# Group sequential monitoring of clinical trials

## The basics

Constantin T. Yiannoutsos, Ph.D.
Indiana University Fairbanks School of Public Health
cyiannou@iu.edu

4 July, 2022

# Outline of the course

- Unit 1: The basics

# Outline of the course

- Unit 1: The basics
  - Introduction

# Outline of the course

- Unit 1: The basics
  - Introduction
    - What we will and what we will not cover

# Outline of the course

- Unit 1: The basics
  - Introduction
    - What we will and what we will not cover
    - Why monitor

# Outline of the course

- Unit 1: The basics
  - Introduction
    - What we will and what we will not cover
    - Why monitor
  - Lest we forget

# Outline of the course

- Unit 1: The basics
  - Introduction
    - What we will and what we will not cover
    - Why monitor
  - Lest we forget
    - Controlling the alpha level

# Outline of the course

- Unit 1: The basics
  - Introduction
    - What we will and what we will not cover
    - Why monitor
  - Lest we forget
    - Controlling the alpha level
  - Monitoring of clinical trials

# Outline of the course

- Unit 1: The basics
  - Introduction
    - What we will and what we will not cover
    - Why monitor
  - Lest we forget
    - Controlling the alpha level
  - Monitoring of clinical trials
    - First approaches of trial monitoring

# Outline of the course

- Unit 1: The basics
  - Introduction
    - What we will and what we will not cover
    - Why monitor
  - Lest we forget
    - Controlling the alpha level
  - Monitoring of clinical trials
    - First approaches of trial monitoring
  - Monitoring 2G

# Outline of the course

- Unit 1: The basics
  - Introduction
    - What we will and what we will not cover
    - Why monitor
  - Lest we forget
    - Controlling the alpha level
  - Monitoring of clinical trials
    - First approaches of trial monitoring
  - Monitoring 2G
    - The game changing idea of the alpha spending functions

# Outline of the course

- Unit 1: The basics
  - Introduction
    - What we will and what we will not cover
    - Why monitor
  - Lest we forget
    - Controlling the alpha level
  - Monitoring of clinical trials
    - First approaches of trial monitoring
  - Monitoring 2G
    - The game changing idea of the alpha spending functions
  - The impact of monitoring on sample size

# Outline of the course

- Unit 1: The basics
  - Introduction
    - What we will and what we will not cover
    - Why monitor
  - Lest we forget
    - Controlling the alpha level
  - Monitoring of clinical trials
    - First approaches of trial monitoring
  - Monitoring 2G
    - The game changing idea of the alpha spending functions
  - The impact of monitoring on sample size
- Unit 2: Advanced topics

# Outline of the course

- Unit 1: The basics
  - Introduction
    - What we will and what we will not cover
    - Why monitor
  - Lest we forget
    - Controlling the alpha level
  - Monitoring of clinical trials
    - First approaches of trial monitoring
  - Monitoring 2G
    - The game changing idea of the alpha spending functions
  - The impact of monitoring on sample size
- Unit 2: Advanced topics
  - Sample size vs. information-based monitoring

# Outline of the course

- Unit 1: The basics
  - Introduction
    - What we will and what we will not cover
    - Why monitor
  - Lest we forget
    - Controlling the alpha level
  - Monitoring of clinical trials
    - First approaches of trial monitoring
  - Monitoring 2G
    - The game changing idea of the alpha spending functions
  - The impact of monitoring on sample size
- Unit 2: Advanced topics
  - Sample size vs. information-based monitoring
  - Monitoring studies of time to event

# Outline of the course

- Unit 1: The basics
  - Introduction
    - What we will and what we will not cover
    - Why monitor
  - Lest we forget
    - Controlling the alpha level
  - Monitoring of clinical trials
    - First approaches of trial monitoring
  - Monitoring 2G
    - The game changing idea of the alpha spending functions
  - The impact of monitoring on sample size
- Unit 2: Advanced topics
  - Sample size vs. information-based monitoring
  - Monitoring studies of time to event
    - Monitoring based on calendar time

# Outline of the course

- Unit 1: The basics
  - Introduction
    - What we will and what we will not cover
    - Why monitor
  - Lest we forget
    - Controlling the alpha level
  - Monitoring of clinical trials
    - First approaches of trial monitoring
  - Monitoring 2G
    - The game changing idea of the alpha spending functions
  - The impact of monitoring on sample size
- Unit 2: Advanced topics
  - Sample size vs. information-based monitoring
  - Monitoring studies of time to event
    - Monitoring based on calendar time
    - Adjustments when event rates are lower than expected

# Outline of the course

- Unit 1: The basics
  - Introduction
    - What we will and what we will not cover
    - Why monitor
  - Lest we forget
    - Controlling the alpha level
  - Monitoring of clinical trials
    - First approaches of trial monitoring
  - Monitoring 2G
    - The game changing idea of the alpha spending functions
  - The impact of monitoring on sample size
- Unit 2: Advanced topics
  - Sample size vs. information-based monitoring
  - Monitoring studies of time to event
    - Monitoring based on calendar time
    - Adjustments when event rates are lower than expected
  - Beta spending and futility

# Outline of the course

- Unit 1: The basics
  - Introduction
    - What we will and what we will not cover
    - Why monitor
  - Lest we forget
    - Controlling the alpha level
  - Monitoring of clinical trials
    - First approaches of trial monitoring
  - Monitoring 2G
    - The game changing idea of the alpha spending functions
  - The impact of monitoring on sample size
- Unit 2: Advanced topics
  - Sample size vs. information-based monitoring
  - Monitoring studies of time to event
    - Monitoring based on calendar time
    - Adjustments when event rates are lower than expected
  - Beta spending and futility
- Unit 3: Case studies

# Outline of the course

- Unit 1: The basics
  - Introduction
    - What we will and what we will not cover
    - Why monitor
  - Lest we forget
    - Controlling the alpha level
  - Monitoring of clinical trials
    - First approaches of trial monitoring
  - Monitoring 2G
    - The game changing idea of the alpha spending functions
  - The impact of monitoring on sample size
- Unit 2: Advanced topics
  - Sample size vs. information-based monitoring
  - Monitoring studies of time to event
    - Monitoring based on calendar time
    - Adjustments when event rates are lower than expected
  - Beta spending and futility
- Unit 3: Case studies
  - The CAST study

# Outline of the course

- Unit 1: The basics
  - Introduction
    - What we will and what we will not cover
    - Why monitor
  - Lest we forget
    - Controlling the alpha level
  - Monitoring of clinical trials
    - First approaches of trial monitoring
  - Monitoring 2G
    - The game changing idea of the alpha spending functions
  - The impact of monitoring on sample size
- Unit 2: Advanced topics
  - Sample size vs. information-based monitoring
  - Monitoring studies of time to event
    - Monitoring based on calendar time
    - Adjustments when event rates are lower than expected
  - Beta spending and futility
- Unit 3: Case studies
  - The CAST study
  - LUN01-24

# Outline of the course

- Unit 1: The basics
  - Introduction
    - What we will and what we will not cover
    - Why monitor
  - Lest we forget
    - Controlling the alpha level
  - Monitoring of clinical trials
    - First approaches of trial monitoring
  - Monitoring 2G
    - The game changing idea of the alpha spending functions
  - The impact of monitoring on sample size
- Unit 2: Advanced topics
  - Sample size vs. information-based monitoring
  - Monitoring studies of time to event
    - Monitoring based on calendar time
    - Adjustments when event rates are lower than expected
  - Beta spending and futility
- Unit 3: Case studies
  - The CAST study
  - LUN01-24
  - The Moderna COVID vaccine study

# Outline of the course

- Unit 1: The basics
  - Introduction
    - What we will and what we will not cover
    - Why monitor
  - Lest we forget
    - Controlling the alpha level
  - Monitoring of clinical trials
    - First approaches of trial monitoring
  - Monitoring 2G
    - The game changing idea of the alpha spending functions
  - The impact of monitoring on sample size
- Unit 2: Advanced topics
  - Sample size vs. information-based monitoring
  - Monitoring studies of time to event
    - Monitoring based on calendar time
    - Adjustments when event rates are lower than expected
  - Beta spending and futility
- Unit 3: Case studies
  - The CAST study
  - LUN01-24
  - The Moderna COVID vaccine study
- **The R Markdown code that generated these slides can be obtained from the https://github.com/cyiannou/Sequential-monitoring/**

# Introduction

# What we will and will not cover

- We will cover:

[1] Although this will be obliquely referred to, particularly in the case studies

# What we will and will not cover

- We will cover:
  - The reasoning behind monitoring of trials

[1] Although this will be obliquely referred to, particularly in the case studies

# What we will and will not cover

- We will cover:
  - The reasoning behind monitoring of trials
  - The technical aspects of clinical trial monitoring

[1] Although this will be obliquely referred to, particularly in the case studies

# What we will and will not cover

- We will cover:
  - The reasoning behind monitoring of trials
  - The technical aspects of clinical trial monitoring
  - Some classic (and emerging classic) case studies

[1] Although this will be obliquely referred to, particularly in the case studies

# What we will and will not cover

- We will cover:
  - The reasoning behind monitoring of trials
  - The technical aspects of clinical trial monitoring
  - Some classic (and emerging classic) case studies
- We will not cover:

[1] Although this will be obliquely referred to, particularly in the case studies

# What we will and will not cover

- We will cover:
  - The reasoning behind monitoring of trials
  - The technical aspects of clinical trial monitoring
  - Some classic (and emerging classic) case studies
- We will not cover:
  - Data Safety Monitoring Boards and the process of review[1]

[1] Although this will be obliquely referred to, particularly in the case studies

# What we will and will not cover

- We will cover:
  - The reasoning behind monitoring of trials
  - The technical aspects of clinical trial monitoring
  - Some classic (and emerging classic) case studies
- We will not cover:
  - Data Safety Monitoring Boards and the process of review[1]
  - Multi-stage designs

[1] Although this will be obliquely referred to, particularly in the case studies

# What we will and will not cover

- We will cover:
  - The reasoning behind monitoring of trials
  - The technical aspects of clinical trial monitoring
  - Some classic (and emerging classic) case studies
- We will not cover:
  - Data Safety Monitoring Boards and the process of review[1]
  - Multi-stage designs
  - Adaptive designs

[1] Although this will be obliquely referred to, particularly in the case studies

# What we will and will not cover

- We will cover:
  - The reasoning behind monitoring of trials
  - The technical aspects of clinical trial monitoring
  - Some classic (and emerging classic) case studies
- We will not cover:
  - Data Safety Monitoring Boards and the process of review[1]
  - Multi-stage designs
  - Adaptive designs
- Our focus is on late-stage (e.g., Phase-III) *randomized* clinical trials with a single (or a single primary) endpoint

[1] Although this will be obliquely referred to, particularly in the case studies

# Why monitor?

- A clinical trial should not continue only by virtue of the fact that it has begun.

# Why monitor?

- A clinical trial should not continue only by virtue of the fact that it has begun.
- Establishing that the ethical considerations that were present at its initiation continue to be present at every point in its implementation is paramount.

# Why monitor?

- A clinical trial should not continue only by virtue of the fact that it has begun.
- Establishing that the ethical considerations that were present at its initiation continue to be present at every point in its implementation is paramount.
- In this lecture we discuss what constitutes appropriate monitoring of a clinical trial.

# Why monitor?

- A clinical trial should not continue only by virtue of the fact that it has begun.
- Establishing that the ethical considerations that were present at its initiation continue to be present at every point in its implementation is paramount.
- In this lecture we discuss what constitutes appropriate monitoring of a clinical trial.
- We focus on randomized comparative trials with a single primary endpoint.

# Reasons for early stopping of a clinical trial

- The following are some *external* reasons to stop a clinical trial early

# Reasons for early stopping of a clinical trial

- The following are some *external* reasons to stop a clinical trial early
  - Treatments are found to be different by experts

# Reasons for early stopping of a clinical trial

- The following are some *external* reasons to stop a clinical trial early
  - Treatments are found to be different by experts
  - Treatments are found to be not different by experts

# Reasons for early stopping of a clinical trial

- The following are some *external* reasons to stop a clinical trial early
  - Treatments are found to be different by experts
  - Treatments are found to be not different by experts
  - Side effects are too severe to continue in light of benefits

# Reasons for early stopping of a clinical trial

- The following are some *external* reasons to stop a clinical trial early
  - Treatments are found to be different by experts
  - Treatments are found to be not different by experts
  - Side effects are too severe to continue in light of benefits
  - Accrual too slow to complete study in a timely fashion

# Reasons for early stopping of a clinical trial

- The following are some *external* reasons to stop a clinical trial early
  - Treatments are found to be different by experts
  - Treatments are found to be not different by experts
  - Side effects are too severe to continue in light of benefits
  - Accrual too slow to complete study in a timely fashion
  - The data are of poor quality

# Reasons for early stopping of a clinical trial

- The following are some *external* reasons to stop a clinical trial early
  - Treatments are found to be different by experts
  - Treatments are found to be not different by experts
  - Side effects are too severe to continue in light of benefits
  - Accrual too slow to complete study in a timely fashion
  - The data are of poor quality
  - Definitive information about the treatment becomes available making the study unnecessary or unethical

# Reasons for early stopping of a clinical trial

- The following are some *external* reasons to stop a clinical trial early
  - Treatments are found to be different by experts
  - Treatments are found to be not different by experts
  - Side effects are too severe to continue in light of benefits
  - Accrual too slow to complete study in a timely fashion
  - The data are of poor quality
  - Definitive information about the treatment becomes available making the study unnecessary or unethical
  - The scientific questions are no longer important

# Reasons for early stopping of a clinical trial

- The following are some *external* reasons to stop a clinical trial early
  - Treatments are found to be different by experts
  - Treatments are found to be not different by experts
  - Side effects are too severe to continue in light of benefits
  - Accrual too slow to complete study in a timely fashion
  - The data are of poor quality
  - Definitive information about the treatment becomes available making the study unnecessary or unethical
  - The scientific questions are no longer important
  - Adherence to treatment is unacceptably poor

# Reasons for early stopping of a clinical trial

- The following are some *external* reasons to stop a clinical trial early
  - Treatments are found to be different by experts
  - Treatments are found to be not different by experts
  - Side effects are too severe to continue in light of benefits
  - Accrual too slow to complete study in a timely fashion
  - The data are of poor quality
  - Definitive information about the treatment becomes available making the study unnecessary or unethical
  - The scientific questions are no longer important
  - Adherence to treatment is unacceptably poor
  - Resources to perform the study are lost or are no longer available

# Reasons for early stopping of a clinical trial

- The following are some *external* reasons to stop a clinical trial early
  - Treatments are found to be different by experts
  - Treatments are found to be not different by experts
  - Side effects are too severe to continue in light of benefits
  - Accrual too slow to complete study in a timely fashion
  - The data are of poor quality
  - Definitive information about the treatment becomes available making the study unnecessary or unethical
  - The scientific questions are no longer important
  - Adherence to treatment is unacceptably poor
  - Resources to perform the study are lost or are no longer available
  - The study integrity has been undermined by fraud or misconduct

# Reasons for early stopping of a clinical trial

- The following are some *external* reasons to stop a clinical trial early
  - Treatments are found to be different by experts
  - Treatments are found to be not different by experts
  - Side effects are too severe to continue in light of benefits
  - Accrual too slow to complete study in a timely fashion
  - The data are of poor quality
  - Definitive information about the treatment becomes available making the study unnecessary or unethical
  - The scientific questions are no longer important
  - Adherence to treatment is unacceptably poor
  - Resources to perform the study are lost or are no longer available
  - The study integrity has been undermined by fraud or misconduct
- In this short course we focus on *internally* generated information which may lead to the interruption of the study

# Lest we forget: A review

# Comparing two means

Suppose that we compare the means of two groups.

In the simplest case, after $2N$ total subjects have been enrolled, we compute the statistic[2]

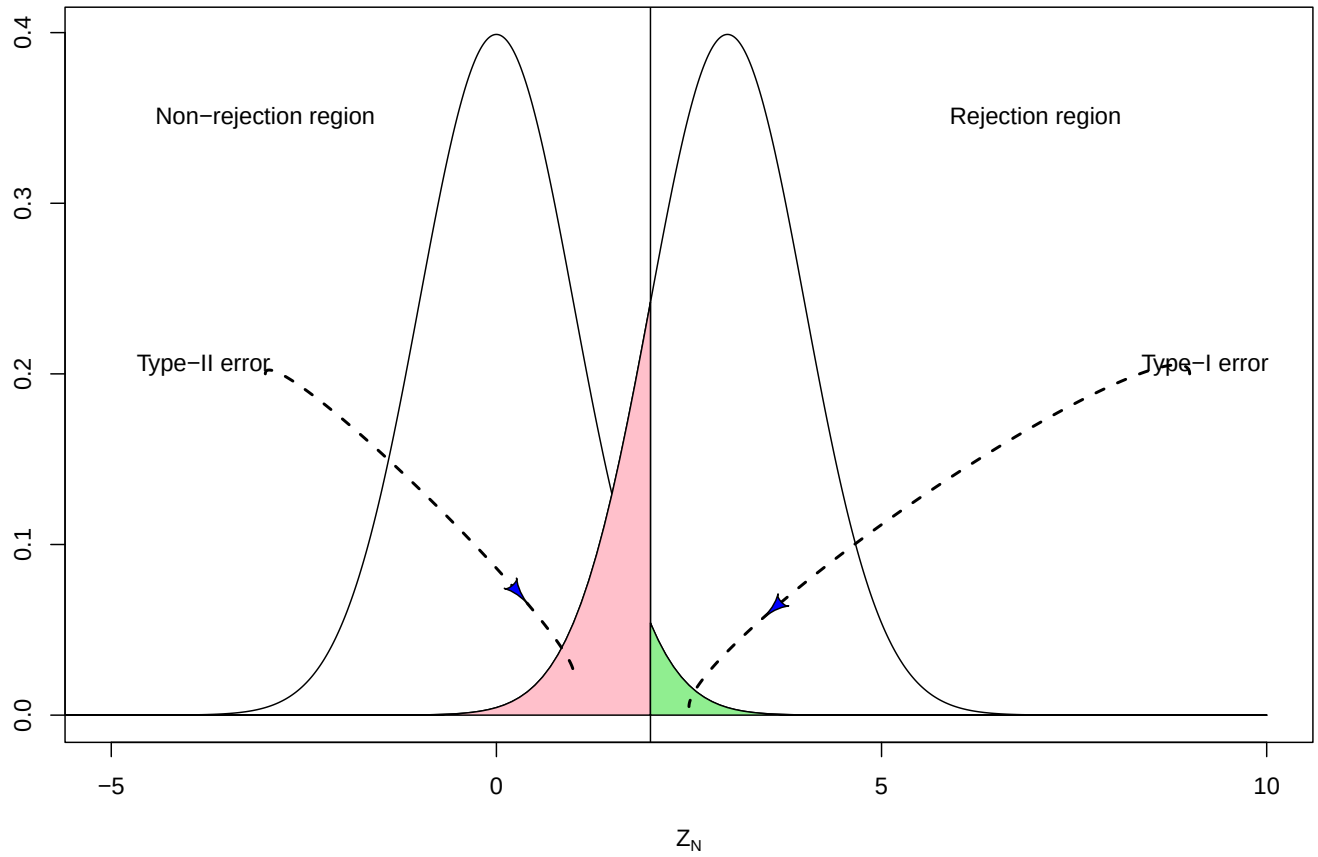
$$Z_N = \frac{\bar{X}_1 - \bar{X}_2}{\left(\frac{2\sigma^2}{N}\right)} = \frac{S_N}{\sqrt{\nu_N}}$$

where $D_i = X_{1i} - X_{2i}$, $S_N = \sum_{i=1}^{N} D_i$ and $\nu_N = \mathrm{var}(S_N) = 2N\sigma^2$.

[2] Here we assume a balanced subject allocation, i.e., $N_1 = N_2 = N$ and a common variance $\mathrm{var}(X_1) = \mathrm{var}(X_2) = \sigma^2$.

# Type-I and Type-II errors

**Figure 1:** Typical situation of a one-sided test comparing the difference of two means. The distribution on the left is consistent to the null hypothesis $H_0 : \mu_2 - \mu_1 \leq 0$, while the one on the right is consistent to the alternative $H_1 : \mu_2 - \mu_1 > 0$



# Controlling the type-I error

The main goal of any design is to control the type-I error by setting an upper limit of the maximum probability of rejecting the null hypothesis when it is true (aka the alpha level of the test).

In a situation like the one shown in Figure 1 above, this is done by setting up the rejection threshold $z$ of the test so that $P(Z_N > z|H_0) \leq \alpha$.

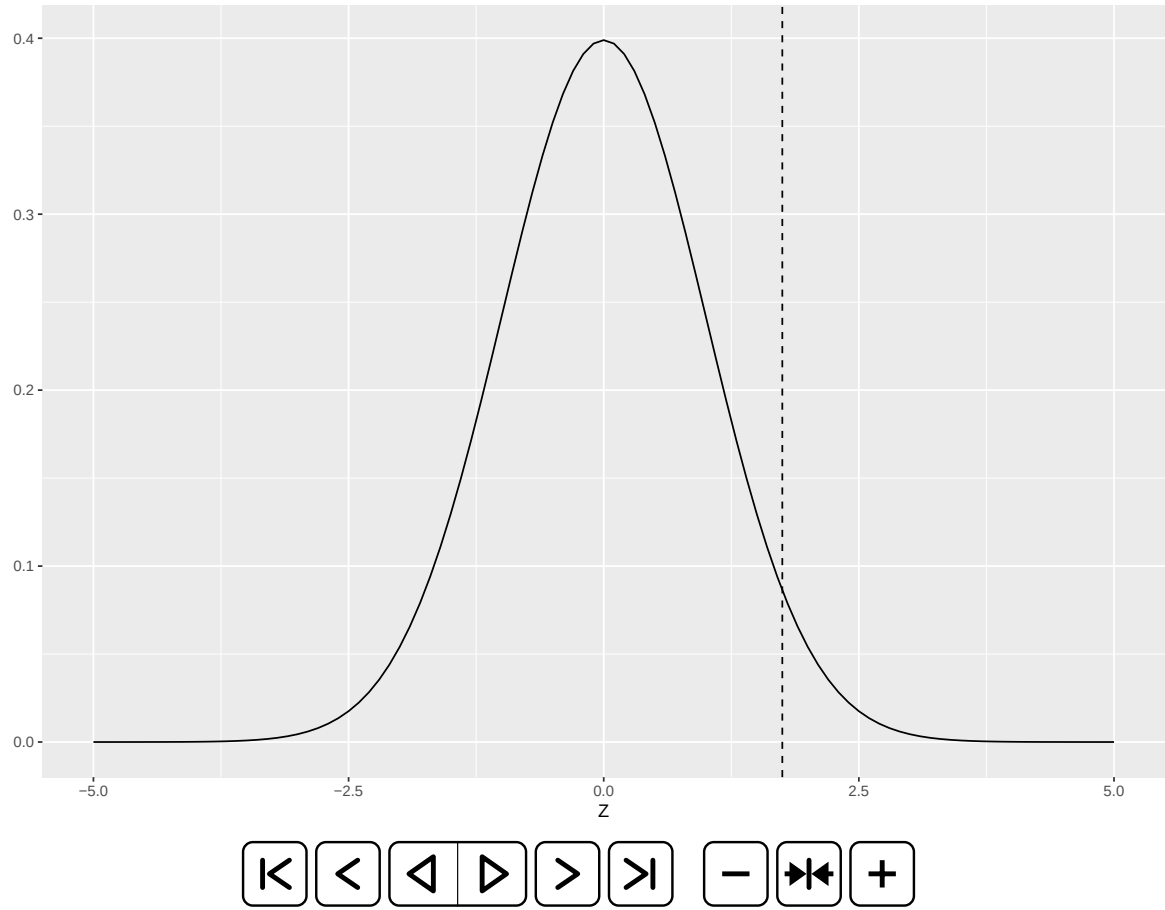
Since $Z_N \sim N(0, 1)$, we can immediately see that $z = Z_{1-\alpha}$. For example, if $\alpha = 0.05$, then $z = 1.645$ $(=Z_{0.95})$ the 95th percentile of the standard normal distribution).

If the one-sided alternative is in the opposite direction, then $z = Z_\alpha$. In the case of two-sided tests, then we set $z = \pm Z_{1-\alpha/2}$, so either too large or too small values of $Z_N$ will result in rejection of $H_0$.

# Type-I error in a single analysis

Here is a simulation of what is supposed to happen, if the null hypothesis is true:

**Figure 2:** One hundred simulated standard normal random variables



# Adding analyses

Now consider the situation where two analyses must be carried out and we conclude that there is a statistically significant difference if at least one of these results in the rejection of the null hypothesis.

Elementary probability tells us that, even under the null hypothesis (i.e., if both tests compare things that are not different), the probability of a (Type-I) error is

$$p = 1 - (1 - \alpha)^2 = 0.0975$$

So, if we carry out both analyses at an alpha level of $\alpha = 0.05$, we would essentially double (!) our Type-I error.

We all know that an (albeit conservative) adjustment that takes care of this is the Bonferroni adjustment, where we carry out each analysis at the revised alpha level $\alpha^* = \alpha/2$, or the original alpha level divided by the number of comparisons.

# Early approaches to monitoring of clinical trials

# Sequential monitoring of clinical trials

Now let's transfer these ideas to interim monitoring of a trial which compares two means (e.g., the mean loss of weight, the reduction in the number of angina episodes, etc.).

In addition to the interim analysis, consider carrying out an interim analysis of the data when $2n$ subjects have been enrolled, with $n < N$, resulting in an interim z-score

$$Z(t) = \frac{Sn}{\sqrt{\nu_n}}$$

where $S_n = \sum_{i=1}^{n} D_i$, $\nu_n = \mathrm{var}(S_n) = 2n\sigma^2$ and $t = n/N$ is the "trial fraction", i.e., the proportion of the total sample size.

We note of course that, at the final analysis, $t = 1$ and $Z(1) = Z_N$.

# Impact on the alpha level

But why stop there?

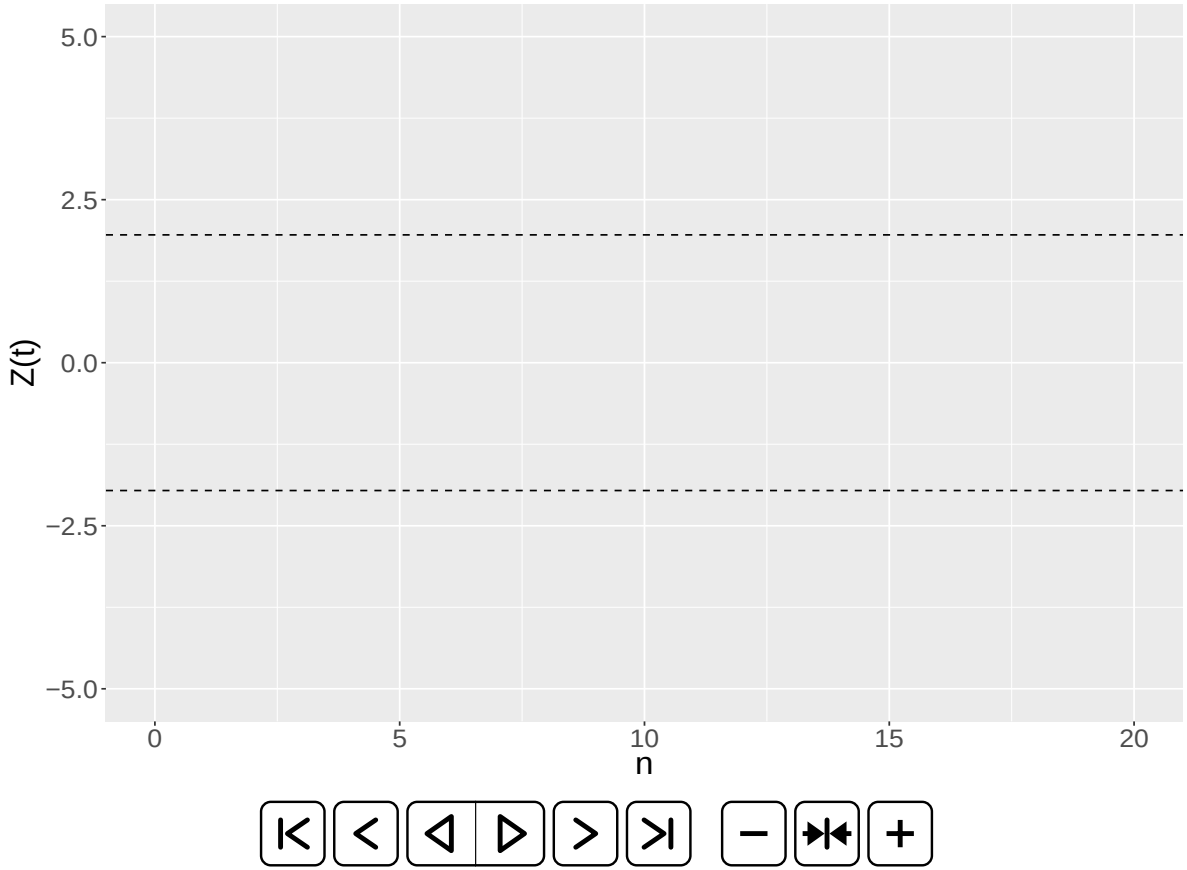
How about we perform an analysis after every single patient pair has been recruited, and obtain the $Z(t)$ statistic after $2n$ patients have been recruited and evaluated $n = 1, \cdots, N$.

More importantly, reject the null hypothesis if $|Z(t)| > z_{1-\alpha/2}$ for, say, $\alpha = 0.05$.

Is there anything wrong with this approach?

# Sampling to a foregone conclusion

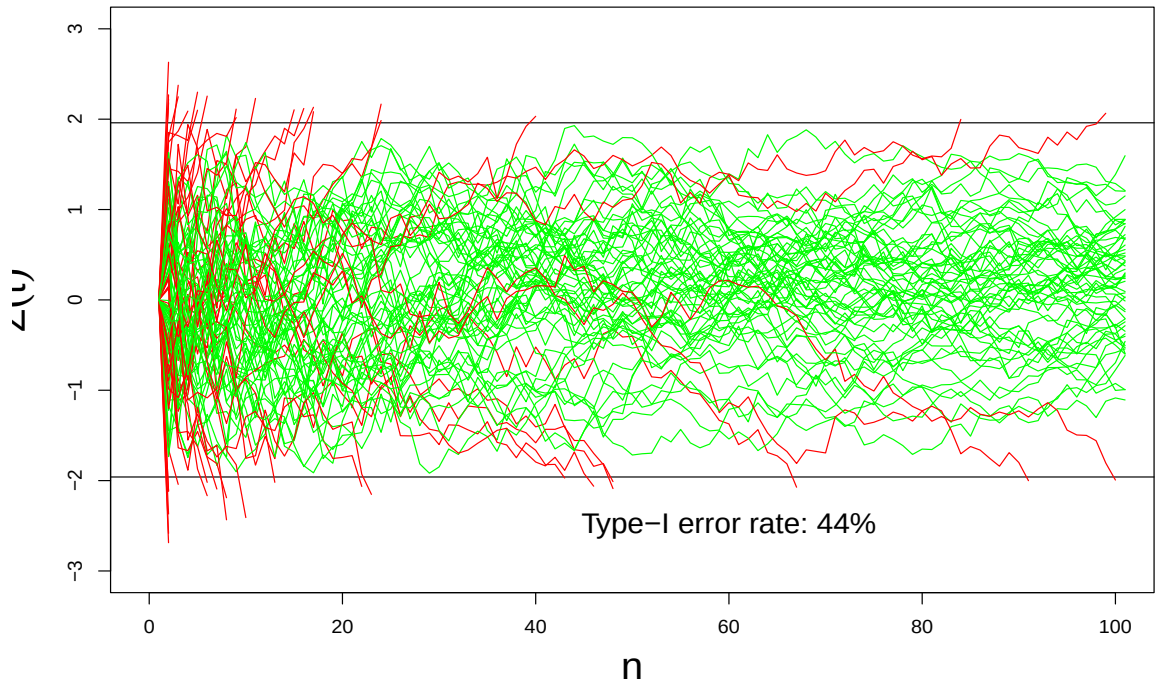
**Figure 3:** Ten simulated trials with $N = 20$, where an interim analysis occurs after every patient pair has been evaluated in the two arms. Red curves symbolize a trial that would declare significance by crossing the upper or lower limit of the rejection region ($\pm z_{0.975} = 1.96$)




From the above figure we see that the impact of repeated analyses is that the type-I error rate is inflated well above the alpha level (maximum allowable rate).

# More simulations

**Figure 4:** One hundred simulated trials of $N = 100$ patients in each treatment arm



# Fix the bounds

A solution is to borrow from the idea of Bonferroni, and adjust the bounds of the rejection region.

To adjust for multiple interim analyses we seek boundaries $c_1, \cdots, c_k$ such that

$$\Pr\left(\cup_{i=1}^{k} Z(t_i) > c_i\right) = \alpha$$

or

$$\Pr\left(\cup_{i=1}^{k} Z(t_i) < -c_i\right) = \alpha$$

for one-tailed tests and

$$\Pr\left(\cup_{i=1}^{k} |Z(t_i)| > c_i\right) = \alpha$$

for two-tailed tests.

In plain language, we seek boundaries such that the total probability of rejecting the null hypothesis in any of the analyses (both interim and final), if the null hypothesis is true, is no higher than $\alpha$.

# Haybittle's idea

Haybittle (Br J Radiol, 1971)[3] proposed the following procedure:

- Use critical value $z = 3$ at the interim analyses
- Use critical value $z = 1.96$ at the final analysis

The author showed by simulation that the alpha level of this procedure does not overly inflate the Type-I error if the number of interim analyses are not too numerous.

[3] Haybittle JL. Repeated assessment of results in clinical trials of cancer treatment. *Br J Radiol*, **44**:793-797. 1971.

# Haybittle's idea: *Improvements*

The latter disadvantage of the inflation of the Type-I error can be fixed by use of the Bonferroni procedure. That is, we proceed as follows:

- At first $k-1$ analyses use $p = 0.001$, that is, reject the null hypothesis at the ith analysis if $|Z(t)| > 3.29$.
- Use Bonferroni to fix last critical value.

For example, if we have $k = 5$ (i.e., 4 interim analyses before the final), we use significance level 0.05 -4(0.001) $= 0.046$ at the final analysis (i.e., reject the null if $|Z(1)| > 1.995$).

This is a very nice procedure because it can be universally applied as long as p values can be computed (i.e., even in hypothesis tests where the distribution theory is difficult to work out).

# Haybittle's idea: *Advantages and disadvantages*

The advantages of this approach are that

- It is simple to implement
- The final test is the same as (or close to) the case of no monitoring. (Note that $z = 1.96 = z_{1-\alpha/2}$ for $\alpha = 0.05$).

The disadvantages of the approach are that

- It is very difficult to stop before the final analysis ($z = 3.0$ is equivalent to a p-value of $p = 0.0013$).
- The procedure, in its original form, still produces a minor inflation of the Type-I error

# The Pocock procedure

Pocock (Biometrika, 1977)[4] suggested the following procedure, based on equally-spaced analyses (i.e., analyses performed at times $t = i/k$ where $i = 1, \cdots, k$).

Determine $c$ such that

$$\Pr\left(\cup_{i=1}^{k} Z(i/k) > c\right) = \alpha$$

[4] Pocock SJ. Group Sequential Methods in the Design and Analysis of Clinical Trials. *Biometrika*, **64**:191-199. 1973.

# The Pocock procedure: *Advantages and disadvantages*

The advantages of this procedure are

- It is a natural extension of the case without monitoring (i.e., going from $z_{1-\alpha}$ to $c$ but still using a constant boundary)
- Uses the same degree of evidence at each analysis
- $c$ is typically smaller than the Haybittle boundary, so the Pocock procedure can stop earlier

The main problem with the Pocock approach is that the p-value at the final analysis is very low (the z-score high)

What's more, Pocock now recommends against his own procedure!

# Boundaries for the Pocock procedure

**Table 1:** Two-tailed boundaries for the Pocock procedure (Proshan, Lan & Wittes, 2006)

| # of looks | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.10$ |
|---|---|---|---|
| 1 | 2.576 | 1.960 | 1.645 |
| 2 | 2.772 | 2.178 | 1.875 |
| 3 | 2.873 | 2.289 | 1.992 |
| 4 | 2.939 | 2.361 | 2.067 |
| 5 | 2.986 | 2.413 | 2.122 |
| 10 | 3.117 | 2.550 | 2.270 |
| 20 | 3.225 | 2.672 | 2.392 |
| $\infty$ | $\infty$ | $\infty$ | $\infty$ |

The conclusion from the table is that the critical values increase significantly with the increase of the number of interim analyses.

# $Z$ and $B$ processes

Now consider the related quantity

$$B(t) = \frac{S_n}{\sqrt{\nu_N}}$$

The interim z-score $Z(t) = \frac{S_n}{\sqrt{\nu_n}}$ at time $t = n/N$ is related to $B(t)$ by the equation

$$\begin{aligned} Z(t) &= \frac{S_n}{\sqrt{\nu_n}} \\ &= \left(\frac{\sqrt{\nu_N}}{\sqrt{\nu_n}}\right)\frac{S_n}{\sqrt{\nu_N}} = \frac{B(t)}{\sqrt{\nu_n}} \end{aligned}$$

The quantity $B(t)$ is related to the so-called "Brownian motion" (a stochastic process with a number of characteristics that help in the modeling of random events (e.g., Lan & Zucker, *Stat Med*, 1993))[5].

While the Brownian motion is the source of fundamental theoretical results in study monitoring, we will not consider it further as it is beyond the scope of this course.

[5] Lan KKG & Zucker DM. Sequential monitoring of clinical trials: the role of information and Brownian motion. *Stat Med*, **12**:753-765. 1993.

# The O'Brien-Fleming method

O'Brien and Fleming (Biometrics, 1979)[6] proposed a related procedure with that of Pocock.

The critical difference of their method is that the boundary is related to $B(t)$ rather than the interim z-score $Z(t)$.

The O'Brien-Fleming boundary is such that

$$\Pr\left(\cup_{i=1}^{k} B(i/k) > c\right) = \alpha$$

Given the relationship between $B(t)$ and $Z(t)$ the above procedure is equivalent to one in terms of $Z(t)$ as follows:

$$\Pr\left(\cup_{i=1}^{k} Z(i/k) > c/\sqrt{i/k}\right) = \alpha$$

So the O-F method scales the level of evidence at each interim analysis by the proportion of the total *information* that has been collected up to that analysis.

[6] O'Brien PC, Fleming TR. A Multiple testing procedure for clinical trials. *Biometrics*, **35**:549-556. 1979.

# Boundaries of the O'Brien-Fleming method

**Table 2:** Two-tailed boundaries for the O'Brien-Fleming procedure (Proshan, Lan & Wittes, 2006)

| # of looks | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.10$ |
|---|---|---|---|
| 1 | 2.576 | 1.960 | 1.645 |
| 2 | 2.580 | 1.977 | 1.678 |
| 3 | 2.595 | 2.004 | 1.710 |
| 4 | 2.609 | 2.024 | 1.733 |
| 5 | 2.621 | 2.040 | 1.751 |
| 10 | 2.660 | 2.087 | 1.801 |
| 20 | 2.695 | 2.126 | 1.842 |
| $\infty$ | 2.807 | 2.241 | 1.960 |

Note that, asymptotically (i.e,. at infinity), the critical values correspond to z-scores with $\alpha^* = \alpha/2$.

# The O'Brien-Fleming method: *Advantages and disadvantages*

The big advantage of the O'Brien-Fleming procedure is that, at the final analysis, the p-value is close to the original alpha level.

This is counter-balanced by the fact that the procedure will stop the trial more infrequently early (when evidence is more limited) compared to the Pocock procedure.

This latter consideration may not be very problematic as, intuitively, there is great resistance for stopping early when information is limited.

# Example: A study with $k = 5$ analyses: *bounds*

For a study with $k = 5$ total analyses, the three boundaries give the following critical values (Table 3) and corresponding p-value boundaries (Table 4):

**Table 3:** Two-tailed boundaries for the O'Brien-Fleming, Pocock and Haybittle- Peto procedures

| # of looks | O'Brien-Fleming | Pocock | Haybittle-Peto |
|---|---|---|---|
| 1 | $\pm$ 4.562 | $\pm$ 2.413 | $\pm$ 3.290 |
| 2 | $\pm$ 3.226 | $\pm$ 2.413 | $\pm$ 3.290 |
| 3 | $\pm$ 2.634 | $\pm$ 2.413 | $\pm$ 3.290 |
| 4 | $\pm$ 2.281 | $\pm$ 2.413 | $\pm$ 3.290 |
| 5 | $\pm$ 2.040 | $\pm$ 2.413 | $\pm$ 1.685 |

# Example: A study with $k = 5$ analyses: *p-values*}

The p-values corresponding to the boundaries Table 3 are given in the following Table:

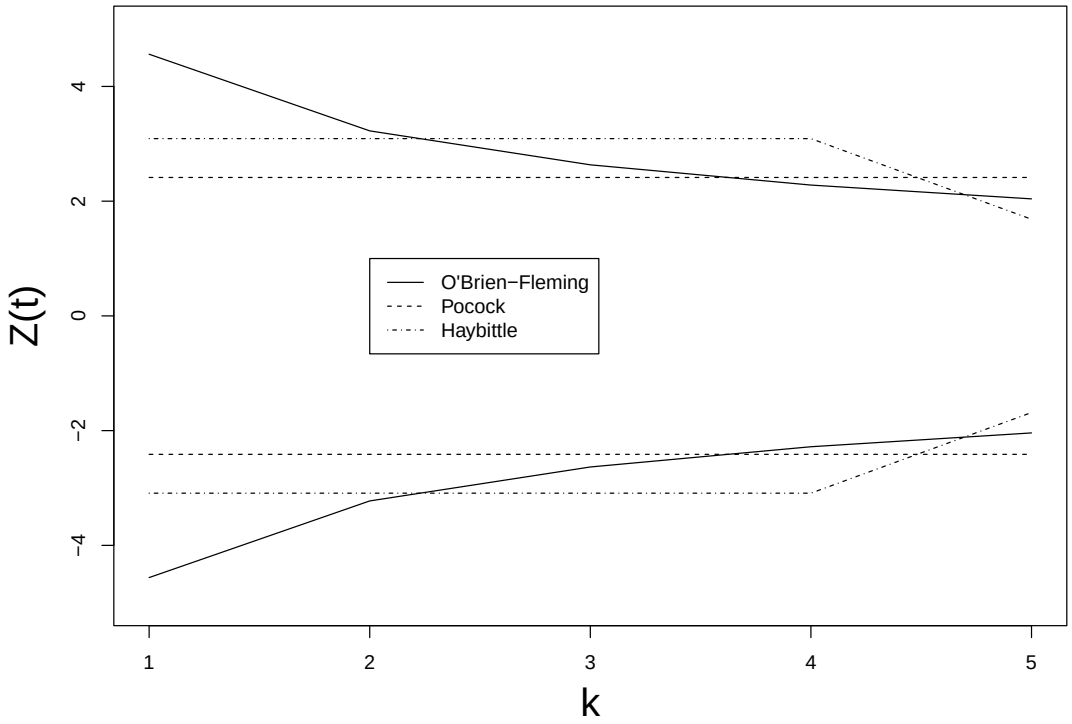
**Table 4:** Two-tailed p-values for the O'Brien-Fleming, Pocock and Haybittle- Peto procedures

| # of looks | O'Brien-Fleming | Pocock | Haybittle-Peto |
|---|---|---|---|
| 1 | 5.067e-06 | 0.016 | 0.001 |
| 2 | 0.001 | 0.016 | 0.001 |
| 3 | 0.008 | 0.016 | 0.001 |
| 4 | 0.023 | 0.016 | 0.001 |
| 5 | 0.041 | 0.016 | 0.046 |

# Example: A study with $k = 5$ analyses: *Graphical representation*

The three boundaries are shown in the following Figure:

**Figure 5:** Two-tailed critical values for the O'Brien-Fleming, Pocock and Haybittle-Peto procedures



# Implementation through R

Implementation through R is done through the gsDesign package (Andersen, 2010).

gsDesign( k = 3, test.type = 4, alpha = 0.025, beta = 0.1, astar = 0, delta = 0, n.fix = 1, timing = 1, sfu = sfHSD, sfupar = -4, sfl = sfHSD, sflpar = -2, tol = 1e-06, r = 18, n.I = 0, maxn.IPlan = 0, nFixSurv = 0, endpoint = NULL, delta1 = 1, delta0 = 0, overrun = 0, usTime = NULL, lsTime = NULL )

# O'Brien-Fleming bounds

We can use the gsDesign package to produce the O'Brien-Fleming bounds we used previously (Table 2):

```
# O-F bounds for k=5 and alpha=0.05
x.OF=gsDesign(k=5, test.type = 2, alpha = 0.025, sfu='OF')
data.frame(`lower bound`=x.OF$lower$bound, `upper bound`=x.OF$upper$bound)
```

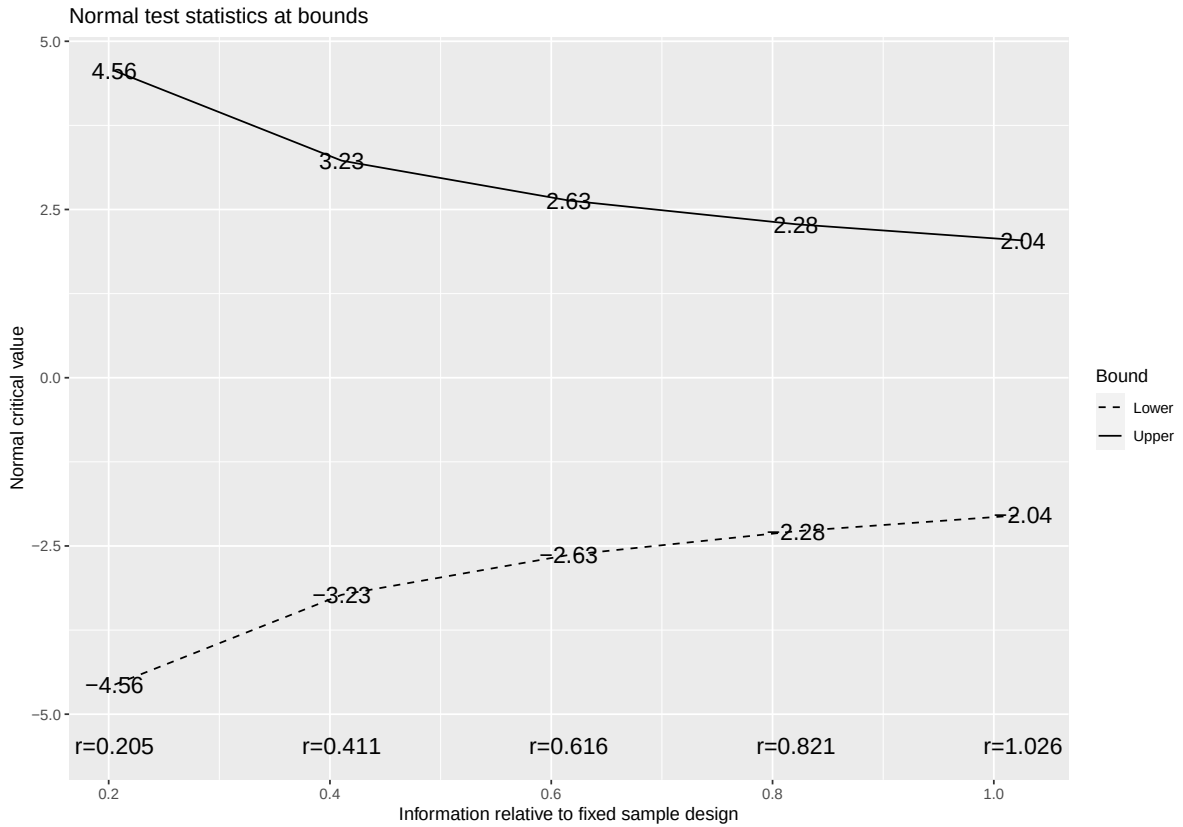
```
##   lower.bound upper.bound
## 1   -4.561743    4.561743
## 2   -3.225639    3.225639
## 3   -2.633723    2.633723
## 4   -2.280871    2.280871
## 5   -2.040073    2.040073
```

# Graphics: *O'Brien-Fleming bounds*

We can produce simple graphics as well.

```
plot(x.OF)
```

**Figure 6:** Simple graphics for the O'Brien-Fleming bounds produced by the `gsDesign` package



# Pocock bounds

We can use the gsDesign package to produce the Pocock bounds we used previously (Table 1):

```
# Pocock bounds
x.PC=gsDesign(k=5, test.type = 2, alpha = 0.025, sfu='Pocock')
data.frame(`lower bound`=x.PC$lower$bound, `upper bound`=x.PC$upper$bound)
```

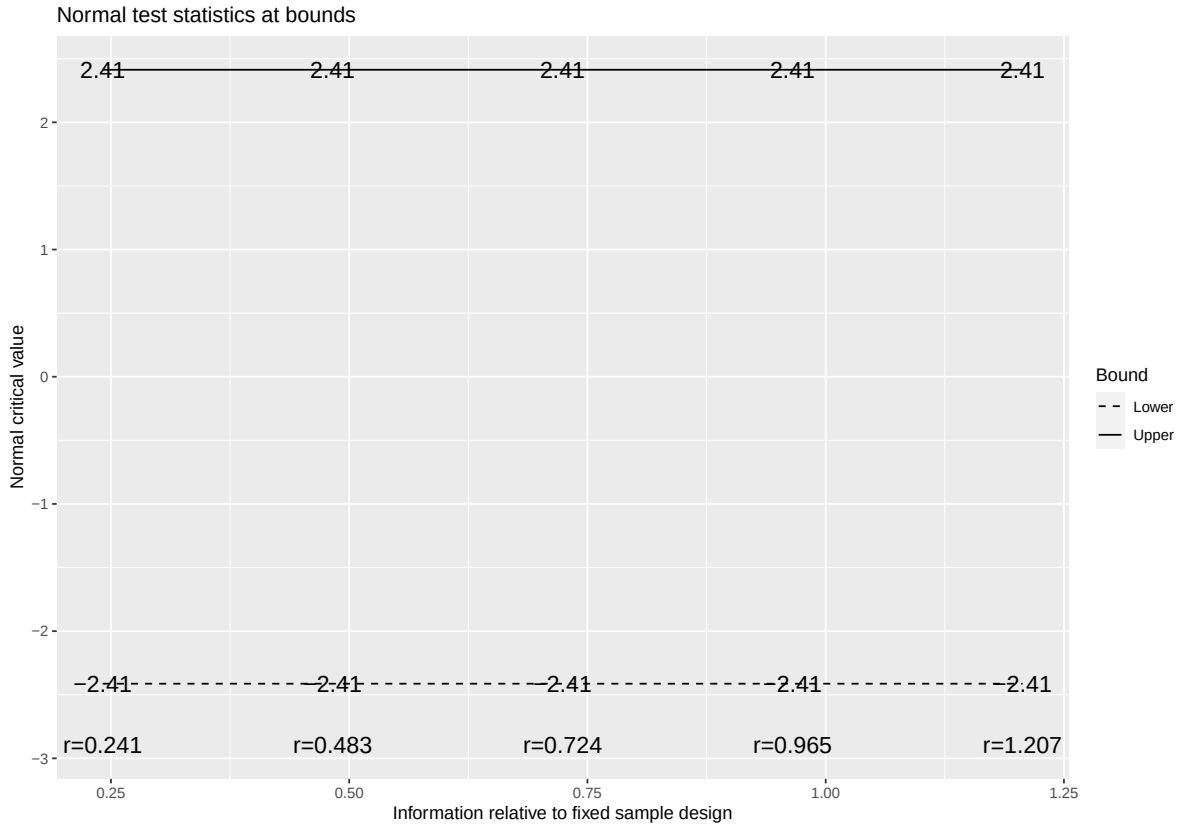
```
##   lower.bound upper.bound
## 1   -2.413176    2.413176
## 2   -2.413176    2.413176
## 3   -2.413176    2.413176
## 4   -2.413176    2.413176
## 5   -2.413176    2.413176
```

# Graphics: *Pocock bounds*

Here is a plot of the Pocock bounds for the same design.

```
plot(x.PC)
```

**Figure 7:** Simple graphics for the Pocock bounds produced by the `gsDesign` package



# Monitoring 2G: Spending functions

# Spending functions

A game-changer in monitoring of clinical trials was the idea of the *spending function*.

The seminal reference for this methodology is the paper by Lan & DeMets (Biometrika, 1983)[7] that showed that sequential boundaries can be computed without knowing the timing of the analyses in advance.

Spending functions show the way that the total alpha is "spent" through the interim and final analyses. They are crucial in monitoring of a trial for the following reasons:

- The Pocock and O'Brien-Fleming boundaries, as originally proposed, require equal spacing of the analyses but DSMB meet when the schedules permit
- Analysis times may not be easily predictable in advance
- Extra analyses may be scheduled during the implementation of the study

[7] Lan KKG and DeMets DL. Discrete sequential boundaries for clinical trials, *Biometrika*, **70**, 659–663. 1983

# Alpha spending functions

An alpha spending function $\alpha(t)$, with $\alpha(0) = 0$ and $\alpha(1) = \alpha$ of the form[8]

$$\alpha(t) = \Pr\left\{\cup_{i=1}^{k} |Z(t)| > c_i | H_0\right\}$$

For a given schedule of analyses $\tau_1, \cdots, \tau_k$, this splits $\alpha$ in probabilities $\alpha(\tau_i)$, $i = 1, \cdots, k$,

$$\alpha(\tau_1) = \Pr\left\{|Z(\tau_1)| > c_1 | H_0\right\}$$

and for $i = 2, \cdots, k$

$$\alpha(\tau_k) = \Pr\left\{|Z(\tau_1)| < c_1, \cdots, |Z(\tau_{k-1})| < c_{k-1}, |Z(\tau_k)| > c_k | H_0\right\}$$

with $\sum_{i=1}^{k} \alpha(\tau_i) = \alpha$. These probabilities are calculated by numerical methods (Armitage, McPherson & Rowe, JRSS A', 1969)[9].

[8] Here we focus on two-sided alternatives. One-sided hypotheses can also be straightforwardly accommodated.

[9] P. Armitage, C. K. McPherson and B. C. Rowe. Repeated Significance Tests on Accumulating Data. *J Roy Stat Soc A*, **132**(2): 235-244. 1969.

# Example: Equal alpha spending over $k = 5$ interim analyses

Consider the situation where sample size calculations have determined that, with $\alpha = 0.05$, the requisite sample size for the fixed (i.e., one-analysis) design is $N = 100$ per group.

Suppose that we want to carry out $k = 5$ total analyses (i.e., four interim and one final analysis) at equal time points (i.e., after $n_1 = 20$ per group, $n_2 = 40$ per group and so on) and we want to spend $\alpha = 0.05$ equally, over these $k = 5$ analysis. In other words, we want

$$\begin{array}{c} \alpha(1) = 0.01 \\ \alpha(2) = 0.01 \\ \alpha(3) = 0.01 \\ \alpha(4) = 0.01 \\ \hline \alpha(5) = 0.01 \end{array}$$

# Example: Equal alpha spending: *Bounds*

The critical values $c_1, \cdots, c_5$ are given from the following output:

```
x.LN<-gsDesign(k=5, sfu=sfHSD, sfupar = 0, test.type = 2)
x.LN
```

```
## Symmetric two-sided group sequential design with
## 90 % power and 2.5 % Type I Error.
## Spending computations assume trial stops
## if a bound is crossed.
##
##            Sample
##             Size
##   Analysis Ratio*  Z   Nominal p Spend
##          1  0.227 2.58    0.0050 0.005
##          2  0.454 2.49    0.0064 0.005
##          3  0.682 2.41    0.0080 0.005
##          4  0.909 2.34    0.0097 0.005
##          5  1.136 2.28    0.0114 0.005
##      Total                       0.0250
##
## ++ alpha spending:
##  Hwang-Shih-DeCani spending function with gamma = 0.
```

# Comments

This is the $\alpha$ spent for the scenarios where the experimental treatment is better.

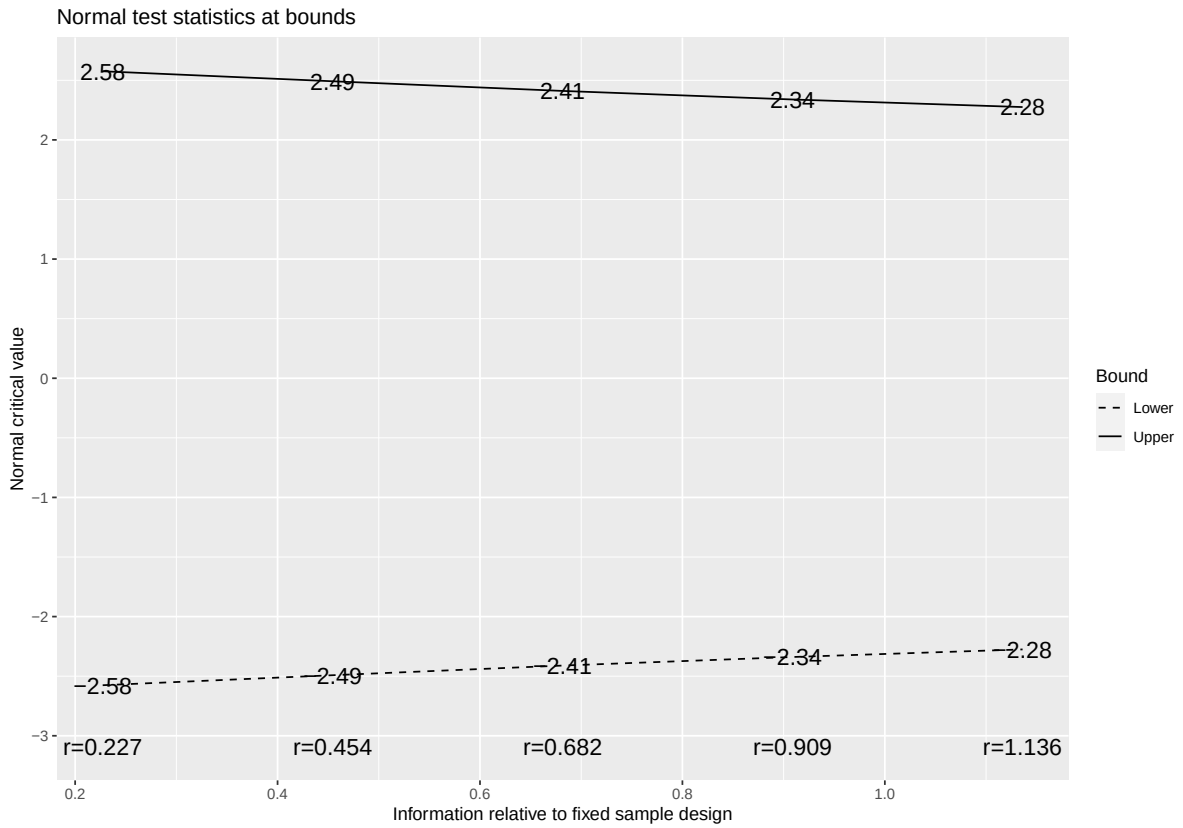
An equal amount (i.e., $\alpha/2 = 0.025$) is spent for scenarios leading to the standard treatment being superior.

Note also that the sample size $N$ has been inflated. More on this later.

# Example: Equal alpha spending: *pictorial representation*

The figure representing the previous analysis is shown below:

**Figure 8:** Boundaries for equal spending of the alpha level over $k = 5$ analyses




Note that the upper and lower bounds are not parallel!

# Continuous spending functions

The advantage of the alpha spending approach is that they can accommodate arbitrary interim analysis schedules.

The limitation of this approach is that the schedule of the interim analyses must be known *a priori*.

In their paper, Lan & DeMets proposed a continuous spending function approach.

They suggested that alpha spending functions can be arbitrary as long as

- $\alpha(0) = 0$
- $\alpha(1) = \alpha$

Note: The main advantage of this approach is that the analyses do not have to be equally-spaced and the timing *or the number* of the analyses do not have to be known in advance!

# L&D O-F and Pocock-like spending functions

The main advantage of the continuous spending-function approach is that interim analyses can be undertaken at any point during the implementation of the study.

Two spending functions, proposed by Lan & DeMets, approximate the Pocock and the O'Brien-Fleming procedures are given below:

- Pocock-like spending function

$$\alpha_P(t) = \alpha \log\{1 + (e-1)t\}$$

- O'Brien-Fleming-like spending function

$$\alpha_{OB}(t) = 4\{1 - \Phi(z_{1-\alpha/4}/\sqrt{t}\}$$

# Example: O-F and Pocock spending functions for $k = 5$: *Cum. alpha level spent*

For example, with $k = 5$ the Pocock and O'Brien-Fleming spending functions are

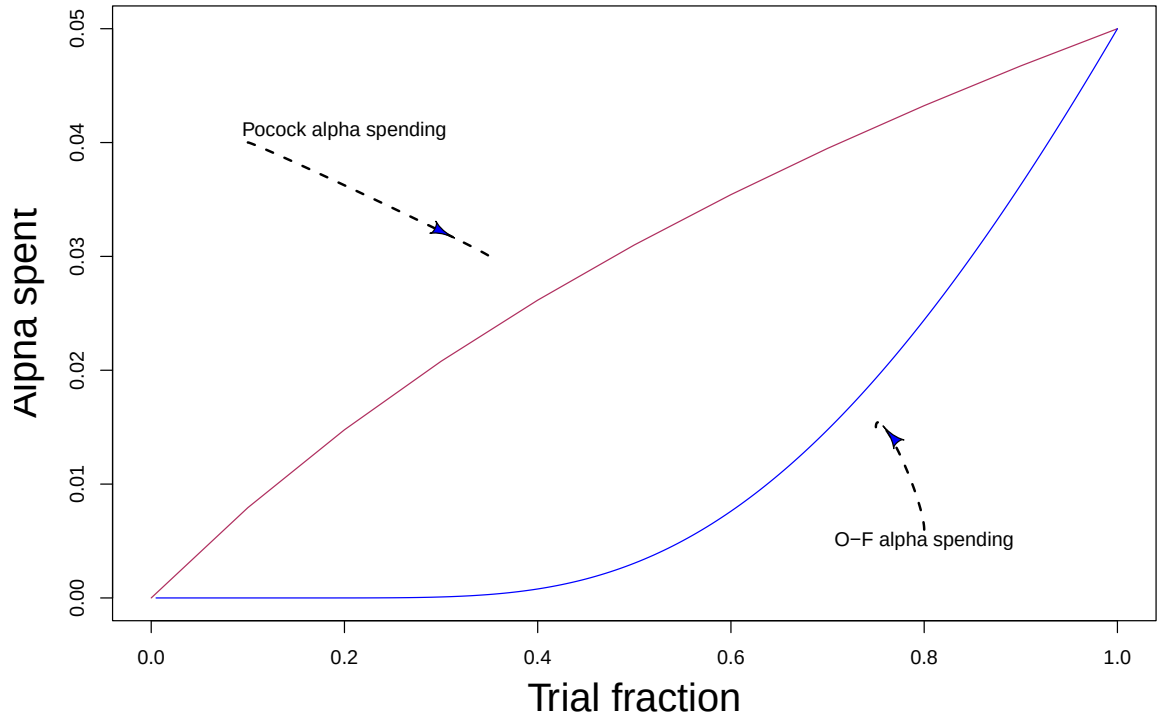
| | Cumulative $\alpha$ spent | | Bounds | |
|---|---|---|---|---|
| Information time | Pocock | O-F | Pocock | O-F |
| 0.2 | 0.016 | 0.000 | $\pm$ 2.41 | $\pm$ 4.56 |
| 0.4 | 0.028 | 0.001 | $\pm$ 2.41 | $\pm$ 3.23 |
| 0.6 | 0.037 | 0.009 | $\pm$ 2.41 | $\pm$ 2.63 |
| 0.8 | 0.044 | 0.026 | $\pm$ 2.41 | $\pm$ 2.28 |
| 1.0 | 0.050 | 0.050 | $\pm$ 2.41 | $\pm$ 2.04 |

So, although both functions spend the same amount of $\alpha$ by the end of the study, the Pocock spending function spends alpha much faster than the O'Brien-Fleming spending function.

# Spending functions: *Pictorial representation*

The cumulative rate of alpha spending of the two spending functions is given pictorially in the following figure:

**Figure 9:** Cum. $\alpha$ spending in Pocock and O'Brien-Fleming spending functions.



# The Hwang, Shih & DeCani family of spending functions

In a 1990 paper[10], Hwang, Shih & DeCani introduced the following general family of alpha spending functions:

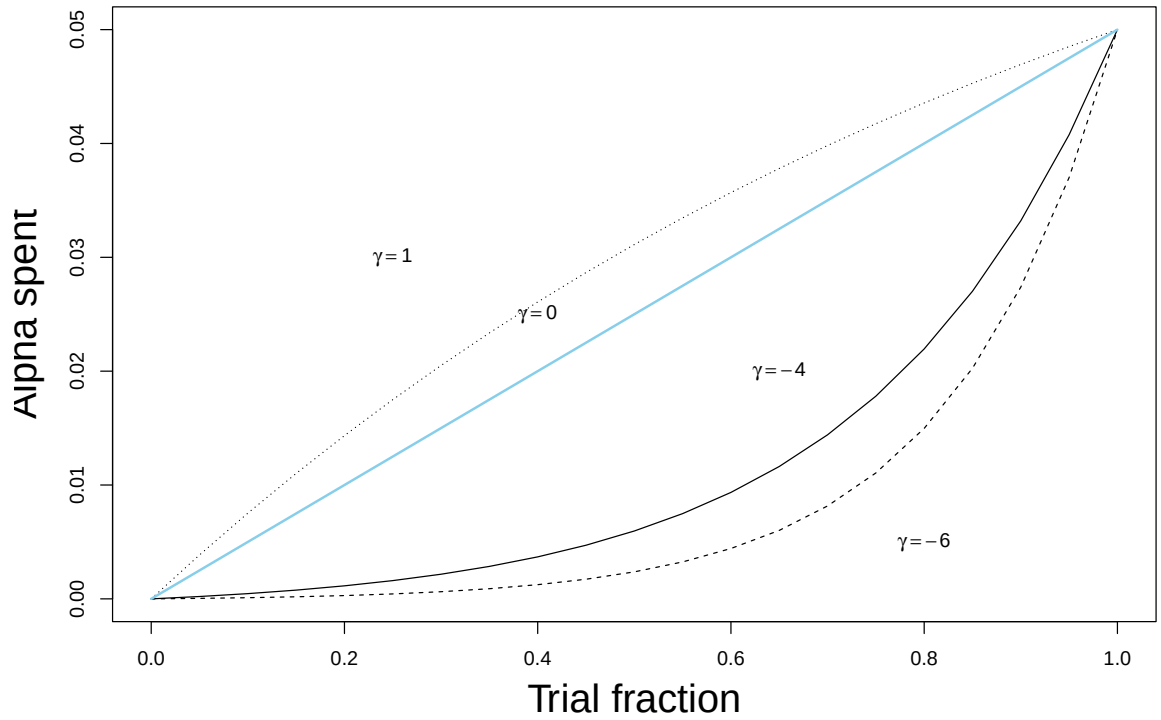
$$\alpha(t) = \begin{cases} \alpha(1 - e^{-\gamma t})/(1 - e^{-\gamma}) & \text{if} \gamma \neq 0 \\ \alpha t & \text{if} \gamma = 0 \end{cases}$$

This is a flexible family of spending functions which can also accommodate approximate functions for the O'Brien & Fleming ($\gamma = -4$) and Pocock approaches ($\gamma = 2$), as well as the linear (equal) spending of the alpha level we saw earlier ($\gamma = 0$, see Figure 8).

[10] Hwang IK, Shih WJ, DeCani JS. Group sequential designs using a family of type I error probability spending functions. *Stat Med*, **9**:1439-1445. 1990.

# The HSD family of spending functions

**Figure 10:** Hwang-Shih-DeCani family of spending functions




From the figure it is clear that alpha is spent more quickly the larger the value of $\gamma$.