,\lambda_i}^{-1}\left(\Phi(l_i^N)\right)$$

Since Nikolakopoulos et al wrote their work in the context of small sample clinical trials, they largely assumed the maximum sample size was fixed. They therefore focussed on the loss of power caused by various designs and did not propose an alternative for adapting the required sample size. This can be done by multiplying the inflation factor of the normal approximation with the required sample size for the fixed sample t-test. Without this increased sample size, the t-transformed analytical approximation is underpowered, as is illustrated in Figure 2. Since the futility bounds also influence the probabilities under the null hypothesis, our improved approximation of the futility bounds also results in a better approximation for the type I error.

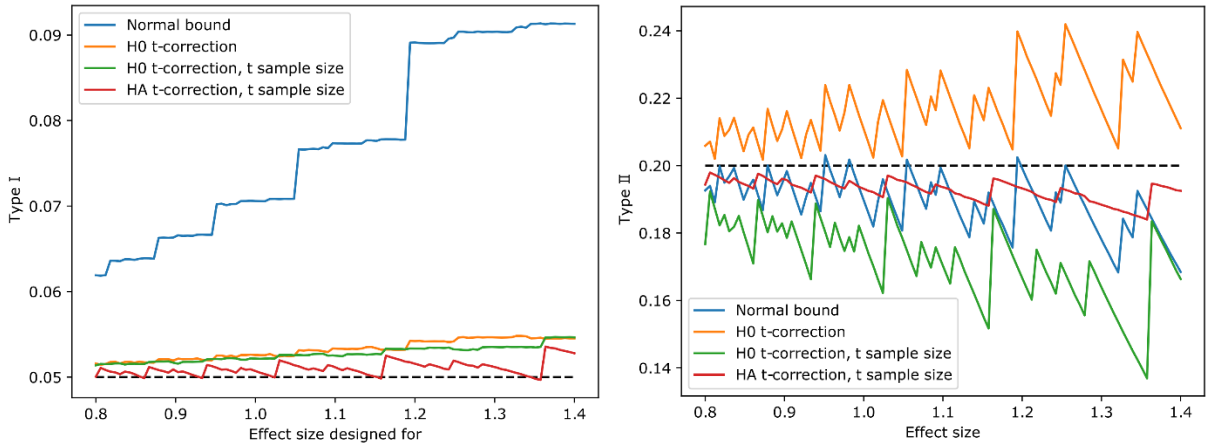

Figure 2 Example of type I and type II errors for analytical approximations In this example, the design consists of three analyses with approximately equal sample sizes, alpha spending function αt and beta spending function βt. The jumps in the graphs are due to the changes in sample sizes.

Simulation

Simulation is currently the only method available to evaluate any possible error spending design at arbitrary accuracy. Previous articles[3,4] that cover simulation, only do this for the Pocock bounds or to evaluate the power of a design. Additionally, they do not discuss how to reach desired levels of accuracy, nor do they discuss how to evaluate designs efficiently. Instead they pick a fixed, large, arbitrary number of simulations to evaluate all designs.

The basic concept to simulate all critical values is straightforward and consists of 4 steps. Generate n_k random normal variables once with means and standard deviations of the null hypothesis and once under the alternative hypothesis. Use these to calculate the corresponding test statistics at each interim analysis. Repeat m times. Determine the quantiles $\hat{u}_i$ and $\hat{l}_i$ of the simulated test statistics that satisfy 1 and 2. We can then use $\hat{u}_i$ and $\hat{l}_i$ to determine for each of the M simulations under $H0$ and HA at which analysis they stopped, what the used

sample size was and whether or not a significant result was obtained, hence allowing us to estimate the expected sample size and the power of the design.

The reliability of the estimates depends on the size of M. One option is to just choose a large but arbitrary number. Of course this gives no guarantee that the desired accuracy was achieved, nor is it very efficient, especially since in many cases researchers may wish to evaluate multiple designs for several scenarios.

The asymptotic distributions of the estimators (proof below) are all normal with variance proportional to $1/m$. Hence taking the average of r simulations with size m has the same asymptotic distribution as one simulation with size rm. Splitting the simulations has several practical advantages. Most importantly, it gives an easy and inexpensive way to calculate the confidence intervals for each estimator and to estimate the number of simulations still required to reach the desired interval lengths. Additionally, it limits the memory usage, which is useful should you want to evaluate quite a few designs in the background while using your computer for other things. Lastly, it makes it very easy to apply parallel computing, although that really is not necessary unless one wishes extremely accurate results or if one wishes to evaluate a large number of designs, e.g. when searching for the optimal alpha spending design.

The asymptotic distributions for the estimators are

$$\sqrt{m}(\hat{u}_i - u_i) \to U_i \sim \mathcal{N}\left(0, \frac{(\alpha_i - \alpha_{i-1})\gamma_i^{-1}(1-(\alpha_i-\alpha_{i-1})\gamma_i^{-1})}{f_{T_i,H_0}(u_i)^2}\right) \tag{5}$$

$$\sqrt{m}(\hat{l}_i - l_i) \to L_i \sim \mathcal{N}\left(0, \frac{(\beta_i - \beta)\zeta_i^{-1}(1-(\beta_i-\beta_{i-1})\zeta_i^{-1})}{f_{T_i,A}(l_i)^2}\right) \tag{6}$$

$$\sqrt{m}(\hat{p} - p) \to P \sim \mathcal{N}\big(0, p(1-p)\big) \tag{7}$$

Where $i = 2, \dots, K$, $\mathcal{N}(\mu, \sigma^2)$ is a normal distribution with mean μ and variance σ^2, $f_{T_i,H0}$ and $f_{T_i,HA}$ are the density functions of the test statistic T_i under the null and alternative hypothesis respectively, and lastly

$$\gamma_i = \mathbb{P}_{H_0}\left(\bigcap_{j=1}^{i-1}\{l_j < T_j < u_j\}\right) \quad \text{and} \quad \zeta_i = \mathbb{P}_{H_A}\left(\bigcap_{j=1}^{i-1}\{l_j < T_j < u_j\}\right)$$

Proof: Statements (5) and (6) can be shown through induction. From the basic properties of convergence of random variables it can easily be derived that for any constant vector $\mathbf{c} \in \mathbb{R}^d$ and for any sequences of random variables $(X_n)_{n\in\mathbb{N}}$ and $(Y_n)_{n\in\mathbb{N}}$:

$$\begin{cases} (X_n, \mathbf{c}) \xrightarrow{d} (X, \mathbf{c}) \\ Y_n \xrightarrow{p} \mathbf{c} \end{cases} \Rightarrow (X_n, Y_n) \xrightarrow{d} (X, \mathbf{c}) \tag{8}$$

Where $\xrightarrow{d}$ and $\xrightarrow{p}$ denote convergence in distribution and in probability respectively.

$\boxed{i = 2}$ Define $M_{i,H0}$ as the sample size at the i^{th} analysis under the null hypothesis. Since u_1 and l_1 are known, $M_{2,H0}$ follows a binomial distribution with m trials and probability γ_2. As a result $M_{2,H_0}/m$ converges in probability to

γ_2 for $m \to \infty$. By construction $\hat{u}_i$ is the $((\alpha_i - \alpha_{i-1})m/M_{i,H0})$ −th quantile of the simulation at analysis *i*. Since the asymptotic distribution of sample quantiles is known [5], it follows that $\left(\sqrt{m}(\hat{u}_2 - u_2), \gamma_2\right) \xrightarrow{d} (U_2, \gamma_2)$. By applying (8) we obtain $\left(\sqrt{m}(\hat{u}_2 - u_2), M_{2,H0}/m\right) \xrightarrow{d} (U_2, \gamma_2)$

The derivation for $\left(\sqrt{m}(\hat{l}_2 - l_2), M_{2,HA}/m\right) \xrightarrow{d} (L_2, \zeta_2)$) is analogous. Through the The derivation for (m(continuous mapping theorem, we can conclude $\hat{u}_2 \xrightarrow{p} u_2$ and $\hat{l}_2 \xrightarrow{p} l_2$.

$i > 2$ Assume that $(\hat{u}_2, \dots, \hat{u}_{i-1}, \hat{l}_2, \dots, \hat{l}_{i-1}) \xrightarrow{p} (u_2, \dots, u_{i-1}, l_2, \dots, l_{i-1})$. From the definition of γ_i it follows that if $\hat{u}_j = u_j$ and $\hat{l}_j = l_j$ for $1 < j \leq i - 1$, then $M_{i,H0}$ has a binomial distribution with m trials and probability γ_i. Through (8) we get $\left(M_{i,H0}/m, \hat{u}_2, \dots, \hat{u}_{i-1}, \hat{l}_2, \dots, \hat{l}_{i-1}\right) \xrightarrow{d} (\gamma_i, u_2, \dots, u_{i-1}, l_2, \dots, l_{i-1})$ and since γ_i is a constant

$$\left(\frac{M_{i,H0}}{m}, \hat{u}_2, \dots, \hat{u}_{i-1}, \hat{l}_2, \dots, \hat{l}_{i-1}\right) \xrightarrow{p} (\gamma_i, u_2, \dots, u_{i-1}, l_2, \dots, l_{i-1})$$

Similarly as for the case $i = 2$, from the asymptotic distribution of quantiles and from (8) it follows that

$$\left(\sqrt{m}(\hat{u}_i - u_i), \frac{M_{i,H0}}{m}, \hat{u}_2, \dots, \hat{u}_{i-1}, \hat{l}_2, \dots, \hat{l}_{i-1}\right) \xrightarrow{d} (U_i, \gamma_i, u_2, \dots, u_{i-1}, l_2, \dots, l_{i-1})$$

and thus

$$(\hat{u}_2, \dots, \hat{u}_i, \hat{l}_2, \dots, \hat{l}_{i-1}) \xrightarrow{p} (u_2, \dots, u_i, l_2, \dots, l_{i-1})$$

The proof for $\sqrt{m}(\hat{l}_i - l_i) \xrightarrow{d} L_i$ and $\hat{l}_i \xrightarrow{p} l_i$ is analogous. Since $(u_2, \dots, u_{i-1}, \hat{u}_i, l_2, \dots, l_{i-1})$ and $(u_2, \dots, u_{i-1}, l_2, \dots, l_{i-1}, \hat{l}_i)$ are independent, it follows that

$$(\hat{u}_2, \dots, \hat{u}_i, \hat{l}_2, \dots, \hat{l}_i) \xrightarrow{p} (u_2, \dots, u_i, l_2, \dots, l_i) \qquad (9)$$

This concludes the proof for (5) and (6).

Given $(u_1, \dots, u_K, l_1, \dots, l_{K-1})$ the odds of obtaining a significant result under the alternative hypothesis is by definition the power of the design p and therefore $\hat{p}$ has a binomial distribution with m trials and probability p. Through (8), (9) and the central limit theorem, we can conclude that (7) is true.

For any fixed set of bounds, the estimators for the expected value of the sample size and of the analysis i at which the experiment stops, are sample averages of i.i.d. variables with finite variance. As for $\hat{p}$ we can therefore apply (8), (9) and the central limit theorem to conclude that each of these estimators is asymptotically normal with variance proportional to $1/m$.

References

1. Rom, D. M. & McTague, J. A. Exact critical values for group sequential designs with small sample sizes. *J. Biopharm. Stat.* **30**, 752–764 (2020).

2. Lan, K. K. G. & Demets, D. L. Discrete sequential boundaries for clinical trials. *Biometrika* **70**, 659–663 (1983).

3. Nikolakopoulos, S., Roes, K. C. & van der Tweel, I. Sequential designs with small samples: Evaluation and recommendations for normal responses. *Stat. Methods Med. Res.* **27**, 1115–1127 (2018).

4. Shao, J. & Feng, H. Group sequential t-test for clinical trials with small sample sizes across stages. *Contemp. Clin. Trials* **28**, 563–571 (2007).

5. Kendall, M. G., Stuart, A., Ord, J. K., Arnold, S. F. & O'Hagan, A. *Kendall's advanced theory of statistics*. (Edward Arnold ; Halsted Press, 1994).