Adaptive Sample Size Re-estimation: Design and Inference

Sarah Emerson and Scott Emerson

Outline

1. Foundations: Group Sequential Designs
   - Clinical Trial Design
   - Example Setting
   - Design Comparison
2. Inference following Group Sequential Designs
   - Inference Goals
   - Inference Approaches
   - Inference Optimality Criteria
3. Adaptive Sequential Designs
   - Forms of Adaptation Considered
   - Considerations in Adapting Future Sampling Path
   - Types of Adaptation Rules
   - Adaptive Designs using Standard Group Sequential Software

1. Foundations: Group Sequential Designs
   - Clinical Trial Design
   - Example Setting
   - Design Comparison
2. Inference following Group Sequential Designs
   - Inference Goals
   - Inference Approaches
   - Inference Optimality Criteria
3. Adaptive Sequential Designs
   - Forms of Adaptation Considered
   - Considerations in Adapting Future Sampling Path
   - Types of Adaptation Rules
   - Adaptive Designs using Standard Group Sequential Software

Considerations in Designing Clinical Trials

- Goal: determine efficacy of a treatment (or difference between treatments)
- One- or two-sided hypothesis test based on a statistic of interest, chosen to be scientifically/clinically relevant

Considerations in Designing Clinical Trials

- Scientific:
  - Answer clinical question of interest with useful estimates and intervals
  - Evaluate mechanistic questions
- Ethical:
  - Quickly identify treatments that cause harm
  - Get effective treatments to patients quickly
  - Release patients from a less promising trial so that they might participate in other trials
- Financial:
  - Patient costs expensive; limit number of patients required
  - Long duration increases operating costs
  - Bringing a good drug to market sooner allows an earlier profit, advantage in competition

1. Foundations: Group Sequential Designs
   - Clinical Trial Design
   - Example Setting
   - Design Comparison
2. Inference following Group Sequential Designs
   - Inference Goals
   - Inference Approaches
   - Inference Optimality Criteria
3. Adaptive Sequential Designs
   - Forms of Adaptation Considered
   - Considerations in Adapting Future Sampling Path
   - Types of Adaptation Rules
   - Adaptive Designs using Standard Group Sequential Software

Example Study Setting

- Goal: Determine efficacy of new/experimental treatment (A) relative to standard of care/placebo control (B).
- Protocol:
  - Accrue n subjects
  - Randomize at 1:r ratio to treatment A or B (we will consider r = 1, so 1:1 randomization)
    - $n_A = \frac{n}{1+r} = \frac{n}{2}$ = Number of subjects receiving treatment A
    - $n_B = \frac{rn}{1+r} = \frac{n}{2}$ = Number of subjects receiving treatment B
  - Measure outcomes
    - X_{Ai} for subject i receiving experimental treatment A
    - X_{Bi} for subject i receiving control treatment B.
    - For now, we assume outcomes are immediately available for all subjects.

Example Study Setting

- Population parameters:
  - $\mu_A = E[X_{Ai}]$ (unknown)
  - $\mu_B = E[X_{Bi}]$ (unknown)
  - $\sigma^2 = \text{Var}[X_{Ai}] = \text{Var}[X_{Bi}]$ (common variance, assumed known for now, and taken to be $\sigma^2 = 1$)
- Parameter of interest: $\theta = \mu_A - \mu_B$
  - Null hypothesis $H_0 : \theta = \theta_0 = 0$ (no difference in mean treatment effect)
  - Alternative hypothesis $H_0 : \theta = \theta_A = 0.46$ (experimental treatment mean is larger, indicating superiority over control)

Example Study Setting

- Sample statistic notation:
  - $\bar{X}_A = \frac{1}{n_A}\sum_{i=1}^{n_A} X_{Ai}$ (sample mean of treatment A group)
  - $\bar{X}_B = \frac{1}{n_B}\sum_{i=1}^{n_B} X_{Bi}$ (sample mean of treatment B group)
- Note:

$$\begin{aligned}\text{Var}[\bar{X}_A - \bar{X}_B] &= \sigma^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right) \\ &= 1\left(\frac{2}{n} + \frac{2}{n}\right) \quad \text{if } n_A = n_B = \frac{n}{2} \text{ and } \sigma^2 = 1 \\ &= \frac{4}{n}\end{aligned}$$

Possible Designs

Fixed Sample

- Gather n = 290 subjects randomized at 1:1 ratio to treatments A and B ($n_A = n_B = 145$)
- Measure outcomes X_{Ai} or X_{Bi} for each subject.
- Compute two-sample z-statistic:

$$z(\theta_0) = \frac{\overline{X}_A - \overline{X}_B - \theta_0}{\sqrt{\sigma^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} = \frac{\overline{X}_A - \overline{X}_B}{\sqrt{4\sigma^2/n}} = \frac{\sqrt{n}}{2}\left(\overline{X}_A - \overline{X}_B\right)$$

- Reject H_0 if $z(\theta_0) > z_\alpha = \Phi^{-1}(1-\alpha) = 1.96$
- Equivalently, reject H_0 if $(\overline{X}_A - \overline{X}_B) > \frac{1.96}{\sqrt{n/2}} = 0.2298$

Significance Level $= 0.025$
Power $= 0.975$ at Design Alternative $\theta_A = 0.46$

Sequential Trials

- Ethical and financial issues in clinical trials may be improved by performing multiple interim analyses during the trial
- Maintain control of the significance level and the power at the design alternative by adjusting the decision criteria at each analysis
- Allowing early stopping of the trial at interim analyses typically reduces the expected trial duration and number of subjects required

Possible Designs

Fixed Sample

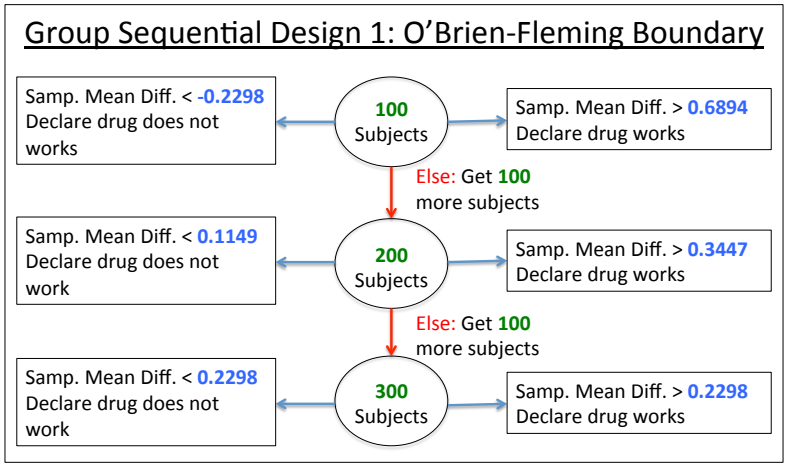
- Gather $n = 290$ subjects randomized at 1:1 ratio to treatments A and B ($n_A = n_B = 145$)
- Measure outcomes X_{Ai} or X_{Bi} for each subject.
- Compute two-sample z-statistic:

$$z(\theta_0) = \frac{\overline{X}_A - \overline{X}_B - \theta_0}{\sqrt{\sigma^2(\frac{1}{n_A} + \frac{1}{n_B})}} = \frac{\overline{X}_A - \overline{X}_B}{\sqrt{4\sigma^2/n}} = \frac{\sqrt{n}}{2}\left(\overline{X}_A - \overline{X}_B\right)$$

- Reject H_0 if $z(\theta_0) > z_\alpha = \Phi^{-1}(1-\alpha) = 1.96$
- Equivalently, reject H_0 if $(\overline{X}_A - \overline{X}_B) > \frac{1.96}{\sqrt{n}/2} = 0.2298$

Significance Level $= 0.025$
Power $= 0.975$ at Design Alternative $\theta_A = 0.46$

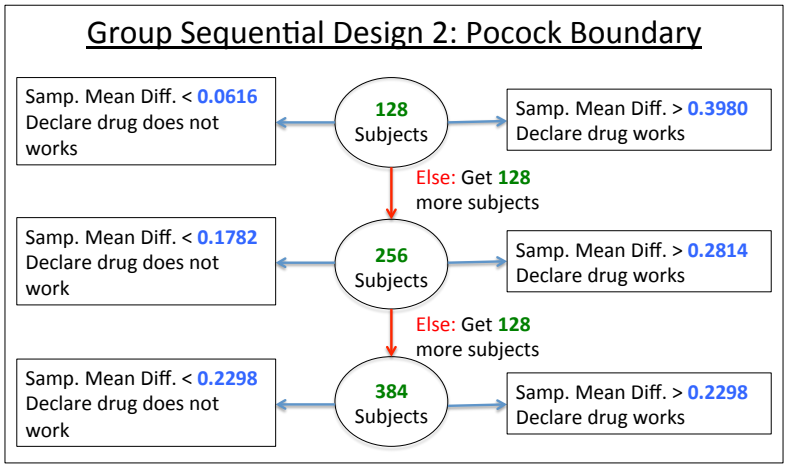
Possible Designs



Significance Level $= 0.025$

Power $= 0.975$ at Design Alternative $\theta_A = 0.46$

Possible Designs



Significance Level $= 0.025$

Power $= 0.975$ at Design Alternative $\theta_A = 0.46$

Group Sequential Trials: Definition

A group sequential design is defined by a **Stopping Rule** consisting of:

1. **Analysis Times:** A set of J analysis times $n_1, n_2, \ldots, n_J$ defined in terms of the amount of *statistical information* accumulated
   - n_j = total number of subjects or number of events (across all arms) observed up to the jth analysis.
   - n_{Aj}, n_{Bj} = number of subjects on arm A or B, respectively, observed up to the jth analysis. $n_{Aj} + n_{Bj} = n_j$.

Group Sequential Trials: Definition

Incremental Analysis Times/Sample Sizes:

- $n_j^* = n_j - n_{j-1}$ = *incremental* number of subjects/events (across all arms) added between $(j-1)$st and jth analyses
- $n_{Aj}^* = n_{Aj} - n_{A(j-1)}$, $n_{Bj}^* = n_{Bj} - n_{B(j-1)}$ = *incremental* number of subjects/events added on each arm

Group Sequential Trials: Definition

2. **Test Statistic:** A test statistic T_j calculated from the data accumulated so far at each analysis time $j = 1, 2, \ldots J$. Examples:
   - Partial Sum/Partial Sum Difference:

$$S_j = \sum_{i=1}^{n_{Aj}} X_{Ai} - \sum_{i=1}^{n_{Bj}} X_{Bi}$$

   - MLE:

$$\begin{aligned} \hat{\theta}_j &= \frac{1}{n_{Aj}} \sum_{i=1}^{n_{Aj}} X_{Ai} - \frac{1}{n_{Bj}} \sum_{i=1}^{n_{Bj}} X_{Bi} \\ &= \bar{X}_{Aj} - \bar{X}_{Bj} \end{aligned}$$

Group Sequential Trials: Definition

   - z-statistic:

$$Z_j = \frac{\hat{\theta}_j - \theta_0}{\sqrt{\frac{\sigma^2}{n_{Aj}} + \frac{\sigma^2}{n_{Bj}}}} = \frac{\hat{\theta}_j - \theta_0}{\sigma\sqrt{\frac{1}{n_{Aj}} + \frac{1}{n_{Bj}}}}$$

   - Fixed-sample p-value:

$$P_j = 1 - \Phi(Z_j)$$

Group Sequential Trials: Definition

Incremental Test Statistics:

- Incremental Partial Sum/Partial Sum Difference:

$$S_j^* = \sum_{i=n_{A(j-1)}+1}^{n_{Aj}} X_{Ai} - \sum_{i=n_{B(j-1)}+1}^{n_{Bj}} X_{Bi}$$

- Incremental MLE:

$$\begin{aligned}\hat{\theta}_j^* &= \frac{1}{n_{Aj}^*} \sum_{n_{A(j-1)}+1}^{n_{Aj}} X_{Ai} - \frac{1}{n_{Bj}^*} \sum_{i=n_{B(j-1)}+1}^{n_{Bj}} X_{Bi} \\ &= \bar{X}_{Aj}^* - \bar{X}_{Bj}^*\end{aligned}$$

Group Sequential Trials: Definition

- Incremental z-statistic:

$$Z_j^* = \frac{\hat{\theta}_j^* - \theta_0}{\sqrt{\frac{\sigma^2}{n_{Aj}^*} + \frac{\sigma^2}{n_{Bj}^*}}} = \frac{\hat{\theta}_j^* - \theta_0}{\sigma\sqrt{\frac{1}{n_{Aj}^*} + \frac{1}{n_{Bj}^*}}}$$

- Incremental Fixed-sample p-value:

$$P_j^* = 1 - \Phi(Z_j^*)$$

Group Sequential Trials: Definition

3. **Stopping Boundary:** A set of boundary values $a_j \leq b_j \leq c_j \leq d_j$ for each analysis time $j = 1, 2, \ldots, J$
   - Decision rule:

$T_j \geq d_j$	Stop trial at j^{th} analysis and accept upper hypothesis
$c_j < T_j < d_j$	Continue trial
$b_j \leq T_j \leq c_j$	Stop trial at j^{th} analysis and accept null hypothesis (two-sided test)
$a_j < T_j < b_j$	Continue trial
$T_j \leq a_j$	Stop trial at j^{th} analysis and accept lower hypothesis

Group Sequential Trials: Definition

- The regions $\mathcal{C}_j = (a_j, b_j) \cup (c_j, d_j)$ are called the *continuation regions* at analysis j.
  - ⋆ If the test statistic belongs to this interval or set of intervals, the trial is continued beyond analysis j.
- The complement of the continuation regions $\mathcal{S}_j = \mathcal{C}_j'$ are called the *stopping regions* at analysis j.
  - ⋆ If the test statistic belongs to this interval or set of intervals, the trial is stopped at analysis j.

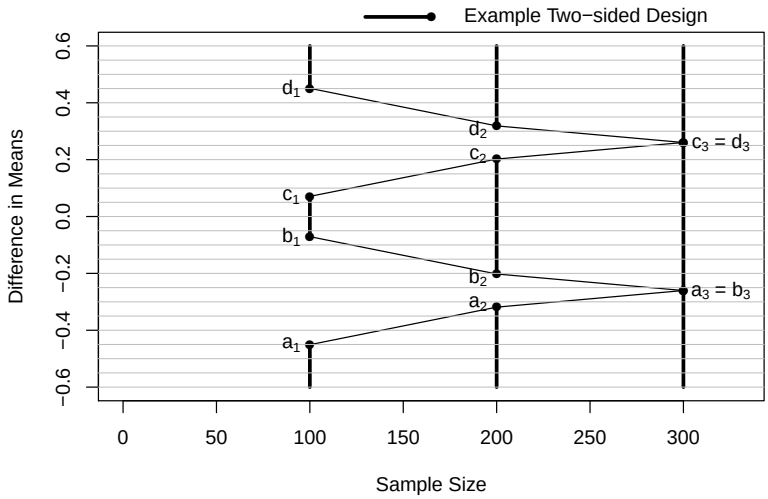
Group Sequential Trials: Definition

- (a_j, b_j, c_j, d_j) for $j = 1, \ldots, J$ must be chosen to obtain desired significance level α:

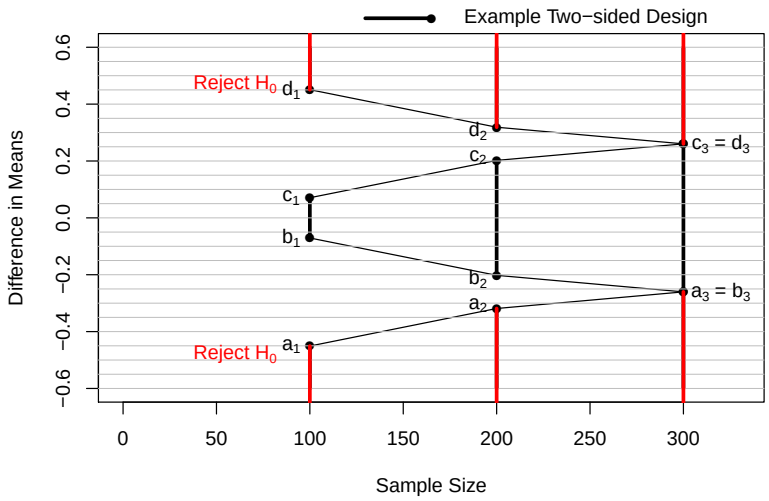
$$P_{\theta_0}(\text{Reject } H_0 \text{ at any } j = 1, \ldots, J) = \alpha$$

- $a_J = b_J$ and $c_J = d_J$ to guarantee that a decision is made by the final analysis
- For a one-sided design, $b_j = c_j$ for all j.

Example Stopping Boundary Figure

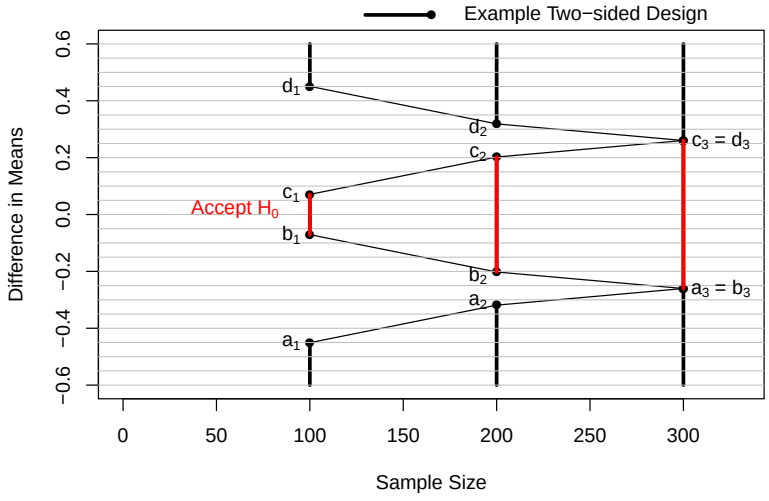


Example Stopping Boundary Figure



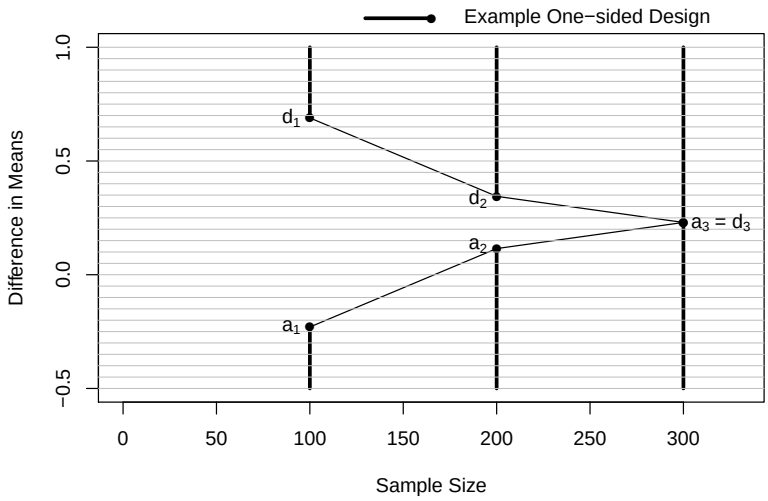
Foundations: Group Sequential Designs
Inference following Group Sequential Designs
Adaptive Sequential Designs
Clinical Trial Design
Example Setting
Design Comparison

Example Stopping Boundary Figure

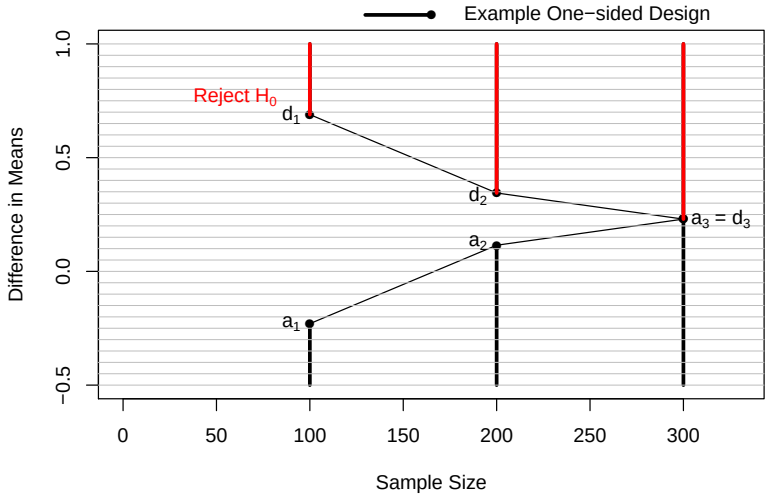


Sarah Emerson and Scott Emerson
Adaptive Designs
24 / 143

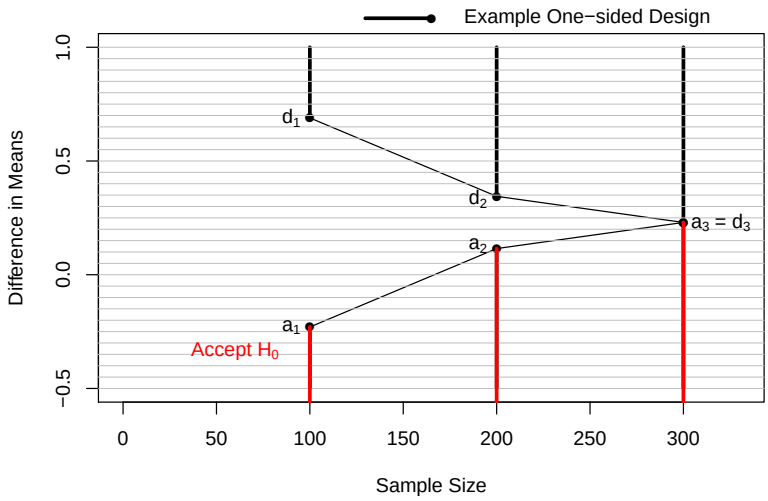
Example Stopping Boundary Figure



Example Stopping Boundary Figure



Example Stopping Boundary Figure



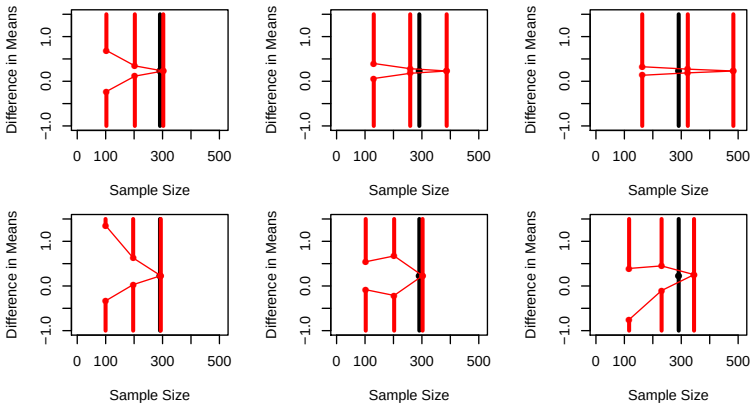
Stopping Boundary Specification

- Great flexibility in choice of boundary
  - Error spending designs
  - Unified family of group sequential designs (Kittelson and Emerson 1999), includes
    - O'Brien-Fleming
    - Pocock
    - Wang and Tsiatis
    - ...and others

  as special cases.

Stopping Boundary Specification

Several Possible Designs with:
$J = 3$ Analyses
Significance Level $= 0.025$
Power $= 0.975$ at Design Alternative $\theta_A = 0.46$



Group Sequential Trials: Sufficient Statistic

When a group sequential trial is stopped, the sufficient statistic is (M, S_M) (or $(M, \hat{\theta}_M)$) where

- M is analysis time at which trial stops, $M \in \{1, 2, \ldots, J\}$; $M = j$ if the trial stops at the jth analysis.
- S_M is the observed partial sum/partial sum difference when the trial stops.
- $\hat{\theta}_M$ is the observed MLE when the trial stops.

This statistic (M, S_M) or $(M, \hat{\theta}_M)$ may be abbreviated as (M, S) or $(M, \hat{\theta})$.

Possible Designs

Fixed Sample

- Gather $n = 290$ subjects randomized at 1:1 ratio to treatments A and B ($n_A = n_B = 145$)
- Measure outcomes X_{Ai} or X_{Bi} for each subject.
- Compute two-sample z-statistic:

$$z(\theta_0) = \frac{\overline{X}_A - \overline{X}_B - \theta_0}{\sqrt{\sigma^2(\frac{1}{n_A} + \frac{1}{n_B})}} = \frac{\overline{X}_A - \overline{X}_B}{\sqrt{4\sigma^2/n}} = \frac{\sqrt{n}}{2}\left(\overline{X}_A - \overline{X}_B\right)$$

- Reject H_0 if $z(\theta_0) > z_\alpha = \Phi^{-1}(1-\alpha) = 1.96$
- Equivalently, reject H_0 if $(\overline{X}_A - \overline{X}_B) > \frac{1.96}{\sqrt{n/2}} = 0.2298$

Significance Level $= 0.025$

Power $= 0.975$ at Design Alternative $\theta_A = 0.46$

Possible Designs

Fixed Sample Design

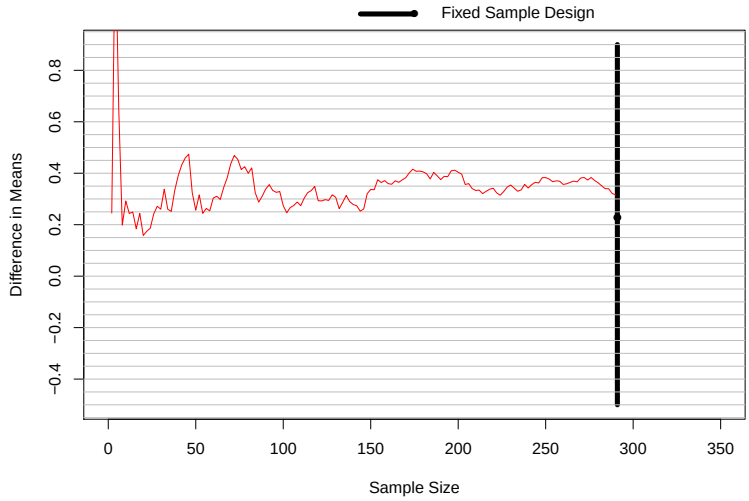
Number of Analyses: $J = 1$

Test Statistic: $T_j = \hat{\theta}_j =$ Sample Mean

j	n_j	a_j	b_j	c_j	d_j
1	290	0.2298	0.2298	0.2298	0.2298

Possible Designs

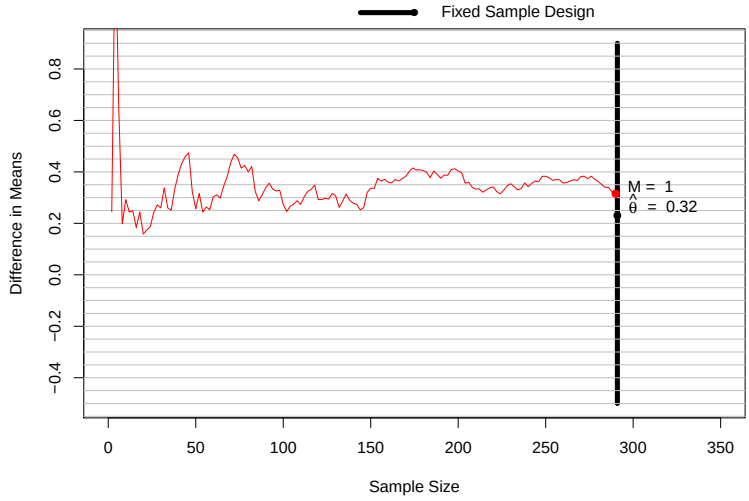
Fixed Sample Design



Foundations: Group Sequential Designs
Inference following Group Sequential Designs
Adaptive Sequential Designs
Clinical Trial Design
Example Setting
Design Comparison

Possible Designs

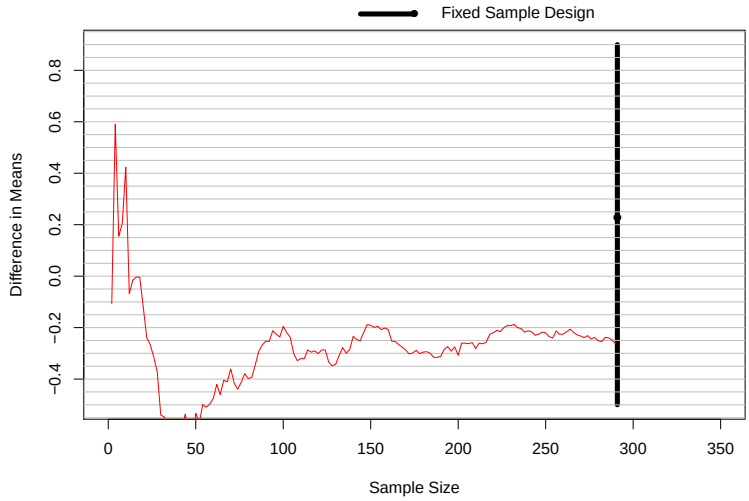
Fixed Sample Design



Sarah Emerson and Scott Emerson
Adaptive Designs
30 / 143

Possible Designs

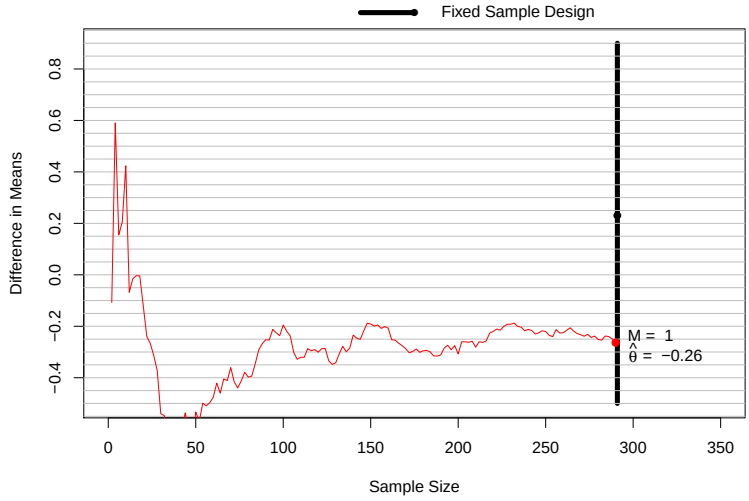
Fixed Sample Design



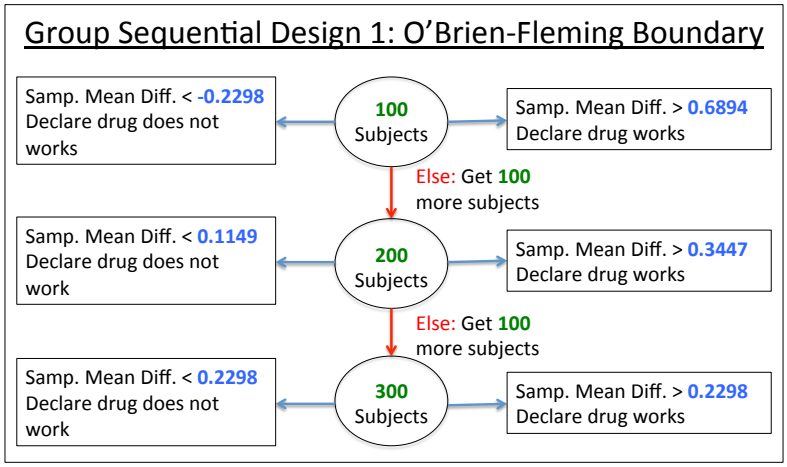
Foundations: Group Sequential Designs
Inference following Group Sequential Designs
Adaptive Sequential Designs
Clinical Trial Design
Example Setting
Design Comparison

Possible Designs

Fixed Sample Design



Possible Designs



Significance Level $= 0.025$

Power $= 0.975$ at Design Alternative $\theta_A = 0.46$

Possible Designs

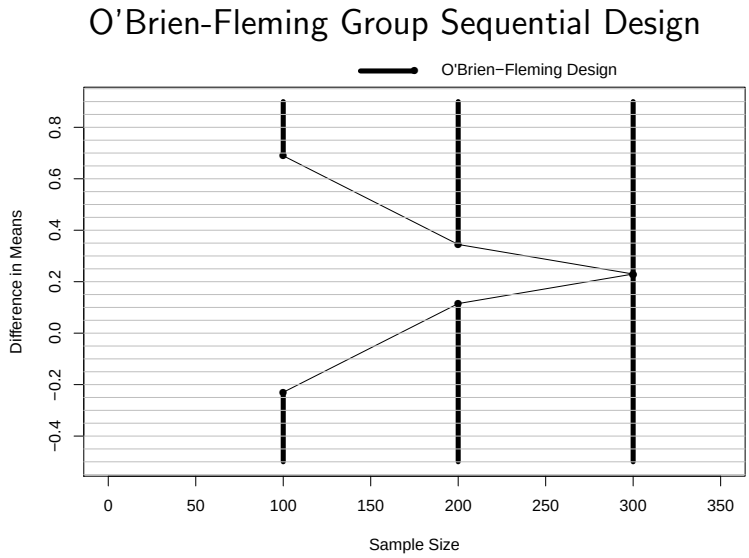
O'Brien-Fleming Group Sequential Design

Number of Analyses: $J = 3$

Test Statistic: $T_j = \hat{\theta}_j =$ Sample Mean

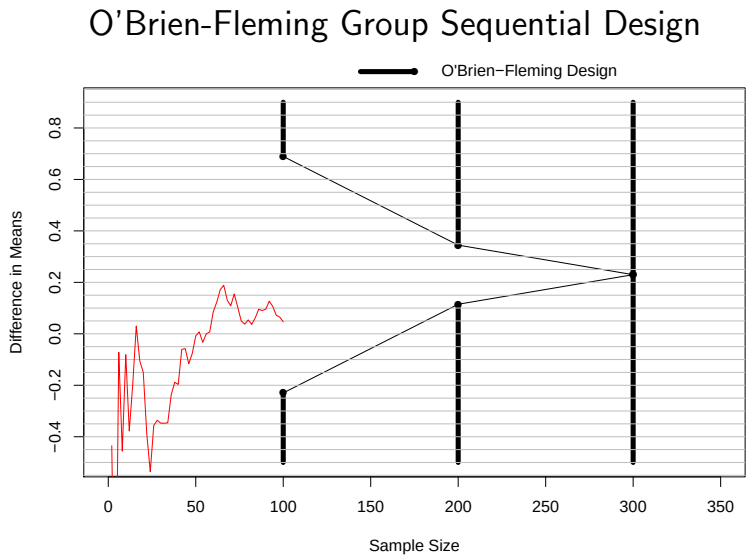
j	n_j	a_j	b_j	c_j	d_j
1	100	-0.2298	0.2298	0.2298	0.6894
2	200	0.1149	0.2298	0.2298	0.3447
3	300	0.2298	0.2298	0.2298	0.2298

Possible Designs



Foundations: Group Sequential Designs
Inference following Group Sequential Designs
Adaptive Sequential Designs
Clinical Trial Design
Example Setting
Design Comparison

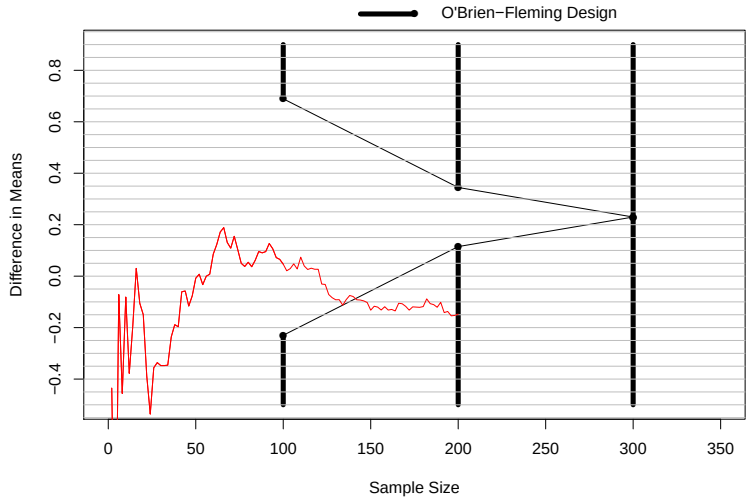
Possible Designs



Sarah Emerson and Scott Emerson
Adaptive Designs
33 / 143

Possible Designs

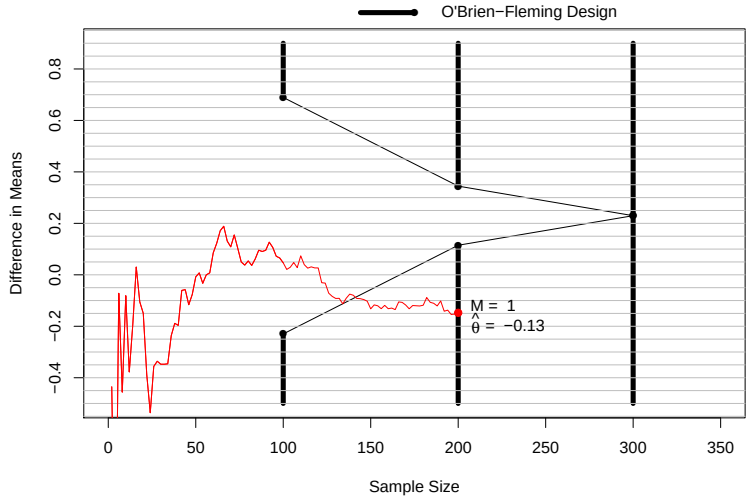
O'Brien-Fleming Group Sequential Design



Foundations: Group Sequential Designs
Inference following Group Sequential Designs
Adaptive Sequential Designs
Clinical Trial Design
Example Setting
Design Comparison

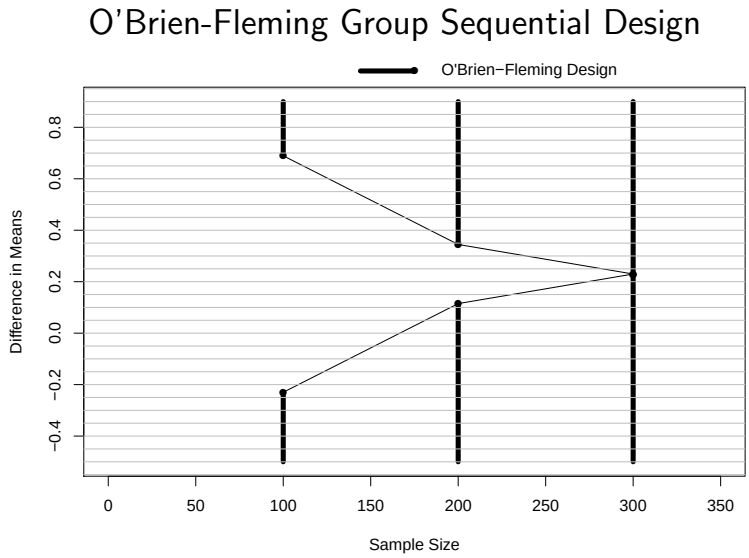
Possible Designs

O'Brien-Fleming Group Sequential Design



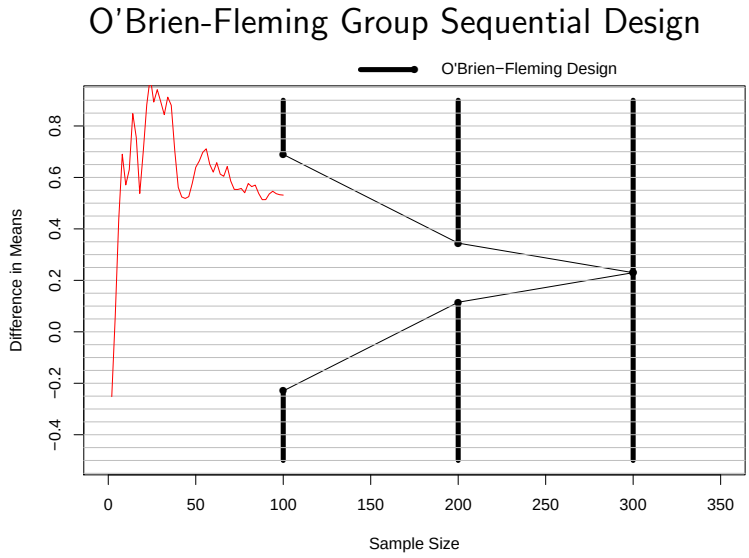
Sarah Emerson and Scott Emerson
Adaptive Designs
33 / 143

Possible Designs



Foundations: Group Sequential Designs
Inference following Group Sequential Designs
Adaptive Sequential Designs
Clinical Trial Design
Example Setting
Design Comparison

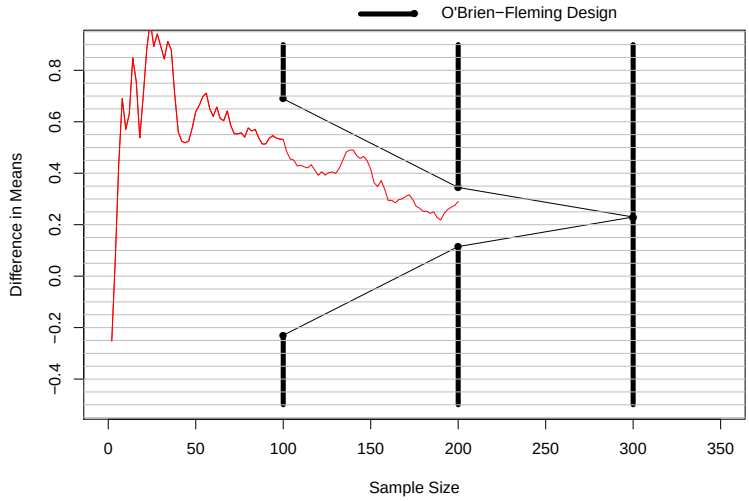
Possible Designs



Sarah Emerson and Scott Emerson
Adaptive Designs
33 / 143

Possible Designs

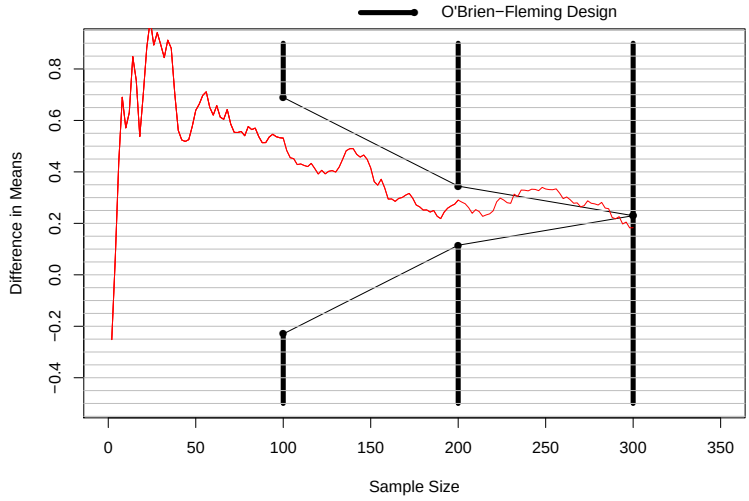
O'Brien-Fleming Group Sequential Design



Foundations: Group Sequential Designs
Inference following Group Sequential Designs
Adaptive Sequential Designs
Clinical Trial Design
Example Setting
Design Comparison

Possible Designs

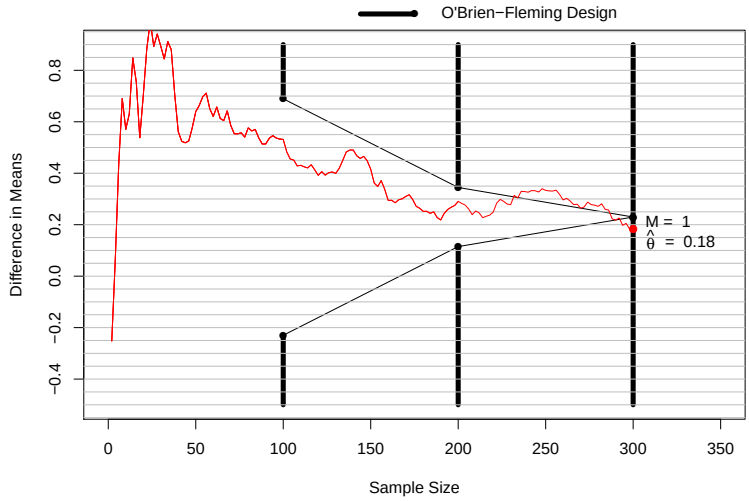
O'Brien-Fleming Group Sequential Design



Sarah Emerson and Scott Emerson
Adaptive Designs
33 / 143

Possible Designs

O'Brien-Fleming Group Sequential Design



Possible Designs

Group Sequential Design 2: Pocock Boundary

Samp. Mean Diff. < **0.0616**
Declare drug does not works

128 Subjects

Samp. Mean Diff. > **0.3980**
Declare drug works

Else: Get **128** more subjects

Samp. Mean Diff. < **0.1782**
Declare drug does not work

256 Subjects

Samp. Mean Diff. > **0.2814**
Declare drug works

Else: Get **128** more subjects

Samp. Mean Diff. < **0.2298**
Declare drug does not work

384 Subjects

Samp. Mean Diff. > **0.2298**
Declare drug works

Significance Level $= 0.025$
Power $= 0.975$ at Design Alternative $\theta_A = 0.46$

Possible Designs

Pocock Group Sequential Design

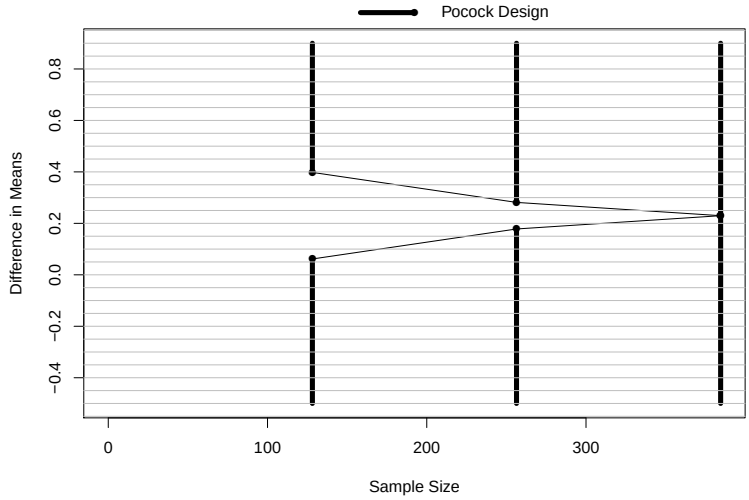
Number of Analyses: $J = 3$

Test Statistic: $T_j = \hat{\theta}_j =$ Sample Mean

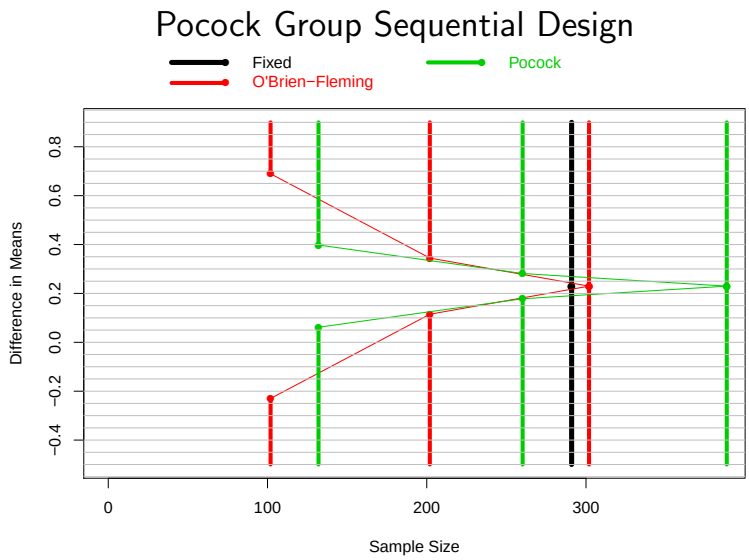
j	n_j	a_j	b_j	c_j	d_j
1	128	0.0616	0.2298	0.2298	0.3980
2	256	0.1782	0.2298	0.2298	0.2814
3	384	0.2298	0.2298	0.2298	0.2298

Possible Designs

Pocock Group Sequential Design



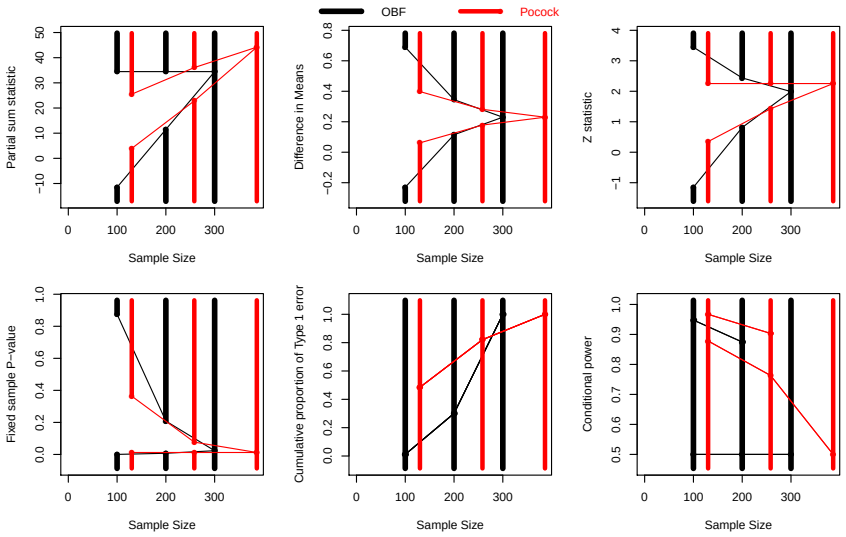
Possible Designs



Boundary Scales

- The same stopping boundary can be represented on many different test-statistic scales, including partial sum difference, sample mean, z-statistic, p-value, etc.

Boundary Scales



1. Foundations: Group Sequential Designs
   - Clinical Trial Design
   - Example Setting
   - Design Comparison
2. Inference following Group Sequential Designs
   - Inference Goals
   - Inference Approaches
   - Inference Optimality Criteria
3. Adaptive Sequential Designs
   - Forms of Adaptation Considered
   - Considerations in Adapting Future Sampling Path
   - Types of Adaptation Rules
   - Adaptive Designs using Standard Group Sequential Software

Clinical Trial Optimality Criteria: Comparing Sequential Designs

- In comparing different types of sequential designs, we must select criteria that we wish to constrain or optimize. Possibilities include:
  - ⋆ Maximal possible sample size n_J
  - For range of parameter values θ:
    - ⋆ Power P_θ(Reject H_0)
    - ⋆ Average sample size (ASN)
    - ⋆ Probability of using more than q subjects
    - ⋆ Median sample size (or any other quantile)
- All but the first of these criteria require knowledge of the sampling distribution of the sufficient statistic (M, S_M) given a parameter value θ.

Clinical Trial Optimality Criteria

Maximal Sample Size

- The maximal sample size for a sequential design is just n_J: the largest sample size at which an analysis may possibly be performed.

Fixed	O'Brien-Fleming	Pocock
291	300	384

Sequential Trial Sampling Density

- For the remaining optimality criteria, the sampling density for the statistic (M, S) is required.
- We will use S_j as the test statistic T_j for ease of discussion; recall that the stopping boundary can be equivalently expressed on many scales.
- For simplicity we will assume $n^*_{Aj} = n^*_{Bj} = \frac{1}{2} n^*_j$ for all j. Analogous formulae for different randomization ratios may be extended from this case.
- We use the fact that

$$S_j = S^*_1 + S^*_2 + \ldots + S^*_j \quad \text{and} \quad S^*_j \sim N\left(\frac{n^*_j}{2}\theta, n^*_j \sigma^2\right)$$

Sequential Trial Sampling Density

To obtain the sampling density at an observed value $(M = j, S = s)$, we have to consider the possible paths that could reach this point.

- If $j = 1$:
  - The test statistic S_1 must have been in the stopping region $\mathcal{S}_j \Leftrightarrow S_1 \notin \mathcal{C}_1$.
  - The value of the test statistic S_1 is $S_1 = s$.
- If $j > 1$:
  - At all analyses $\ell = 1, 2, \ldots, j-1$, the test statistic S_ℓ must have been in the continuation region $\mathcal{C}_\ell$
  - At analysis j the test statistic S_j must have been in the stopping region $\mathcal{S}_j \Leftrightarrow S_j \notin \mathcal{C}_j$.
  - The value of the test statistic S_j is $S_j = s$.

Sequential Trial Sampling Density

Following Armitage et al. (1969), the density of $(M = j, S = s)$ is

$$p_{M,S}(j, s;\, \theta) = \begin{cases} f_{M,S}(j, s;\, \theta) & \text{if } s \in \mathcal{S}_j \\ 0 & \text{otherwise} \end{cases}$$

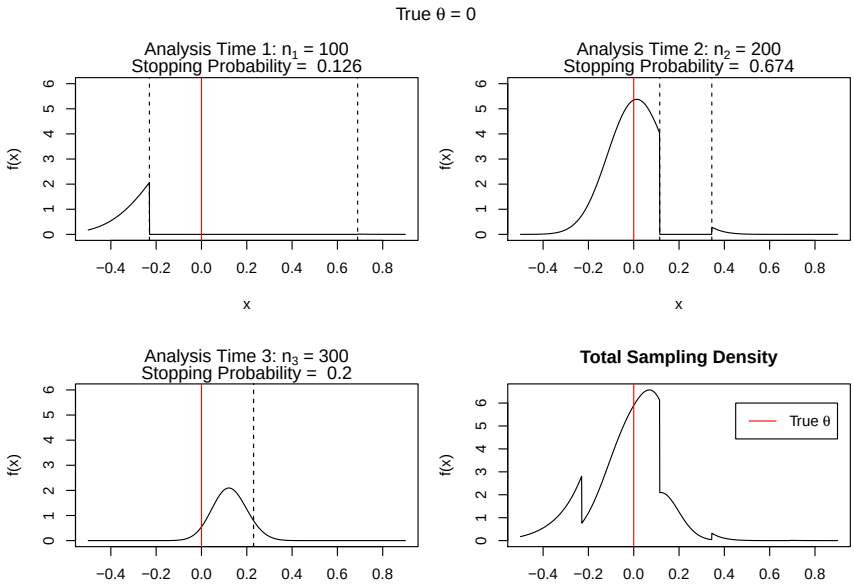
where the (sub)density $f_{M,S}(j,\, s;\, \theta)$ is recursively defined as

$$f_{M,S}(1, s;\, \theta) \;=\; \frac{1}{\sigma\sqrt{n_1}}\, \phi\left(\frac{s - n_1\,\theta/2}{\sigma\sqrt{n_1}}\right)$$

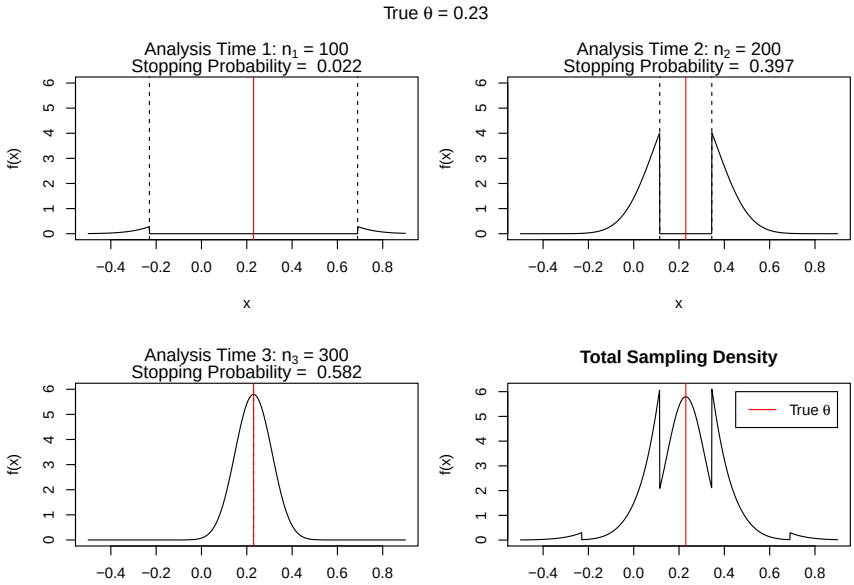
$$f_{M,S}(j, s;\, \theta) \;=\; \int_{\mathcal{C}_{j-1}} \frac{1}{\sigma\sqrt{n_j^*}}\, \phi\left(\frac{s - u - n_j^*\,\theta/2}{\sigma\sqrt{n_j^*}}\right)\, f_{M,S}(j-1,\, u,\, ;\theta)\, du$$

for $j = 2, \ldots, J$

Sequential Trial Sampling Density

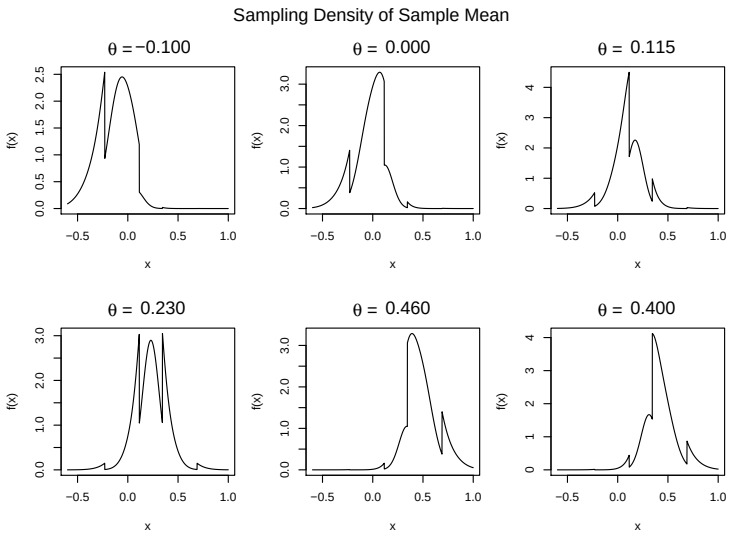


Sequential Trial Sampling Density



Foundations: Group Sequential Designs
Inference following Group Sequential Designs
Adaptive Sequential Designs
Clinical Trial Design
Example Setting
Design Comparison

Sequential Trial Sampling Density



Sarah Emerson and Scott Emerson
Adaptive Designs
45 / 143

Sequential Trial Stopping Probabilities

- Using the density $p_{M,S}(j, s; \theta)$, analysis time stopping probabilities may be obtained as

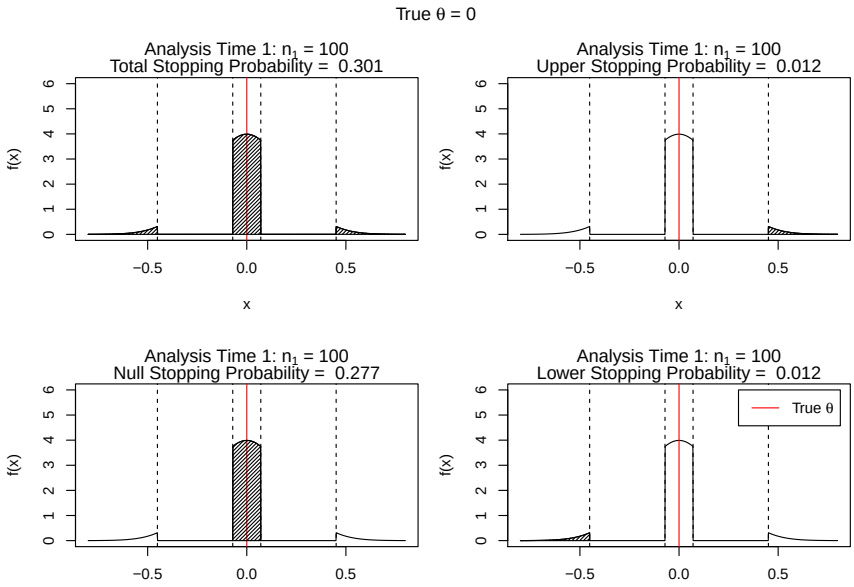
$$\text{Total:} \quad P_\theta(M = j \text{ Total}) = \int_{\mathcal{S}_j} p_{M,S}(j, u)\, du$$

$$\text{Upper:} \quad P_\theta(M = j,\ \text{Upper}) = \int_{u \geq d_j} p_{M,S}(j, u)\, du$$

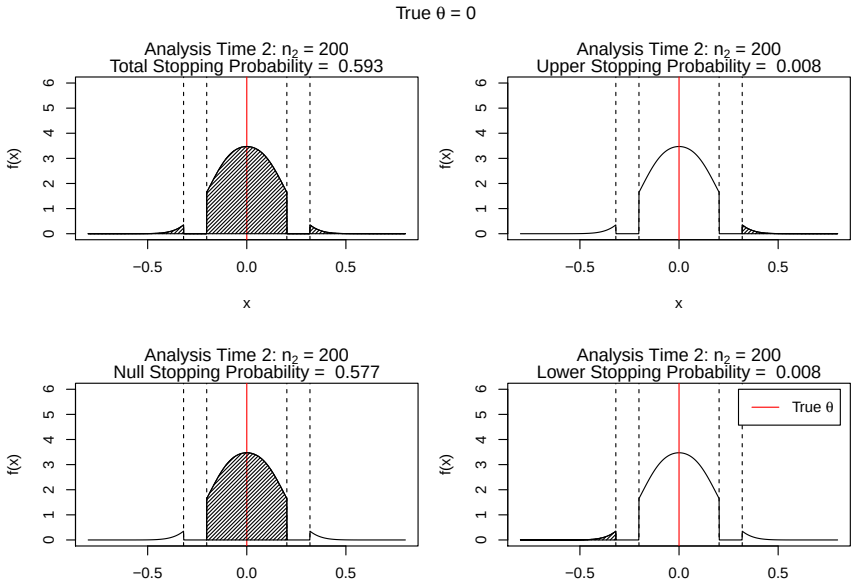
$$\text{Null:} \quad P_\theta(M = j,\ \text{Null}) = \int_{b_j \leq u \leq c_j} p_{M,S}(j, u)\, du$$

$$\text{Lower:} \quad P_\theta(M = j,\ \text{Lower}) = \int_{u \leq a_j} p_{M,S}(j, u)\, du$$

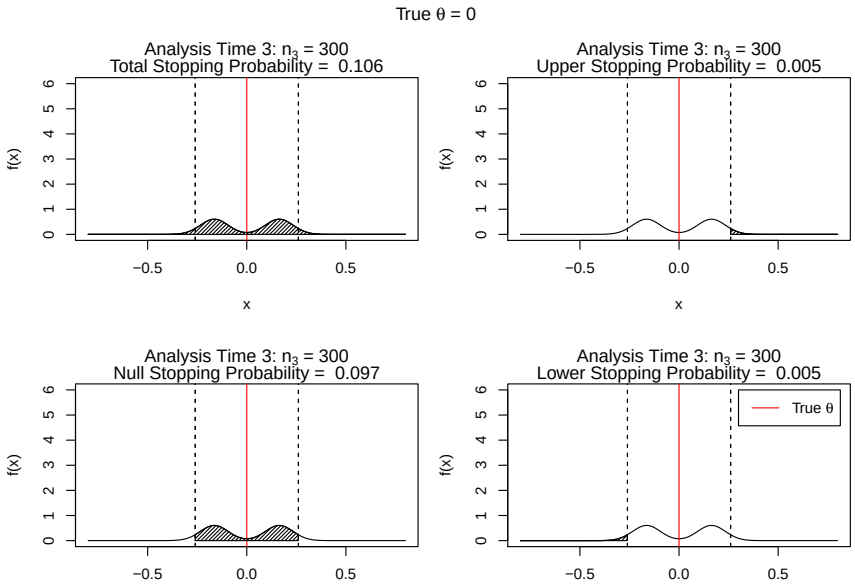
Sequential Trial Stopping Probabilities



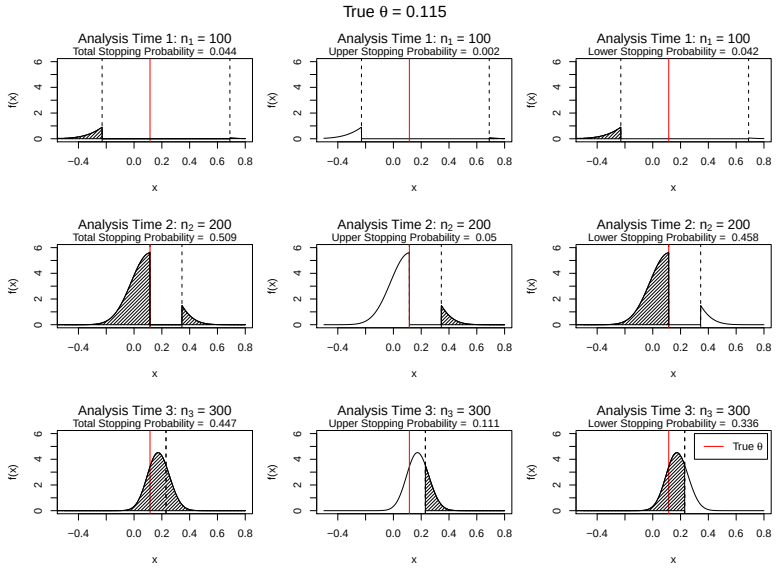
Sequential Trial Stopping Probabilities



Sequential Trial Stopping Probabilities

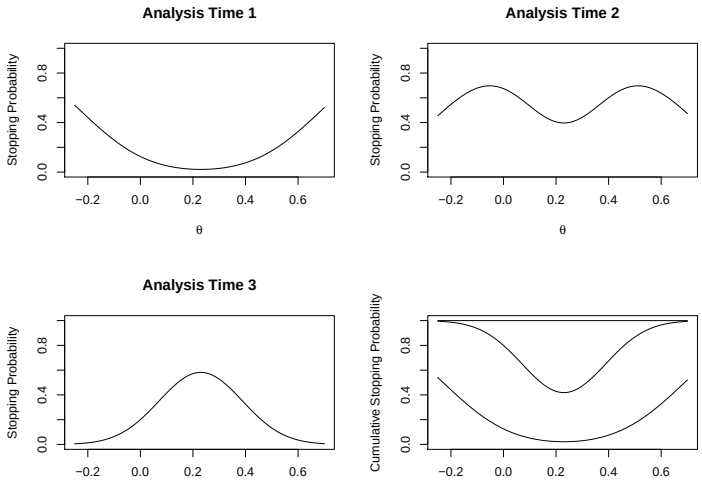


Sequential Trial Stopping Probabilities



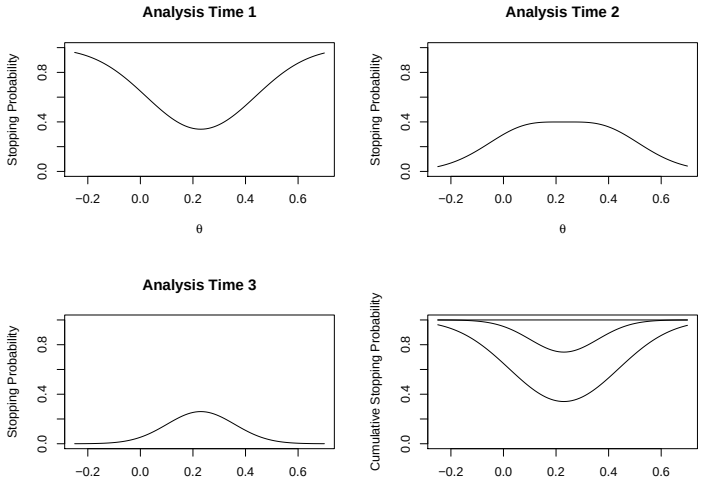
Sequential Trial Stopping Probabilities

Analysis Time Stopping Probabilities: O'Brien-Fleming Group Sequential Design



Sequential Trial Stopping Probabilities

Analysis Time Stopping Probabilities:
Pocock Group Sequential Design



Clinical Trial Optimality Criteria

Power

- Here we consider the upper power for a one-sided test of a greater alternative.
- Using the total analysis time stopping probabilities $P_\theta(M = j,\ \text{Total})$, the power may be obtained as

$$\text{Power}(\theta) = 1 - \beta(\theta) = \sum_{j=1}^{J} P_\theta(M = j,\ \text{Upper})$$

Clinical Trial Optimality Criteria

Power

- For example, under the design alternative $\theta = 0.4596$, the O'Brien-Fleming design has the following upper stopping probabilities:

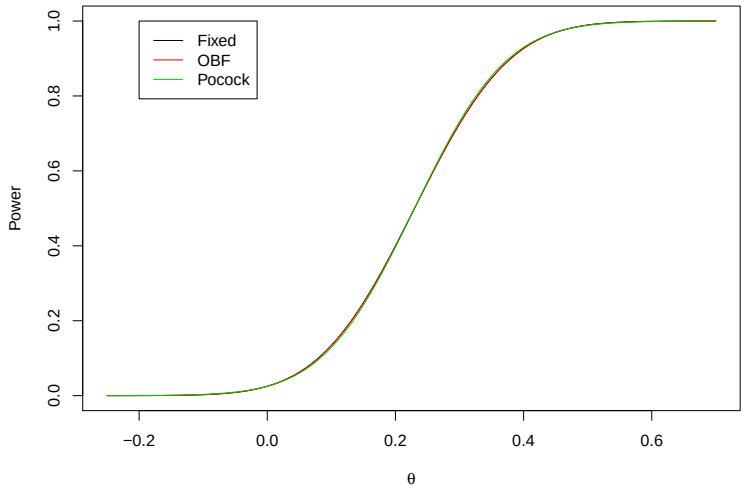
j	N_j	$P_{\theta=0.4596}(M = j,\ \text{Upper})$
1	100	0.1253
2	200	0.6670
3	300	0.1827

The power when $\theta = 0.4596$ is therefore

$$\begin{aligned}\text{Power}(\theta = 0.4596) &= 0.1253 + 0.6670 + 0.1827 \\ &= 0.975\end{aligned}$$

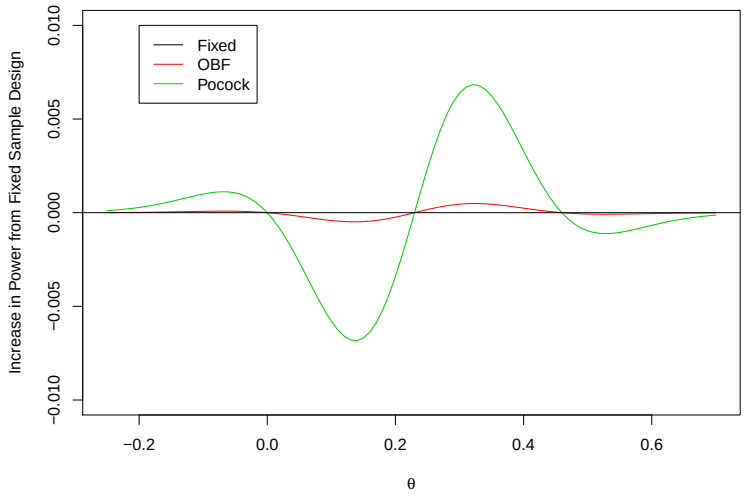
Clinical Trial Optimality Criteria

Power



Clinical Trial Optimality Criteria

Power



Clinical Trial Optimality Criteria

Average Sample Size (ASN)

- Using the total analysis time stopping probabilities $P_\theta(M = j, \text{ Total})$, the average sample size may be obtained as

$$\text{ASN}(\theta) = \sum_{j=1}^{J} P_\theta(M = j, \text{ Total}) n_j$$

Clinical Trial Optimality Criteria

Average Sample Size (ASN)

- For example, under the null hypothesis $\theta = 0$ the O'Brien-Fleming design has the following total stopping probabilities:

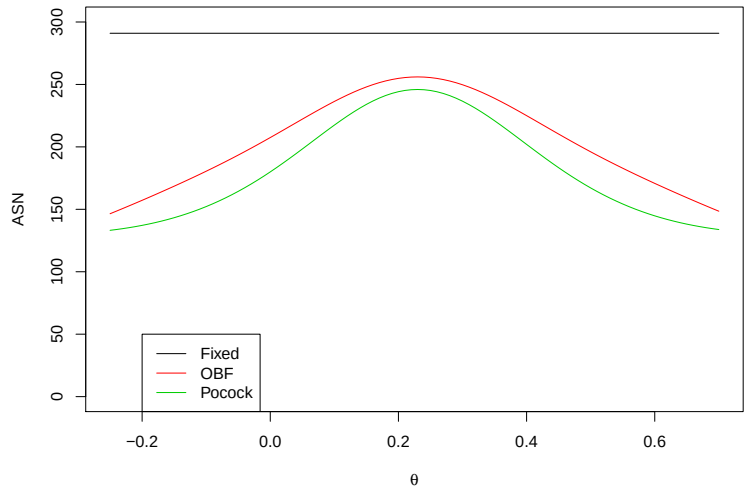
j	n_j	$P_{\theta=0}(M = j, \text{ Total})$
1	100	0.1256
2	200	0.6742
3	300	0.2002

The average sample size (ASN) when $\theta = 0$ is therefore

$$\begin{aligned} \text{ASN}(\theta = 0) &= 100(0.1256) + 200(0.6742) + 300(0.2002) \\ &= 207.4663 \end{aligned}$$

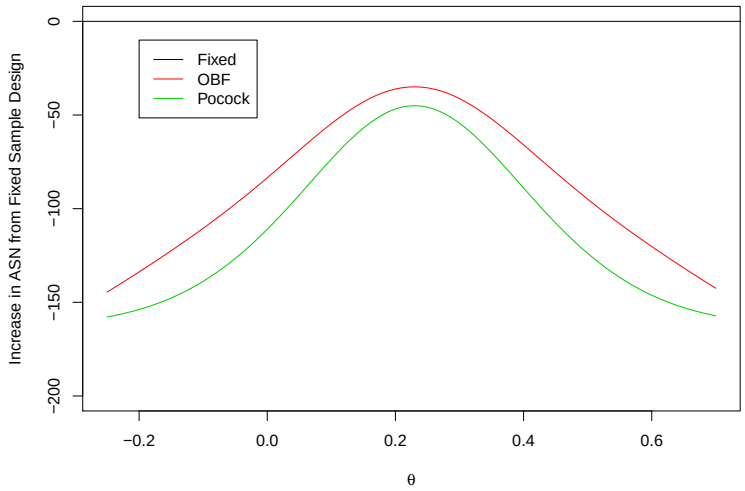
Clinical Trial Optimality Criteria

Average Sample Size (ASN)



Clinical Trial Optimality Criteria

Average Sample Size (ASN)



Clinical Trial Optimality Criteria

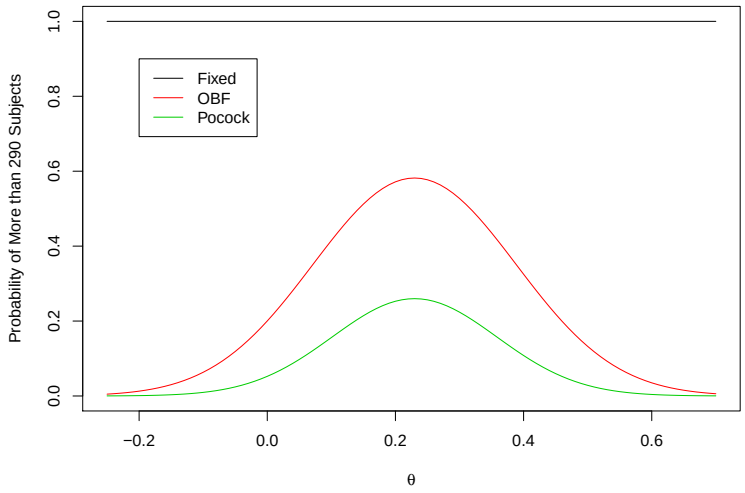
Probability of More Than q Subjects

- Using the total analysis time stopping probabilities $P_\theta(M = j, \text{ Total})$, the probability of using more than q subjects may be obtained as

$$P_\theta(n_M \geq q) = \sum_{j:n_j > q} P_\theta(M = j, \text{ Total})$$

Clinical Trial Optimality Criteria

Probability of More Than 290 Subjects



Clinical Trial Optimality Criteria

Percentile of Sample Size Distribution

- Using the total analysis time stopping probabilities $P_\theta(M = j,\ \text{Total})$, the pth percentile of the sample size distribution may be obtained as

$$q_p(\theta) = \min\{n_j : \sum_{\ell=1}^{j} P_\theta(M = \ell,\ \text{Total}) \geq p\}$$

1. Foundations: Group Sequential Designs
   - Clinical Trial Design
   - Example Setting
   - Design Comparison
2. Inference following Group Sequential Designs
   - Inference Goals
   - Inference Approaches
   - Inference Optimality Criteria
3. Adaptive Sequential Designs
   - Forms of Adaptation Considered
   - Considerations in Adapting Future Sampling Path
   - Types of Adaptation Rules
   - Adaptive Designs using Standard Group Sequential Software

Group Sequential Design Inference Goals

- Trial designed to decide between $H_0 : \theta = \theta_0$ and $H_A : \theta = \theta_A > \theta_0$ with significance level α and power $1 - \beta$ when $\theta = \theta_A$.
- Decision at end of trial:
  - Reject $H_0 : \theta = \theta_0$
  - Fail to reject $H_0 : \theta = \theta_0$ ('Accept' H_0)
- Almost always want more information than just this binary decision:
  - How large is the effect?
  - How confident are we in the estimated effect?

Group Sequential Design Inference Goals

- Hypothesis Testing: Decide between $H_0 : \theta = \theta_0$ and $H_A : \theta > \theta_0$.
  - Design constructed to test particular value of θ_0 at desired level α, with desired power $1 - \beta$ to detect a particular $\theta_A > \theta_0$.
  - We may want to perform a test of a different null hypothesis at the conclusion of the test.
- Point Estimates: Estimates of θ satisfying various optimality criteria. (How large is the effect?)
- Confidence Intervals: Interval estimates $\mathcal{C}_{1-\alpha}$ of θ satisfying $P_\theta(\theta \in \mathcal{C}_{1-\alpha}) = 1 - \alpha$ (How confident are we in the estimated effect?)

1. Foundations: Group Sequential Designs
   - Clinical Trial Design
   - Example Setting
   - Design Comparison
2. Inference following Group Sequential Designs
   - Inference Goals
   - Inference Approaches
   - Inference Optimality Criteria
3. Adaptive Sequential Designs
   - Forms of Adaptation Considered
   - Considerations in Adapting Future Sampling Path
   - Types of Adaptation Rules
   - Adaptive Designs using Standard Group Sequential Software

Obtaining p-values for tests of general null hypotheses

- First consider fixed sample inference: $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ where μ is unknown but σ^2 is known.
- $H_0 : \mu = 0$ vs. $H_A : \mu = 1$
- Recall the interpretation of a p-value for a fixed sample test of $H_0 : \theta = \theta_0'$ when the observed statistic is $X = x$:

$$p_{\theta_0} = \text{Probability of a more 'extreme' result than } X = x \text{ when } \theta = \theta_0'$$

Obtaining p-values for tests of general null hypotheses

- Need to decide which of the possible sample results (outcomes) at the end of the trial are more 'extreme': More convincing for the alternative/less convincing for the null.
  - Larger values of $\bar{X}_n$ are more convincing for H_A and less convincing for H_0.
  - e.g., $\bar{X}_n = 0.7$ is stronger evidence for the alternative/against the null than $\bar{X}_n = 0.3$.

Obtaining p-values for tests of general null hypotheses

- This concept of a p-value may be used in the group sequential design setting to obtain tests of null hypotheses other than the design null hypothesis θ_0.
- Ordering of the sample space (outcome space): Define an ordering or partial ordering of all possible outcomes (M, S) to specify which results will be considered more extreme under the null/stronger evidence for the alternative.
- Unlike in fixed sample inference (at least in normal setting), no obvious ordering exists since sufficient statistic is bivariate (outcome space is 2-dimensional)

Group Sequential Design Inference Approaches

- Suppose, as in earlier example, testing $H_0 : \theta = \theta_0 = 0$ vs. $H_A : \theta = \theta_A = 0.4596$.
- Recall O'Brien-Fleming design with $\alpha = 0.025$, power $1 - \beta = 0.975$:

Number of Analyses: $J = 3$

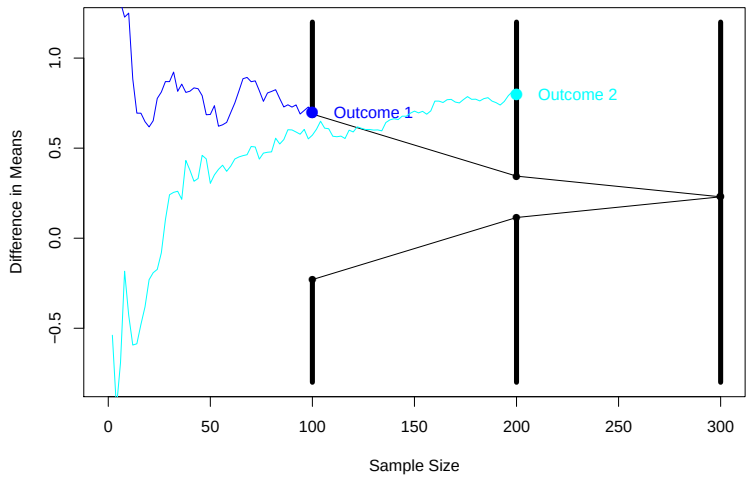
Test Statistic: $T_j = \hat{\theta}_j =$ Sample Mean

j	n_j	a_j	b_j	c_j	d_j
1	100	-0.2298	0.2298	0.2298	0.6894
2	200	0.1149	0.2298	0.2298	0.3447
3	300	0.2298	0.2298	0.2298	0.2298

Group Sequential Design Inference Approaches

- Consider two possible outcomes:
  - Outcome 1: $(M_1 = 1, \hat{\theta}_1 = 0.7)$
  - Outcome 2: $(M_2 = 2, \hat{\theta}_2 = 0.8)$
- Which of these outcomes would you consider stronger evidence for the alternative, $\theta = 0.45$/weaker evidence for the null $\theta = 0$?

Group Sequential Design Inference Approaches



Sample Mean Ordering

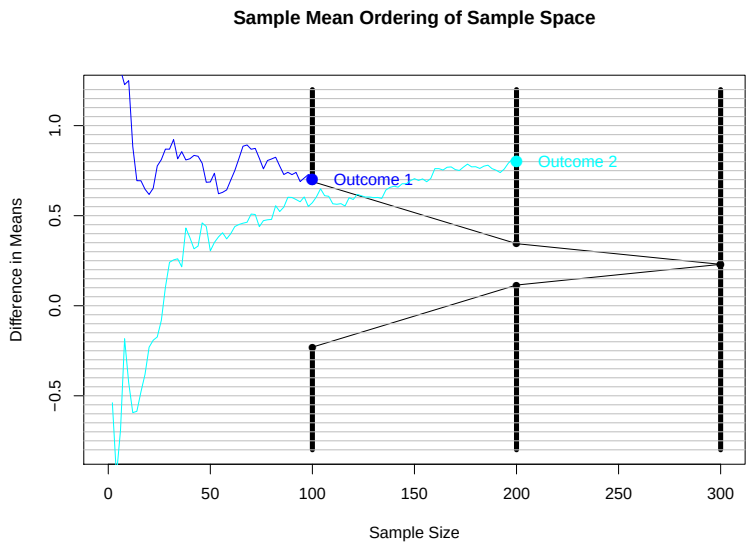
Sample Mean Ordering:

- Outcomes are ordered according to the value of the MLE $\hat{\theta}_M = \hat{\theta}$.
- Consider two outcomes
  - Outcome 1: $(M = j_1, \hat{\theta} = t_1)$
  - Outcome 2: $(M = j_2, \hat{\theta} = t_2)$

Outcome 1 would be considered more extreme under the Sample Mean ordering as follows:

$$(j_1, t_1) \succ_{SM} (j_2, t_2) \text{ if } t_1 > t_2$$

Sample Mean Ordering



Analysis Time Ordering

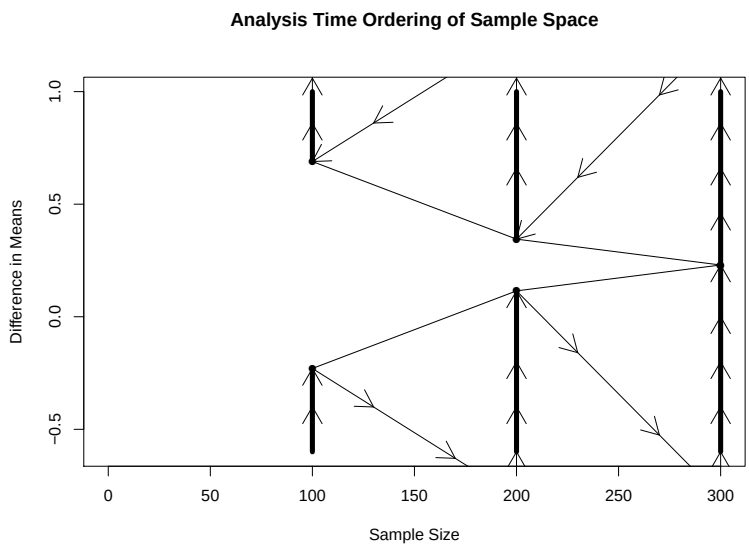
Analysis Time Ordering:

- Outcomes are ordered according to
  1. Stopping time M
  2. MLE $\hat{\theta}$
- Consider two outcomes:
  - Outcome 1: $(M = j_1, \hat{\theta} = t_1)$
  - Outcome 2: $(M = j_2, \hat{\theta} = t_2)$

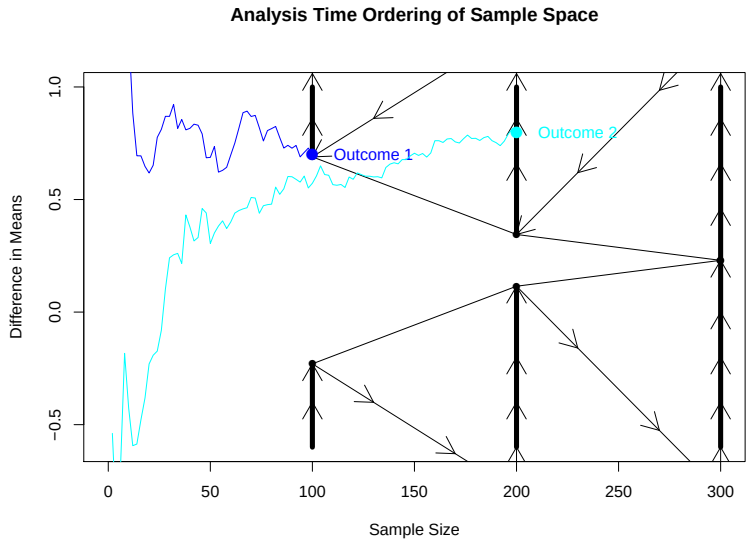
Outcome 1 would be considered more extreme under the Analysis Time ordering as follows:

$$(j_1, t_1) \succ_{AT} (j_2, t_2) \text{ if } \begin{cases} j_1 < j_2 & \text{and } t_1 \geq d_{j_1} \\ j_1 > j_2 & \text{and } t_2 \leq a_{j_2} \\ j_1 = j_2 & \text{and } t_1 > t_2 \end{cases}$$

Analysis Time Ordering



Analysis Time Ordering



Likelihood Ratio Ordering

(Signed) Likelihood Ratio Ordering:

- Outcomes are ordered according to signed likelihood ratio test statistic for hypothesized θ_0'
- Consider two outcomes:
  - Outcome 1: $(M = j_1, \hat{\theta} = t_1)$
  - Outcome 2: $(M = j_2, \hat{\theta} = t_2)$

Outcome 1 would be considered more extreme under the Likelihood Ratio ordering as follows:

$$(j_1, t_1) \succ_{AT} (j_2, t_2) \text{ if}$$

$$\text{sign}(t_1 - \theta_0') \frac{p_{M,T}(j_1, t_1;\, \theta = t_1)}{p_{M,T}(j_1, t_1;\, \theta = \theta_0')} > \text{sign}(t_2 - \theta_0') \frac{p_{M,T}(j_2, t_2;\, \theta = t_2)}{p_{M,T}(j_2, t_2;\, \theta = \theta_0')},$$

$$\text{i.e., if} \quad \sqrt{n_{j_2}}(t_2 - \theta_0') > \sqrt{n_{j_1}}(t_1 - \theta_0')$$

Outcome Space Orderings

- For both the Sample Mean ordering and the Analysis Time ordering, the ordering does not depend upon the null hypothesis being tested.
- In contrast, note that the Likelihood Ratio ordering depends on the value of θ_0' being tested, and therefore may order the outcome space differently for different θ_0' values.

Outcome Space Orderings

- It can be shown that the Sample Mean and Analysis Time orderings produce *stochastically ordered* distributions of the outcomes under the proposed ordering:

$$P_\theta\left((M,\hat{\theta}) \succ (j,t)\right) \text{ is an increasing function of } \theta$$

  for both $\succ_{SM}$ and $\succ_{AT}$ orderings.
- In contrast, stochastic ordering has not been proven for the Likelihood Ratio ordering.

Confidence Intervals from p-values

- Construct one-sided p-values $p_1(\theta_0)$ for test of $H_0 : \theta = \theta_0$ vs. $H_A : \theta > \theta_0$ using chosen ordering of sample space.
- Obtain two-sided p-values as $p(\theta_0) = 2 * \min(p_1(\theta_0), 1 - p_1(\theta_0))$
- Construct confidence intervals using hypothesis test/confidence interval duality:

$$\mathcal{C} = \{\theta_0 : p(\theta_0) > \alpha\}$$

Confidence Intervals from Sample Space Orderings

- Using the O'Brien-Fleming Boundary (Group Sequential Design 1)
- Observe Outcome 1: $(M = 1, \hat{\theta} = 0.7)$

Method	95% CI for θ
Sample Mean Ordering	(0.305, 0.971)
Analysis Time Ordering	(0.308, 1.090)
Likelihood Ratio Ordering	(0.265, 1.02)

Confidence Intervals from Sample Space Orderings

- Using the O'Brien-Fleming Boundary (Group Sequential Design 1)
- Observe Outcome 2: $(M = 2, \hat{\theta} = 0.8)$

Method	95% CI for θ
Sample Mean Ordering	(0.407, 1.170)
Analysis Time Ordering	(0.297, 1.010)
Likelihood Ratio Ordering	(0.515, 1.030)

Alternative Confidence Interval Approach: Repeated Confidence Intervals

Repeated Confidence Intervals (Jennison and Turnbull, 1989):

- Invert a level α two-sided group sequential test at each stage $j = 1, \ldots, J$ to obtain intervals $\mathcal{I}_j$ such that

$$P_\theta \left(\theta \in \mathcal{I}_j \text{ for all } j = 1, \ldots, J \right) = 1 - \alpha$$

- $\mathcal{I}_j$ is the set of all values of θ_0' for which a group sequential test of $H_0 : \theta = \theta_0'$ would not reject at stage j.

Alternative Confidence Interval Approach: Repeated Confidence Intervals

- $\mathcal{I}_j$ can be rephrased in terms of the test statistic $T_j(\theta_0')$ which depends upon the null hypothesis value.
- The group sequential stopping rule can be expressed as

$$\text{Reject } H_0 : \theta = \theta_0' \text{ if } T_j(\theta_0') < a_j \text{ or } T_j(\theta_0') > d_j.$$

- Thus we have

$$\mathcal{I}_j = \left\{\theta_0' : a_j \leq T_j(\theta_0') \leq d_j\right\}$$

Alternative Confidence Interval Approach: Repeated Confidence Intervals

- Consider the normal setting (with no mean variance relationship)
- Let $\{J, n_j, T_j, (a_j, b_j, c_j, d_j) \text{ for } j = 1, \ldots, J\}$ be a level α group sequential test of $H_0 : \theta = 0$ vs. $H_A : \theta \neq 0$
- Consider the boundary scale $T_j = \hat{\theta}_j$
- The interval $\mathcal{I}_j$ at stage j is

$$\mathcal{I}_j = \left\{\theta_0' : a_j \leq \hat{\theta}_j - \theta_0' \leq d_j\right\}$$

- The repeated confidence interval for θ is therefore

$$\left\{\theta_0' : a_j \leq \hat{\theta}_j - \theta_0' \leq d_j \text{ for all } j = 1, \ldots, J\right\}$$

Point Estimation

- Given that we observe an outcome $(M, \hat{\theta}) = (j, t)$, we would like to provide a point estimate for the parameter θ.
- Several options have been proposed:
  - Maximum Likelihood Estimator
  - Bias-Adjusted Mean
  - Median-Unbiased Estimator
  - (And several others)

Point Estimation

- Maximum Likelihood Estimate:

$$\begin{aligned}\hat{\theta}_{\text{MLE}} &= \hat{\theta} \\ &= \bar{X}_A - \bar{X}_B\end{aligned}$$

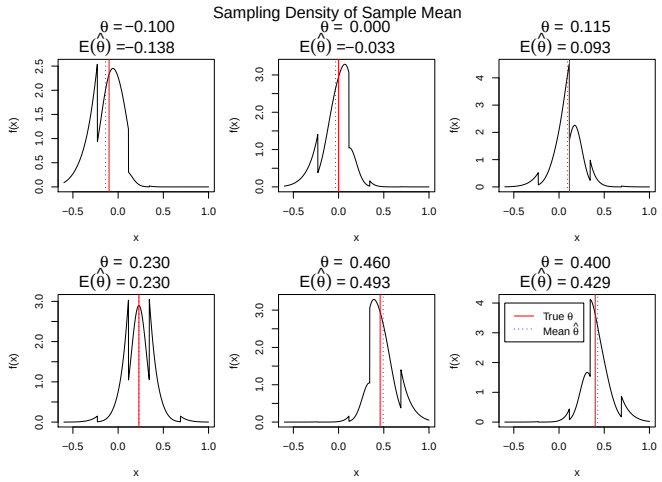
- The MLE is typically a biased estimate of θ:

$$E_\theta[\hat{\theta}] \neq \theta$$

- For example, when $\theta = 0$, the expected value of the difference in sample means when the trial stops is

$$E_{\theta=0}[\hat{\theta}] = -0.033 \neq 0$$

Point Estimation



Point Estimation

- Bias-adjusted Mean: $\hat{\theta}_{\text{BAM}}$ is the value of θ' satisfying

$$\mathrm{E}\left[\hat{\theta}; \theta'\right] = t;$$

that is, the value of the parameter for which the observed statistic is the expected value under that parameter value.

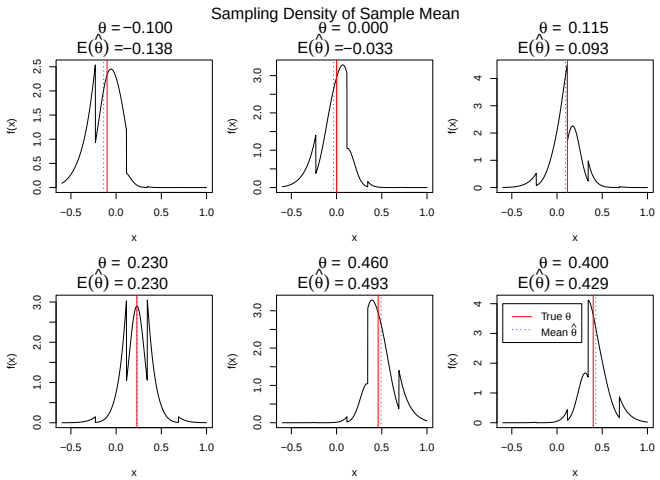
Point Estimation

Example:

- Using the O'Brien-Fleming Boundary (Group Sequential Design 1)
- Suppose the trial stops with $(M, \hat{\theta}) = (2, 0.093)$. Then the BAM is found by searching for θ' such that

$$\mathrm{E}\left[\hat{\theta}; \theta'\right] = 0.093$$

Point Estimation



Point Estimation

Example:

- Using the O'Brien-Fleming Boundary (Group Sequential Design 1)
- Suppose the trial stops with $(M, \hat{\theta}) = (2, 0.093)$. Then the BAM is found by searching for θ' such that

$$\mathrm{E}\left[\hat{\theta}; \theta'\right] = 0.093$$

- We see from the previous slide that when $\theta = 0.115$,

$$\mathrm{E}\left[\hat{\theta}; \theta = 0.115\right] = 0.093.$$

Therefore, the BAM when $(M, \hat{\theta}) = (2, 0.093)$ is

$$\hat{\theta}_{\mathrm{BAM}} = 0.115$$

Point Estimation

- Median-unbiased Mean: $\hat{\theta}_{\mathrm{MUE}}$ is the value of θ' satisfying

$$P\left((M, S) \succ (m, s); \theta'\right) = 0.5;$$

that is, the value of the parameter for which the observed statistic would be the median of the sampling distribution under that parameter value.

- Note that this estimator depends on the ordering of the outcome space.

Point Estimation

- Observe Outcome 1: $(M = 1, \hat{\theta} = 0.7)$

Method	Estimate of θ
MLE	0.700
BAM	0.659
MUE (SM)	0.653
MUE (AT)	0.700
MUE (LR)	0.644

Point Estimation

- Observe Outcome 2: $(M = 2, \hat{\theta} = 0.8)$

Method	Estimate of θ
MLE	0.800
BAM	0.762
MUE (SM)	0.780
MUE (AT)	0.679
MUE (LR)	0.786

1. Foundations: Group Sequential Designs
   - Clinical Trial Design
   - Example Setting
   - Design Comparison
2. Inference following Group Sequential Designs
   - Inference Goals
   - Inference Approaches
   - Inference Optimality Criteria
3. Adaptive Sequential Designs
   - Forms of Adaptation Considered
   - Considerations in Adapting Future Sampling Path
   - Types of Adaptation Rules
   - Adaptive Designs using Standard Group Sequential Software

Inference Optimality Criteria: Confidence Intervals

- How should we decide which method of CI construction is better?

Inference Optimality Criteria: Confidence Intervals

- How should we decide which method of CI construction is better?
  - Coverage probability close to nominal level $1 - \alpha$
  - Confidence interval width: narrow intervals preferred, more efficient
  - Convexity: Does the method produce a true interval?
  - Agreement with design hypothesis test decision
  - Agreement with a reasonable point estimate

Confidence Interval Optimality Criteria

Coverage Probability:

- Let $\mathcal{C}_{1-\alpha}$ be a nominal $(1-\alpha)100\%$ confidence interval for the parameter θ.
- Recall that the confidence interval is random: when the experiment is repeated, we will obtain different limits for the interval.
- The coverage probability is $P(\theta \in \mathcal{C}_{1-\alpha})$.
- This should be the target level $(1-\alpha)100\%$ by construction, but it is important to assess whether that is actually being achieved.

Confidence Interval Optimality Criteria

Interval Width:

- Methods that produce shorter/narrower intervals with the same coverage probability are preferred: more precision about the estimate of θ.
- In fixed sample setting with known variance, the width is constant for a given sample size.
- In contrast, in the group sequential setting interval width is random and its distribution depends upon the true value of θ.
- Interval length may be compared on basis of
  - Average width
  - Median/other quantile of width
  - Probability that the width exceeds some given size

Confidence Interval Optimality Criteria

Convexity:

- Are the confidence regions true intervals?
- If $\theta_1 \in \mathcal{C}$ and $\theta_2 \in \mathcal{C}$, then we would like to have all parameter values θ^* between θ_1 and θ_2 also in $\mathcal{C}$.
- That is, we want

$$\beta\theta_1 + (1-\beta)\theta_2 \in \mathcal{C}$$

for any $\beta \in (0,1)$.

Confidence Interval Optimality Criteria

Agreement with Decision:

- If the study is stopped for efficacy then we would prefer that θ_0 not be in the $(1-\alpha)100\%$ confidence region, where α is level for which the stopping boundaries were designed.
  - That is, if the design null hypothesis $H_0 : \theta = \theta_0$ is rejected by the level α by the stopping boundary, $\theta_0 \notin \mathcal{C}$.
- If the study is stopped for futility then the design alternative at which the design has power $1-\beta = 1-\alpha$ should not be in the confidence region.
  - That is, if the design null hypothesis $H_0 : \theta = \theta_0$ is accepted by the stopping boundary that has power $1-\alpha$ to detect the alternative $\theta = \theta_A$, then $\theta_A \notin \mathcal{C}$

Confidence Interval Optimality Criteria

Agreement with Point Estimate:

- If $\tilde{\theta}$ is an estimate of the parameter θ, and $\mathcal{C}$ is a confidence interval for θ, it is preferable to have $\tilde{\theta} \in \mathcal{C}$.
- This is an optimality criterion for both the estimate and the interval
- It is more important to have agreement with well-behaved estimators like the Bias-adjusted Mean than with poorer estimators like the MLE.
- Some reasonable confidence intervals may not contain the MLE with non-negligible probability, which is more acceptable due to the bias of the MLE.

Inference Optimality Criteria: Point Estimates

- How should we decide which method of point estimate construction is better/which to use?

Foundations: Group Sequential Designs
Inference following Group Sequential Designs
Adaptive Sequential Designs
Inference Goals
Inference Approaches
Inference Optimality Criteria

Inference Optimality Criteria: Point Estimates

- How should we decide which method of point estimate construction is better/which to use?
  - Bias
  - Mean-squared Error
  - Agreement with reasonable confidence interval
  - Agreement with design hypothesis test decision
  - (Consistency)

Point Estimates Optimality Criteria

Bias:

- Is the expected value of the estimator equal to the true parameter value?
- The bias of an estimator $\tilde{\theta}$ for the parameter θ is

$$B(\theta;\theta) = \mathrm{E}(\tilde{\theta}) - \theta$$

- An estimator $\tilde{\theta}$ is unbiased if

$$E(\tilde{\theta}) = \theta$$

- Low or zero bias is desirable, other properties being equal.

Point Estimates Optimality Criteria

Mean-squared Error:

- What is the expected squared distance between the estimator and the true parameter value?
- The mean-squared error of an estimator $\tilde{\theta}$ for the parameter θ is

$$\mathrm{MSE}(\tilde{\theta};\theta) = \mathrm{E}\left[(\tilde{\theta} - \theta)^2\right]$$

- It can be shown that

$$\mathrm{MSE}(\tilde{\theta};\theta) = \left[B(\tilde{\theta};\theta)\right]^2 + \mathrm{Var}[\tilde{\theta}]$$

- Small mean-squared error is desirable.

Point Estimates Optimality Criteria

Agreement with Confidence Interval:

- If $\tilde{\theta}$ is an estimate of the parameter θ, and $\mathcal{C}$ is a confidence interval for θ, it is preferable to have $\tilde{\theta} \in \mathcal{C}$.
- This is an optimality criterion for both the estimate and the interval
- It is more important to have agreement with well-behaved confidence intervals (i.e. those that are narrower, form true intervals, etc.)

Point Estimates Optimality Criteria

Agreement with Decision:

- Is it possible that the estimate be in the null hypothesis region of the parameter space, but the decision based on the boundary is to reject the null?
- That is, if $\tilde{\theta} \leq \theta_0$ we do not want to reject $H_0 : \theta = \theta_0$ in favor of a greater alternative.

Point Estimates Optimality Criteria

Consistency

- Does the estimator converge (in probability) to the true value as the sample size increases to infinity?
- This property is less emphasized for sequential designs, as we are primarily interested in the sample size for which the study is planned.
- (You may, nevertheless, encounter papers where consistency of estimators in a group sequential design setting is considered.)

1. Foundations: Group Sequential Designs
   - Clinical Trial Design
   - Example Setting
   - Design Comparison
2. Inference following Group Sequential Designs
   - Inference Goals
   - Inference Approaches
   - Inference Optimality Criteria
3. Adaptive Sequential Designs
   - Forms of Adaptation Considered
   - Considerations in Adapting Future Sampling Path
   - Types of Adaptation Rules
   - Adaptive Designs using Standard Group Sequential Software

Types of Design Adaptation

- Many possible ways to adaptively modify future analysis plan at an interim analysis. Examples:
  - Sample size re-estimation
  - Adaptive randomization
  - Dropping inferior treatment groups
  - Change of endpoint
  - Change of hypothesis

Types of Design Adaptation

- The common theme among adaptive designs is the use of an interim effect estimate to adjust the plans for future analyses.
- Here we focus solely on adaptive sample size and stopping boundary modification based on interim effect estimate.
- Note that modifying sample sizes due to updated information on ancillary statistics/information growth is not considered in this setting, and does not require *as* careful attention to protecting Type I error rate.

Adaptive Sample Size and Stopping Boundary Modification

- The type of design adaptation we consider here involves using an interim estimate of effect size to modify the future analyses.
- Modification may affect any or all of the following components of the future analysis plan:
  - Number of future analyses
  - Timing/sample size for future analyses
  - Stopping boundary/critical value(s) for future analyses

1. Foundations: Group Sequential Designs
   - Clinical Trial Design
   - Example Setting
   - Design Comparison
2. Inference following Group Sequential Designs
   - Inference Goals
   - Inference Approaches
   - Inference Optimality Criteria
3. Adaptive Sequential Designs
   - Forms of Adaptation Considered
   - Considerations in Adapting Future Sampling Path
   - Types of Adaptation Rules
   - Adaptive Designs using Standard Group Sequential Software

Why Adapt?

- Proposed Benefits of Adaptive Sample Size and Stopping Boundary modification?

Foundations: Group Sequential Designs
Inference following Group Sequential Designs
Adaptive Sequential Designs
Forms of Adaptation Considered
Considerations in Adapting Future Sampling Path
Types of Adaptation Rules
Adaptive Designs using Standard Group Sequential Software

Why Adapt?

- Proposed Benefits of Adaptive Sample Size and Stopping Boundary modification?
  - Re-power study to detect smaller/larger effect size if interim estimate indicates a value substantially different from design hypotheses
  - Increased flexibility in accrual decisions, justification for sample size
  - Possibly improve efficiency
  - Potential cost reduction, particularly in time-to-event setting

Sarah Emerson and Scott Emerson Adaptive Designs 109 / 143

Considerations in Adapting Sample Size/Stopping Boundary

- If adaptation is performed without careful adjustment of stopping boundary, Type I error can be greatly inflated.
- Proschan and Hunsberger (1995):
  - Two-stage design: n_1 in first stage
  - Interim effect size estimate at first stage used to choose n_2 for second stage
  - Depending upon how n_2 chosen, Type I error probability can more than double: $0.05 \rightarrow 0.1146$
  - Even Bonferroni correction would not fix this inflation of Type I error rate.

1. Foundations: Group Sequential Designs
   - Clinical Trial Design
   - Example Setting
   - Design Comparison
2. Inference following Group Sequential Designs
   - Inference Goals
   - Inference Approaches
   - Inference Optimality Criteria
3. Adaptive Sequential Designs
   - Forms of Adaptation Considered
   - Considerations in Adapting Future Sampling Path
   - Types of Adaptation Rules
   - Adaptive Designs using Standard Group Sequential Software

Some Proposed Adaptation Rules

- Proschan and Hunsberger (conditional error)
- Lehmacher and Wassmer 1999 (reweighted statistic)
- Cui, Hung, Wang 1999 (reweighted statistic)
- Muller and Schafer 2001 (conditional error)
- Brannath, Posch, Bauer 2002 (recursive combination tests/conditional error)
- Gao, Ware, Mehta 2008 (conditional error, sample size guided by conditional power)
- Mehta and Pocock 2010 (conditional error, sample size guided by conditional power)
- More general: Any path of group sequential designs chosen to have correct rejection rate under null

Some Proposed Adaptation Rules

- Adaptation rules in literature can be categorized into three general approaches:
  - Reweighting the test statistic: using the same stopping boundary (critical values) with different sample sizes
  - Conditional error preservation: using possibly different stopping boundary (critical values) and different sample sizes
  - General pre-specified design such that overall type I error rate is controlled
- We will see that these are listed in order of increasing flexibility.

Adaptation Rules: Reweighted Statistic/Combining p-values

- The first type of adaptation rule starts with a group sequential design:
  - $\{J, n_j, T_j, (a_j, b_j, c_j, d_j)$ for $j = 1, \ldots, J\}$
  - Let T_j be either the z-statistic Z_j or the fixed sample p-value P_j.
  - Incremental test statistics Z_j^* and P_j^*, computed only from the data acquired in the jth group.
- At some interim analysis $h\,(1 \leq h < J)$, the future incremental sample sizes may be modified:
  - $n_j^* \rightarrow \tilde{n}_j^*$ for $j = h+1, h+2, \ldots, J$
  - For notational conveniene, we let $\tilde{n}_j^* = n_j^*$ for $j = 1, \ldots, h$.
  - Let $\tilde{T}_j^*$ be the incremental test statistic computed using the new sample size for the jth stage, $\tilde{n}_j^*$.

Adaptation Rules: Reweighted Statistic/Combining p-values

- First consider the normal mean setting with $n_{Aj} = n_{Bj} = \frac{n_j}{2}$, so

$$Z_j = \frac{(\hat{\theta}_j - \theta_0)\sqrt{n_j}}{2\sigma}$$

$$Z_j^* = \frac{(\hat{\theta}_j^* - \theta_0)\sqrt{n_j^*}}{2\sigma}$$

- Note that we can write

$$\hat{\theta}_j = \frac{\frac{n_1^*}{2}\hat{\theta}_1^* + \frac{n_2^*}{2}\hat{\theta}_2^{\,*} + \ldots + \frac{n_j^*}{2}\hat{\theta}_j^*}{\frac{n_1^*}{2} + \frac{n_2^*}{2} + \ldots + \frac{n_j^*}{2}} = \frac{\sum_{\ell=1}^{j} n_\ell^* \hat{\theta}_\ell^*}{\sum_{\ell=1}^{j} n_\ell}$$

Adaptation Rules: Reweighted Statistic/Combining p-values

- Therefore, we can decompose Z_j in terms of the incremental Z_ℓ^* as

$$\begin{aligned} Z_j &= \frac{\sqrt{n_j}}{2\sigma}\left(\frac{\sum_{\ell=1}^{j} n_\ell^* \hat{\theta}_\ell^*}{\sum_{\ell=1}^{j} n_\ell} - \theta_0\right) \\ &= \frac{\sqrt{n_j}}{2\sigma}\left(\frac{\sum_{\ell=1}^{j} n_\ell^* (\hat{\theta}_\ell^* - \theta_0)}{n_j}\right) \\ &= \frac{\sum_{\ell=1}^{j} \sqrt{n_\ell^*} Z_\ell^*}{\sqrt{n_j^*}} \end{aligned}$$

Adaptation Rules: Reweighted Statistic/Combining p-values

- If the incremental group sizes are modified, we still have

$$\tilde{Z}_j^* = \frac{(\tilde{\hat{\theta}}_j^* - \theta_0)\sqrt{\tilde{n}_j^*}}{2\sigma} \sim N(0,1) \quad \text{under } H_0 : \theta = \theta_0$$

- Note, however, that if the sample sizes $\tilde{n}_j$ and the incremental sample sizes $\tilde{n}_j^*$ depend on interim effect estimates, we do not have $\tilde{n}_j$ independent of Z_ℓ^* for $\ell \neq j$.

Adaptation Rules: Reweighted Statistic/Combining p-values

- Therefore, the statistic

$$\tilde{Z}_j = \frac{\sum_{\ell=1}^{j} \sqrt{\tilde{n}_\ell^*}\tilde{Z}_\ell^*}{\sqrt{\tilde{n}_j^*}}$$

may not be $N(0,1)$ under H_0, as it is no longer a standardized sum of independent normal random variables.

Adaptation Rules: Reweighted Statistic/Combining p-values

- If instead we use pre-specified weights (variances) w_ℓ for each Z_ℓ^* in computing the test statistic, we do obtain a standard normal random variable under H_0:

$$Y_j = \frac{\sum_{\ell=1}^{j} \sqrt{w_\ell}\tilde{Z}_\ell^*}{\sqrt{\sum_{\ell=1}^{j} w_\ell}}$$

since $\sqrt{w_\ell}Z_\ell^* \sim N(0, w_\ell)$ so $\sum_{\ell=1}^{j} \sqrt{w_\ell}Z_\ell^* \sim N(0, \sum_{\ell=1}^{j} w_\ell)$

- A natural choice for the weights w_ℓ is the originally planned sample sizes

$$w_\ell = n_\ell^*$$

Adaptation Rules: Reweighted Statistic/Combining p-values

- The statistic Y_j is compared to the originally planned stopping boundary critical values for the jth stage.
- This procedure has the same Type I error rate as the original group sequential design.

Adaptation Rules: Reweighted Statistic/Combining p-values

- The reweighting approach can also be expressed, more generally, as an approach of combining p-values.
- Setting
  - A total of J potential analyses are allowed.
  - The data gathered in each stage is independent of all other stages (independent increments)
  - Incremental p-values P_j^* for each stage are exact (or at least near-exact) in the sense that

$$P_{H_0}(P_j^* \leq u) \approx u \quad \text{for all } u \in [0,1]$$

Adaptation Rules: Reweighted Statistic/Combining p-values

- Then the following test statistic may be compared to a level α stopping boundary with J analyses on the Z-scale (reject H_0 for large $Q_j \Leftrightarrow$ greater alternative).

$$Q_j = \frac{1}{\sqrt{j}} \sum_{i=1}^{j} \Phi(1 - P_j^*)$$

- This approach protects the type I error rate at level α, no matter what incremental sample sizes n_j^* are used for each stage.
- Incremental sample sizes may be modified at any time
- Number of possible future analyses may *not* be changed

Adaptation Rules: Conditional Error/Modified Critical Value

- The second type of adaptation rule also starts with a group sequential design:
  - $\{J, n_j, T_j, (a_j, b_j, c_j, d_j)$ for $j = 1, \ldots, J\}$
  - At some interim analysis $h\,(1 \leq h < J)$, the entire future sampling plan and stopping boundary may be modified:
    - $n_j^* \rightarrow \tilde{n}_j^*$ for $j = h+1, h+2, \ldots, J$
    - $(a_j, b_j, c_j, d_j) \rightarrow (\tilde{a}_j, \tilde{b}_j, \tilde{c}_j, \tilde{d}_j)$ for $j = h+1, h+2, \ldots, \tilde{J}$
  - Note that the entire **sample path**: (sample size, boundary, and number of future analyses) may change.

Adaptation Rules: Conditional Error/Modified Critical Value

- The **conditional rejection rate**
  - at a specified true value of the parameter θ,
  - given the current test statistic value/estimate of effect size $\hat{\theta}_h = t_h$, and
  - using a particular future sampling path k: $\{\tilde{J}^{(k)}, \tilde{n}_j^{(k)}, \tilde{T}_j^{(k)}, (\tilde{a}_j^{(k)}, \tilde{b}_j^{(k)}, \tilde{c}_j^{(k)}, \tilde{d}_j^{(k)}) \text{ for } j = h+1, \ldots, \tilde{J}^{(k)}\}$

  is given by

$$CP_\theta(\text{Sampling Path } k | \hat{\theta}_h = t_h) =$$
$$P_\theta(\text{Reject } H_0 \text{ at any } j = h+1, \ldots, \tilde{J}^k \text{ using sampling path } k | \hat{\theta}_h = t_h)$$

Adaptation Rules: Conditional Error/Modified Critical Value

- **Conditional type I error rate** is the conditional rejection rate under the null hypothesis $H_0 : \theta = \theta_0$
- If we constrain our future sampling paths to match the original conditional type I error rate, i.e. ensure that

$$CP_{\theta_0}(\text{Original Sampling Path } k = 0 | \hat{\theta}_h = t_h) =$$
$$CP_{\theta_0}(\text{Sampling Path } k | \hat{\theta}_h = t_h)$$

  then the overall type I error rate of the adaptive design is controlled at the original level α.

Adaptation Rules: Conditional Error/Modified Critical Value

- Popular methods of conditional error adaptation:
  - Consider adaptation at the next to last stage $h = J - 1$.
  - Modification of final sample size only; no increase in number of future analyses is considered.
  - To maintain conditional type I error rate, the critical values $(\tilde{a}_J, \tilde{b}_J, \tilde{c}_J, \tilde{d}_J)$ must be adjusted based on the new final sample size $\tilde{n}_J = n_{J-1} + \tilde{n}_J^*$
  - Typically, a one-sided design is considered, so $\tilde{a}_J = \tilde{d}_J$ and therefore a single critical value $\tilde{a}_J(\tilde{n}_J)$ must be solved for, in terms of the new incremental sample size $\tilde{n}_J^*$.

Adaptation Rules: Conditional Error/Modified Critical Value

- Gao, Ware, and Mehta 2008 provide formulae for the critical value $\tilde{a}(\tilde{n}_J^*)$ given an observed test statistic $Z_{J-1} = z_{J-1}$ and a new incremental sample size $\tilde{n}_J^*$:

$$\tilde{a}_J(\tilde{n}_J^*) = \frac{1}{\sqrt{\tilde{n}_J}} \left[\frac{\sqrt{\tilde{n}_J^*}}{\sqrt{n_J^*}} \left(a_J \sqrt{n_J} - z_h \sqrt{n_{J-1}} \right) + z_h \sqrt{n_{J-1}} \right]$$

- It can be shown that this is equivalent to reweighting the z-statistic and using the original critical value a_J:
  - Changing the statistic, keeping the boundary $\Leftrightarrow$ Keeping the statistic, changing the boundary

Adaptation Rules: Conditional Error/Modified Critical Value

- Contrary to the statement in Gao, Ware, and Mehta 2008:

  > *"The equivalence of the three methods demonstrates that the sample size re-estimation method of Cui, Hung, and Wang is valid and does not truly down-weight any portion of the data."*

  this equivalence instead demonstrates that modifying the critical value based on a new sample size *is the same as down-weighting some of the data* and is therefore likely to be an inefficient approach.

Example Adaptation Rules: Conditional Error/Modified Critical Value

- Adaptive final sample size may be chosen according to any desired criteria
- Popular choice of adaptive final sample size is to chose $\tilde{n}_J$ to attain a desired level of **conditional power**, where the conditional power is evaluated using the current effect size estimate $\hat{\theta}_h$ as the true parameter.

Adaptation Rules: Conditional Error/Modified Critical Value

- Given desired conditional power $1-\beta$, we find $\tilde{n}_J^*$ such that

$$P_{\hat{\theta}_L}\left(\tilde{Z}_J > \tilde{a}_J(\tilde{n}_J^*)\right) = 1-\beta$$

  where $\tilde{Z}_J$ is the cumulative z-statistic using $\tilde{n}_J = n_{J-1} + \tilde{n}_J^*$ observations.
- Since $\tilde{a}_J(\tilde{n}_J^*)$ is a function of $\tilde{n}_J^*$, this expression can be solved for the desired value $\tilde{n}_J^*$.

Adaptation Rules: Conditional Error/Modified Critical Value

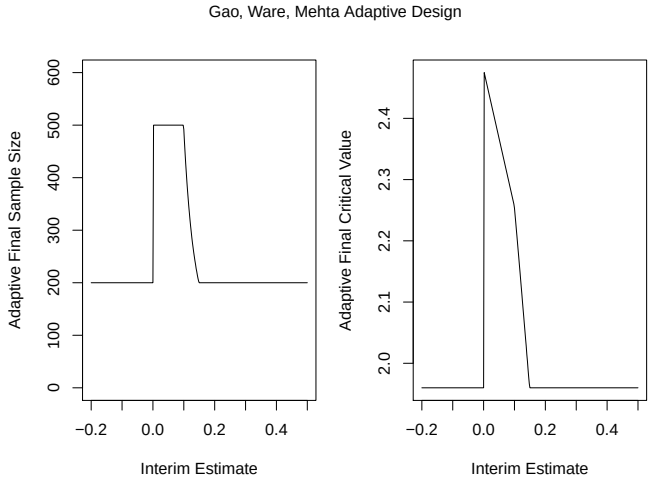
- Gao, Ware, and Mehta 2008 also provide formulae for the new final sample size $\tilde{n}_J$ needed to obtain conditional power of $1-\beta$, given that $Z_{J-1} = z_{J-1}$:

$$\tilde{n}_J = \frac{n_{J-1}}{z_{J-1}^2}\left[\frac{\left(a_J\sqrt{n_J} - z_{J-1}\sqrt{n_{J-1}}\right)}{\sqrt{n_J^*}} + z_\beta\right]^2 + n_{J-1}$$

Adaptation Rules: Conditional Error/Modified Critical Value

- Consider a design to test $H_0 : \theta = 0$ vs. $H_A : \theta = 0.5544$ at level $\alpha = 0.025$ with power 0.975 to detect the alternative.
- Sample size $N = 200$ with critical value 0.2772 on the sample mean scale (1.96 on the z-statistic scale)
- After $N_1 = 100$ subjects, an interim analysis is performed to adaptively modify the final sample size.
- The final sample size is chosen to obtain conditional power 0.90 at the interim estimate of effect size.
- A maximum sample size of 500 is allowed.

Adaptation Rules: Conditional Error/Modified Critical Value



Adaptation Rules: Adaptive Switching between Sampling Path

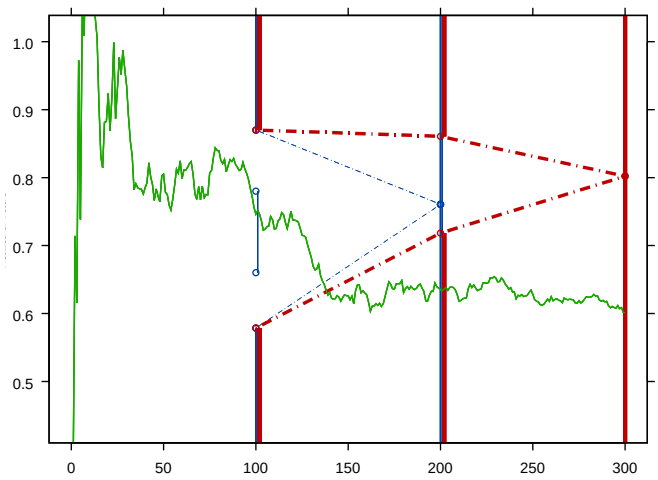
- The third, most general type of adaptation rule can be thought of as adaptively switching between different group sequential designs (different **sampling paths**).
- Starting with a group sequential design:
  - $\{J^{(0)}, n_j^{(0)}, T_j^{(0)}, (a_j^{(0)}, b_j^{(0)}, c_j^{(0)}, d_j^{(0)})$ for $j = 1, \ldots, J^{(0)}\}$
- At some interim analysis $h\,(1 \leq h < J)$, adaptively select one of r possible future sampling paths.
- Valid adaptive design controlling overall type I error rate as long as:
  - Total probability under $H_0 : \theta = \theta_0$ of rejecting H_0 is constrained to be $\leq \alpha$
  - Exactly one future sampling path is selected.

Adaptation Rules: Adaptive Switching between Sampling Path

Details:

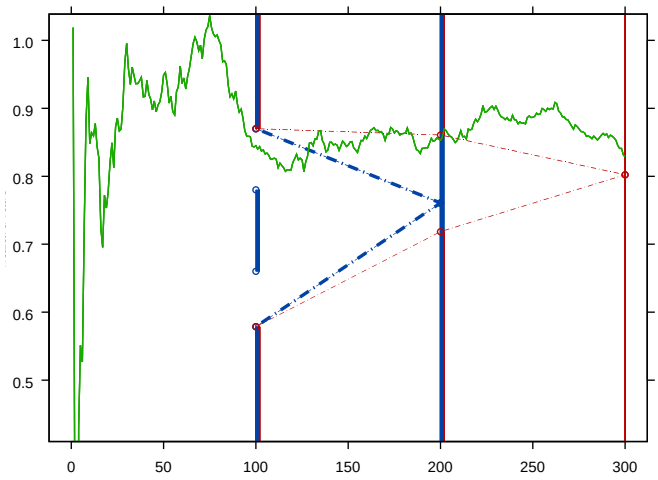
- At interim analysis $h\,(1 \leq h < J)$, the continuation region for $T_h^{(0)}$ is partitioned into r disjoint continuation sets $\mathcal{C}_h^{(k)}$, for $k = 1, \ldots, r$.
- If $T_h^{(0)} \in \mathcal{C}_h^{(k)}$, then the future stopping boundary will be the kth future group sequential sampling path:
  - $\{J^{(k)}, n_j^{(k)}, T_j^{(k)}, (a_j^{(k)}, b_j^{(k)}, c_j^{(k)}, d_j^{(k)})$ for $j = h+1, \ldots, J^{(k)}\}$, for $k = 1, \ldots, r$
- Let K be the random variable denoting which path is chosen, $K \in \{1, \ldots, r\}$: $K = k$ if $T_h^{(0)} \in \mathcal{C}_h^{(k)}$.

Adaptation Rules: Adaptive Switching between Sampling Path



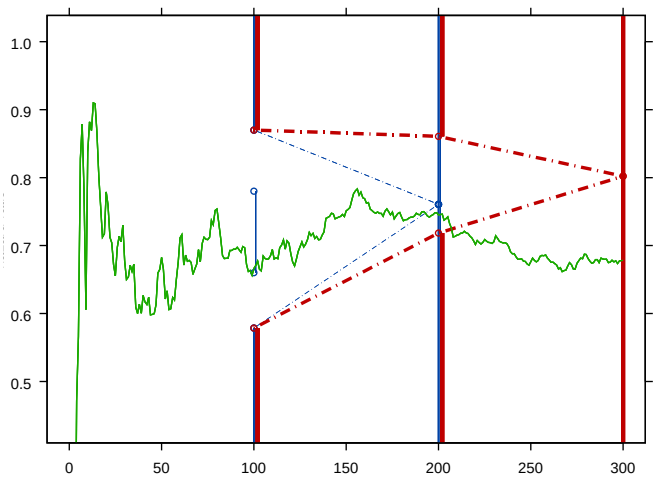
Foundations: Group Sequential Designs
Inference following Group Sequential Designs
Adaptive Sequential Designs
Forms of Adaptation Considered
Considerations in Adapting Future Sampling Path
Types of Adaptation Rules
Adaptive Designs using Standard Group Sequential Software

Adaptation Rules: Adaptive Switching between Sampling Path

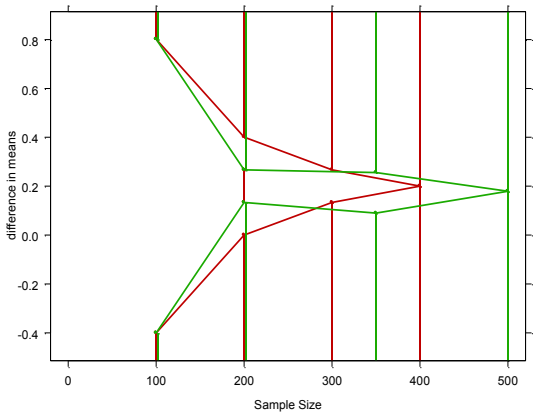


Sarah Emerson and Scott Emerson
Adaptive Designs
136 / 143

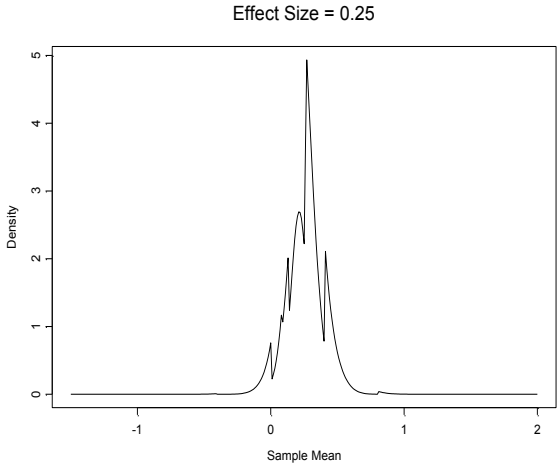
Adaptation Rules: Adaptive Switching between Sampling Path



Adaptation Rules: Adaptive Switching between Sampling Path

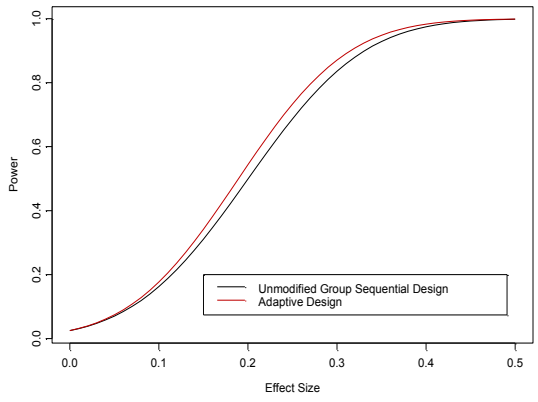


Adaptation Rules: Adaptive Switching between Sampling Path



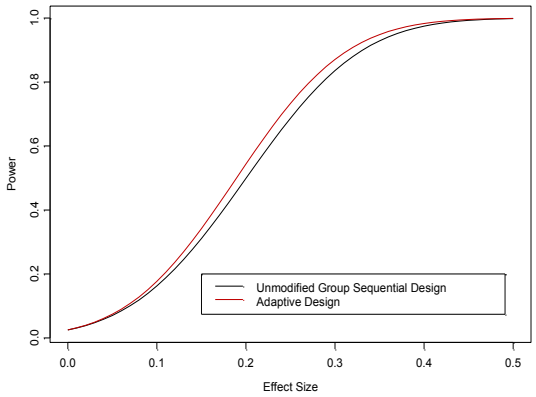
Foundations: Group Sequential Designs
Inference following Group Sequential Designs
Adaptive Sequential Designs
Forms of Adaptation Considered
Considerations in Adapting Future Sampling Path
Types of Adaptation Rules
Adaptive Designs using Standard Group Sequential Software

Adaptation Rules: Adaptive Switching between Sampling Path



Sarah Emerson and Scott Emerson
Adaptive Designs
137 / 143

Adaptation Rules: Adaptive Switching between Sampling Path



Adaptation Rules: Adaptive Switching between Sampling Path

- Compared to the previous two approaches (reweighting and preserving conditional type I error rates), the adaptive switching approach is more flexible.
- Control of the unconditional type I error rate may be accomplished without constraining the conditional type I error rates.

Adaptation Rules: Adaptive Switching between Sampling Path

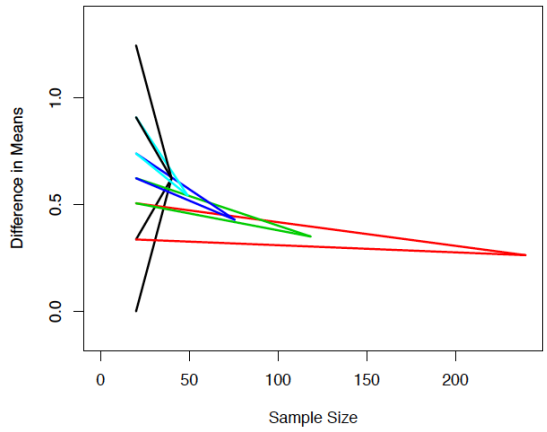
- Adaptive designs as proposed in Gao, Ware, Mehta 2008 (GWM) and others can be represented in this adaptive switching framework.
- Since
  - sample sizes must be discrete, and
  - there is almost always (always) a maximal possible sample size set

  any adaptive rule can be regarded as switching between a finite number of future group sequential sampling paths.

Adaptation Rules: Adaptive Switching between Sampling Path

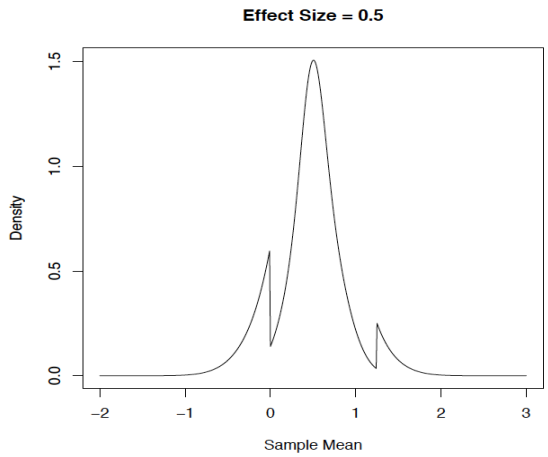
- In practice, we have found that the performance of an adaptive rule with a large number r of different possible group sequential sampling paths is not much different from an adaptive rule with a small number of possible sampling paths.
- e.g. A discretized GWM design with just $r = 4$ different possible group sequential sampling paths has practically identical performance to one with $r = 100$ different sampling paths.

Adaptation Rules: Adaptive Switching between Sampling Path



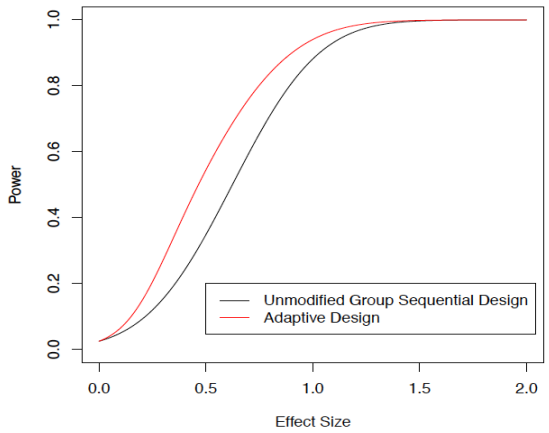
Foundations: Group Sequential Designs
Inference following Group Sequential Designs
Adaptive Sequential Designs
Forms of Adaptation Considered
Considerations in Adapting Future Sampling Path
Types of Adaptation Rules
Adaptive Designs using Standard Group Sequential Software

Adaptation Rules: Adaptive Switching between Sampling Path



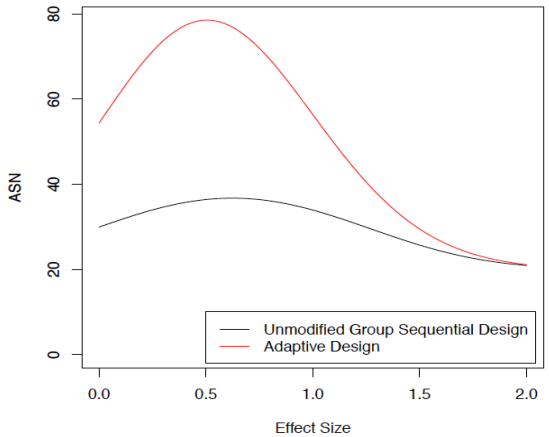
Sarah Emerson and Scott Emerson
Adaptive Designs
141 / 143

Adaptation Rules: Adaptive Switching between Sampling Path



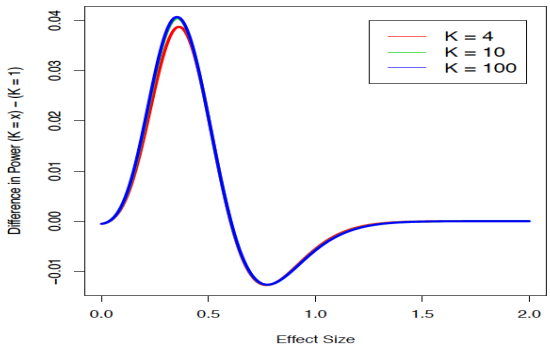
Foundations: Group Sequential Designs
Inference following Group Sequential Designs
Adaptive Sequential Designs
Forms of Adaptation Considered
Considerations in Adapting Future Sampling Path
Types of Adaptation Rules
Adaptive Designs using Standard Group Sequential Software

Adaptation Rules: Adaptive Switching between Sampling Path



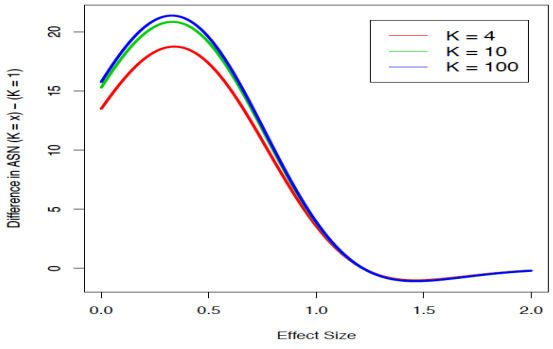
Sarah Emerson and Scott Emerson
Adaptive Designs
141 / 143

Adaptation Rules: Adaptive Switching between Sampling Path



Foundations: Group Sequential Designs
Inference following Group Sequential Designs
Adaptive Sequential Designs
Forms of Adaptation Considered
Considerations in Adapting Future Sampling Path
Types of Adaptation Rules
Adaptive Designs using Standard Group Sequential Software

Adaptation Rules: Adaptive Switching between Sampling Path



Sarah Emerson and Scott Emerson
Adaptive Designs
141 / 143

1. Foundations: Group Sequential Designs
   - Clinical Trial Design
   - Example Setting
   - Design Comparison
2. Inference following Group Sequential Designs
   - Inference Goals
   - Inference Approaches
   - Inference Optimality Criteria
3. Adaptive Sequential Designs
   - Forms of Adaptation Considered
   - Considerations in Adapting Future Sampling Path
   - Types of Adaptation Rules
   - Adaptive Designs using Standard Group Sequential Software

Adaptive Switching Designs using Standard Group Sequential Software

- Any pre-specified adaptive design can be represented and calculated using standard group sequential software that allows:
  - Arbitrarily spaced analyses
  - Constrained and partially constrained boundary searches
  - Numerical integration to find stopping probabilities and stopping densities
- Basic idea:
  - Specify each possible group sequential sampling path after analysis time h as a different group sequential design with first analysis at time n_h.

Adaptive Sample Size Re-estimation: Design and Inference

Sarah Emerson and Scott Emerson

Outline

1. Statistical Efficiency of Adaptation
2. Complete Inference after Adaptation
   - Inference for Pre-specified Design
   - Inference after Unplanned Adaptation
3. Evaluating Inferential Methods
4. Additional Issues
5. Adaptive Time-to-event Setting

1 Statistical Efficiency of Adaptation

2 Complete Inference after Adaptation
- Inference for Pre-specified Design
- Inference after Unplanned Adaptation

3 Evaluating Inferential Methods

4 Additional Issues

5 Adaptive Time-to-event Setting

Competing Issues in Clinical Trials

- Ethics: individual and collective
- Clinical science: overall patient health
- Basic Science: mechanisms
- Statistical: reliable and precise answers
- Economic/Operational: feasibility, profits and/or costs

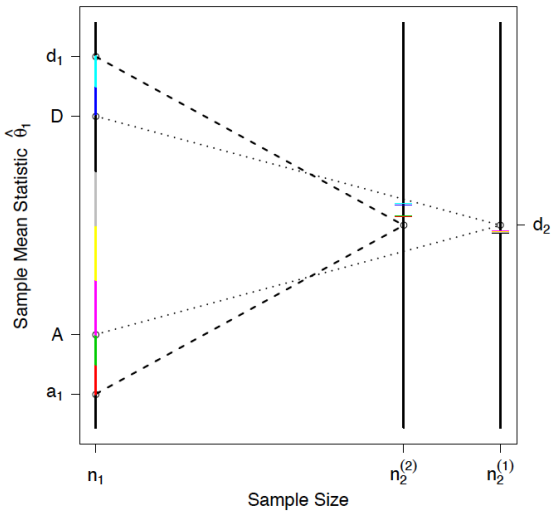
Considering Adaptation

- What do we gain?
  - Efficiency?
  - Flexibility?
- What do we lose?
  - Efficiency?
  - Interpretability?
  - Ease of implementation?
- How do we make fair comparisons?
  - Same number or schedule of analyses, or trial duration?
  - Same power at the alternative? Same power curve?
  - How to measure efficiency?

Efficiency of Adaptive Testing

- Methods of adaptive hypothesis testing based on combination or conditional error functions violate sufficiency principle
  - Same sample mean and N at stopping could lead to opposite decisions (see next slide)
- Suffer efficiency losses compared to GSDs
  - Losses of $\sim 40\%$ in certain cases (Jennison and Turnbull 2006)
- Efficiency loss due to testing method or poor sample size modification rules?

Violation of Sufficiency Principle



Our Research on Efficiency

- Consider completely pre-specified adaptive designs with testing adhering to sufficiency principle
  - Differences in operating characteristics due to adaptation rule, not testing method
- Explore efficiency gains over group sequential designs
- Explore efficient types of adaptations
- Compare to frequently proposed adaptation rules

Setting and Notation

- Potential observations X_{Ai} on treatment A and X_{Bi} on treatment B, for $i = 1, 2, ...,$ independently distributed
  - Means μ_A and μ_B and common known variance σ^2
- Parameter of interest: $\theta = \mu_A - \mu_B$
  - Positive values of θ indicate superiority of new treatment
- Up to J interim analyses with sample sizes $N_1, N_2, N_3, ..., N_J$
- At the jth analysis, let
  - Partial Sum: $S_j = \sum_{i=1}^{N_{Aj}} X_{Ai} - \sum_{i=1}^{N_{Bj}} X_{Bi}$
  - MLE: $\hat{\theta}_j = \overline{X}_{Aj} - \overline{X}_{Bj}$

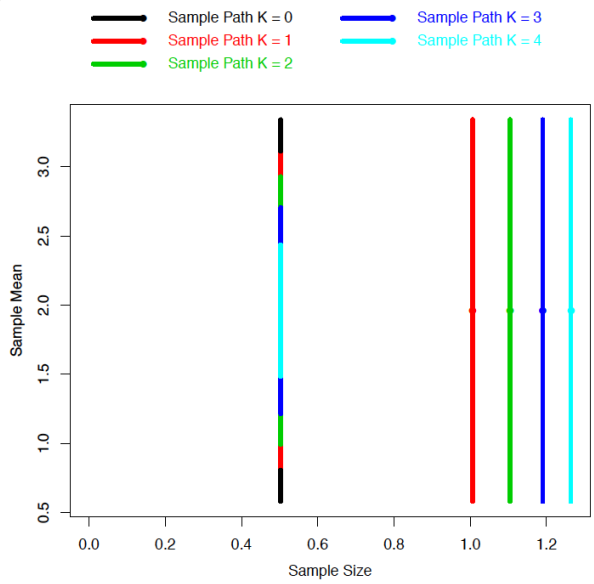
Setting and Notation

- Upper-case letters for random variables, lower-case for fixed quantities
- Use a * to denote incremental data
  - $N_j^* = N_j - N_{j-1}$ (with $N_0 = 0$)
  - $S_j^* = \sum_{i=N_{Aj-1}+1}^{N_{Aj}} X_{Ai} - \sum_{i=N_{Bj-1}+1}^{N_{Bj}} X_{Bi}$
  - $\hat{\theta}_j^* = \overline{X}_{Aj}^* - \overline{X}_{Bj}^*$ and $Z_j^* = \frac{(\hat{\theta}_j^* - \theta_0)}{\sqrt{\frac{\sigma^2}{N_{Aj}^*} + \frac{\sigma^2}{N_{Bj}^*}}}$
- Outcomes immediately observed
- Test null $H_0 : \theta = \theta_0 = 0$ against one-sided alternative $\theta > 0$

A Class of Pre-specified Adaptive Designs

- Single adaptation occurs at analysis time $j = h$
- At adaptation analysis ($j = h$), there are r mutually exclusive continuation sets, denoted C_h^k, $k = 1, \ldots, r$
- Each continuation set C_h^k at adaptation analysis corresponds to future group sequential path k
- Random sample path variable K can take values $0, 1, \ldots, r$
- Define three-dimensional test statistic (M, S, K)
  - M is stage, S is partial sum, K is path at stopping

Example of Adaptive Design



Sampling Density

- $N_j^* = n_j^{k*}$ is fixed conditional on $S_{j-1} = s \in C_{j-1}^k$
- Assume equal allocation ($N_{Aj} = N_{Bj}$). Appealing to the central limit theorem,
  - $S_1^* \sim N(n_{A1}^0\,\theta,\ 2\,n_{A1}^0\,\sigma^2)$
  - $S_j^* \mid S_{j-1} \sim N(n_{Aj}^{k*}\,\theta,\ 2\,n_{Aj}^{k*}\,\sigma^2)$

Sampling Density

Following Armitage et al. (1969), density of $(M = j, S = s, K = k)$ is

$$p_{M,S,K}(j,\ s,\ k;\ \theta) = \begin{cases} f_{M,S,K}(j,\ s,\ k;\ \theta) & \text{if } s \in \mathcal{S}_j^k \\ 0 & \text{otherwise} \end{cases}$$

where the (sub)density is recursively defined as

$$f_{M,S,K}(1,\ s,\ 0;\ \theta) = \frac{1}{\sqrt{2\,n_{A1}^0}\,\sigma}\,\phi\left(\frac{s - n_{A1}^0\,\theta}{\sqrt{2\,n_{A1}^0}\,\sigma}\right)$$

$$f_{M,S,K}(j,\ s,\ k;\ \theta) = \int_{C_{j-1}^k} \frac{1}{\sqrt{2\,n_{Aj}^{k*}}\,\sigma}\,\phi\left(\frac{s - u - n_{Aj}^{k*}\,\theta}{\sqrt{2\,n_{Aj}^{k*}}\,\sigma}\right) f_{M,S,K}(j,\ u,\ k;\theta)\,du$$

for $k = 0, j = 2, \ldots, h$ (if $h > 1$) and $k = 1, \ldots, r, j = h+1, \ldots, J_k$

Sampling Density

Easy to show the following relation:

$$p_{M,S,K}(j,\ s,\ k;\ \theta) = p_{M,S,K}(j,\ s,\ k;\ 0) \exp\left(\frac{s\,\theta}{2\,\sigma^2} - \frac{\theta^2}{4\,\sigma^2}\, n_{Aj}^k\right)$$

$\Rightarrow$ MLE is sample mean $\hat{\theta} = \overline{X}_A - \overline{X}_B$

$\Rightarrow$ (N, S) minimally sufficient for θ

Computations

- Can compute density of sample mean, $\beta(\theta)$, ASN(θ), etc.
- All computations just functions of density and/or operating characteristics (OC) of a set of $r + 1$ group sequential designs
- Can modify existing group sequential software to carry out computations
- All our results using R package RCTdesign built from S-Plus module S+SeqTrial

Efficiency of Adaptive Testing

Our research on efficiency...

- Define optimality criteria in two simple, realistic RCT settings with different scientific constraints
- Derive optimal competing fixed sample, GS, adaptive designs
  - Restrict attention to symmetric designs
- Compare operating characteristics
- Describe in detail sampling plan of optimal adaptive designs

Setting 1: Optimality Criteria

- Number of analyses constrained to max of two
- Type I error $\alpha = 0.025$, power $\beta = 0.975$ at $\theta = \Delta$
- Initial candidate design: fixed $n = 4\,\frac{(z_{1-\alpha}+z_{\beta})^2}{\Delta^2}$ (WLOG, $\sigma^2 = 1$)
- Primary interest: find most efficient design meeting constraints
  - Efficiency measured by average sample size in presence of truly ineffective (under null) or effective (under alternative) treatment

Setting 1: Optimal GSD

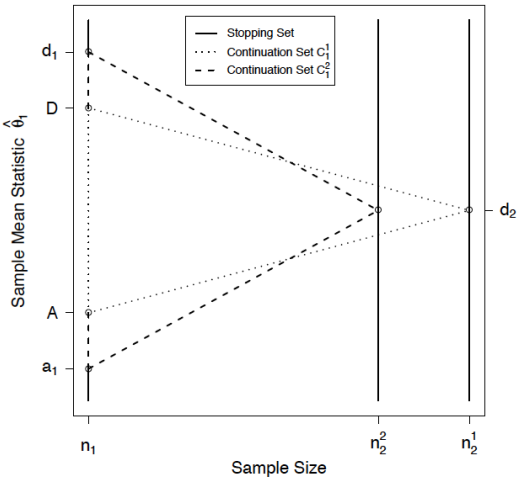
- 2-analysis GSD with Pocock-like stopping boundaries
- Analyses at 50% and 118% of original fixed sample size n
- Stopping boundaries for futility and efficacy at first analysis 0.21Δ and 0.79Δ on sample mean scale
  - $(0.57, 2.21)$ on Z-scale
  - $(4.9\%, 95.1\%)$ on conditional power scale assuming $\theta = \hat{\theta}_1$
  - $(81.8\%, 99.0\%)$ on conditional power scale assuming $\theta = \Delta$
- ASN of 68.54% of fixed sample size n at design alternatives

Finding the "Optimal" Adaptive Design

Find optimal adaptive designs with increasing number r of continuation regions...

1. Holding constant α, β, first-stage stopping bounds of optimal GSD, choose C_1^1 and n_2^1 to minimize ASN at design alternatives based on numerical grid search
2. Proceed to 3 continuation regions by holding C_1^1 constant and finding optimal split of C_1^2 into 2 continuation regions
3. Proceed to 4 continuation regions by optimally splitting $C_1^1 \ldots$

Finding the "Optimal" Adaptive Design



Setting 1: Results

Table: Average, maximal sample sizes of competing designs in units of n

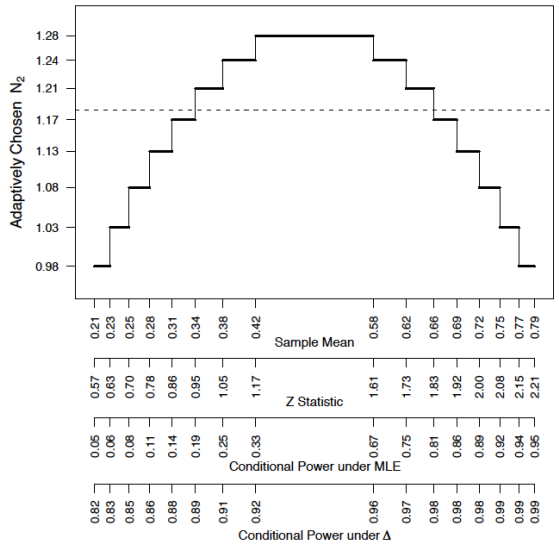
	Number of Continuation Regions								
	0^a	1^b	2	3	4	5	6	7	8
$ASN_{\theta=0,\Delta}$	1	0.6854	0.6831	0.6828	0.6825	0.6824	0.6824	0.6824	0.6824
% Difference	+45.9%	*Ref*	-0.34%	-0.38%	-0.42%	-0.43%	-0.43%	-0.44%	-0.44%
Maximal N	1	1.18	1.24	1.24	1.26	1.26	1.26	1.26	1.28

a. Fixed Sample Design

b. Group Sequential Design (*Reference* design)

- Efficiency gain by optimal adaptive design minimal ($< 0.5\%$)
- Gain largely achieved with $r = 2$, negligible decreases with $r > 4$

Setting 1: The Optimal Adaptive Design



Setting 1: Describing the Design

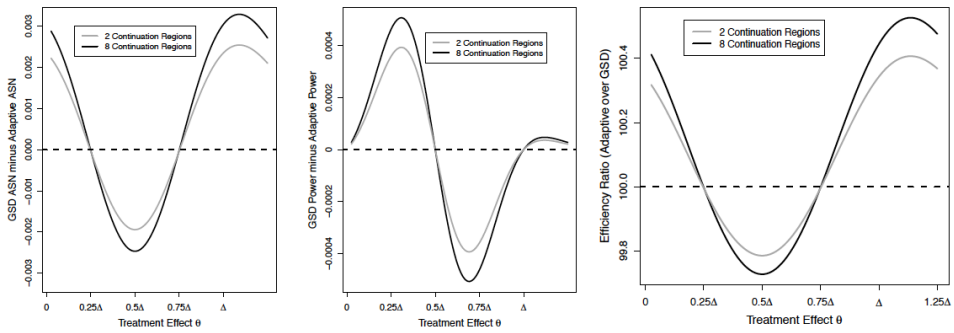
- Increasing # regions only modestly increases maximal N
  - Designs frequently proposed in literature allow $\geq$ 2-fold increase
- Largest maximal sample sizes chosen near center of group sequential continuation region, smallest near boundaries
- $\sim$ Optimal thresholds for increasing N_2 (relative to GSD)
  - $(0.34\Delta, 0.66\Delta)$ on sample mean scale
  - $(0.95, 1.83)$ on Z scale
  - $(0.19, 0.81)$ on CP($z1$; MLE) scale
  - $(0.89, 0.98)$ on CP($z1$; Δ) scale
- Thresholds on conditional power scale change substantially based on presumption of MLE or Δ as true treatment effect

Setting 1: Other Efficiency Considerations

- Efficiency gain at alternatives ($\theta = 0$ and $\theta = \Delta$) offset by losses at intermediate treatment effects ($0.25\Delta - 0.75\Delta$)
  - ASN increases $\sim$ same magnitude as efficiency gains
- Negligible power differences (< 0.0005) between adaptive design and GSD at intermediate θs
- Adding additional analysis to GSD leads to much larger efficiency gain than allowing adaptivity
  - Reduces ASN of GSD by 6.3% as compared to $< 0.5\%$

Setting 1: Other Efficiency Considerations

- Efficiency index of design A: ratio of fixed sample size needed to match its power over its ASN



Setting 2: Optimality Criteria

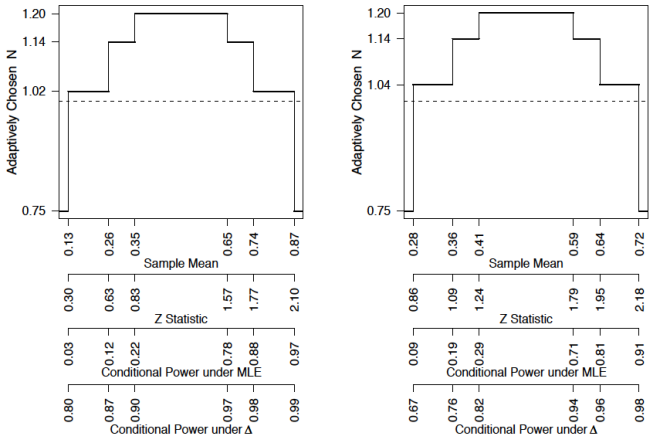
- Only one "stopping analysis"
- Earlier "adaptation" analysis to determine optimal sample size
- $\alpha = 0.025$, $\beta = 0.975$ at $\theta = \Delta$, candidate fixed $n = 4\,\frac{(z_{1-\alpha}+z_{\beta})^2}{\Delta^2}$
- Minimum sample size for stopping of $n_{min} < n$ required for adequate safety profile
  - Assume $n_{min} = 0.75n$ (similar patterns with other choices)
- "Adaptation" analysis may occur at range of time points n_{adapt}
  - Let $n_{adapt} = q * n_{min}$ and consider $q \in \{0.1, 0.2, ..., 0.9, 1.0\}$
- Primary interest: find most efficient design meeting constraints

Setting 2: Results

	q (Proportion of n_{min} at which adaptation occurs)									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$ASN_{\theta=0,\Delta}$	0.99	0.97	0.94	0.91	0.88	0.86	0.84	0.82	0.80	0.78
Maximal N	1.07	1.12	1.16	1.18	1.20	1.21	1.21	1.20	1.18	1.17

- Adding "adaptation" analysis leads to meaningful efficiency gains over fixed sample test, reducing ASN by $\sim 20\%$
- Best design allows stopping at "adaptation" analysis
- Behavior improves as statistical info at adaptation increases

Setting 2: The Optimal Adaptive Designs



(a) q = 0.5 (b) q = 0.8

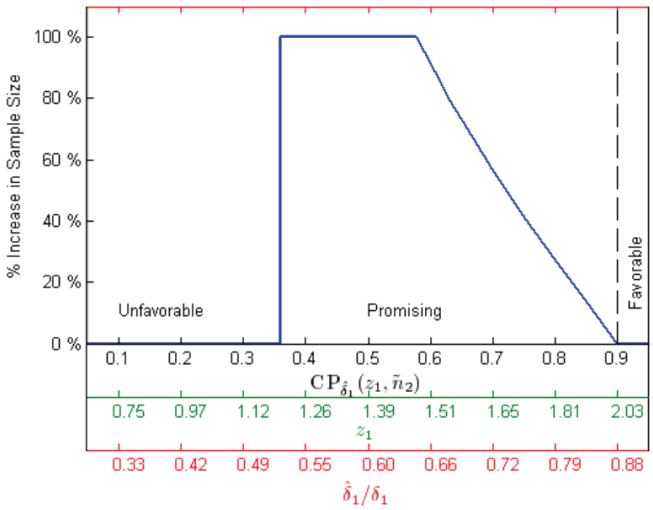
Setting 2: Describing the Designs

- Largest maximal sample size chosen near middle
- N increased up to $\sim 20\%$, much less than frequently proposed
- $\sim$ Optimal thresholds for increasing N ($q = 0.5$)
  - $(0.13\Delta, 0.87\Delta)$ on sample mean scale
  - $(0.30, 2.10)$ on Z scale
  - $(0.03, 0.97)$ on CP($z1$; MLE) scale
  - $(0.80, 0.99)$ on CP($z1$; Δ) scale
- Thresholds on CP scales depend heavily on presumed θ and may not represent intuitive thresholds
- Thresholds on CP($z1$; MLE) scale deviate from designs proposed in literature - have set lower threshold to 36% (MP 2010)

Results on Efficiency

- Optimal adaptive designs attain very small efficiency gains ($< 0.5\%$) over group sequential designs with same # analyses
  - Offset by losses at other plausible treatment effects
  - Far outpaced by adding an analysis to group sequential design
- Insight into good and bad choices of adaptive sampling plans
  - Only few continuation regions and possible final Ns necessary
  - Better to adapt with more information and when stopping permitted
  - Efficient designs qualitatively different than those in literature

Design in Literature (MP 2010)



Limitations

- Many parameters can vary
  - Number, timing of analyses, family of stopping boundaries, definition of "efficiency," scientific constraints
- We covered fraction of this space
  - Focused on symmetric designs in two settings
  - Defined "efficiency" and "optimal" based on ASN at design alternatives, holding power constant
- True minimum ASN not guaranteed for $r > 2$
  - Sensitivity procedures iterating between adjacent regions do not provide further reduction
- Statistical efficiency not only (or most important) concern...

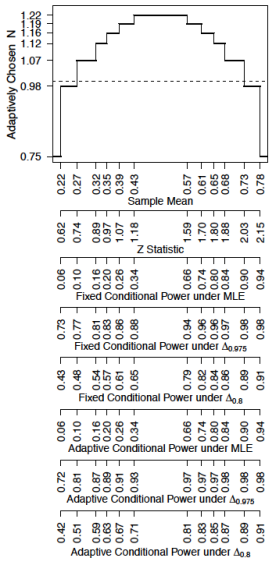
Other Results: Jennison and Turnbull 2006

- Compared optimal pre-specified adaptive designs derived under Bayesian framework to optimal group sequential designs
- Sample size adaptation led to efficiency gains of $< 1.5\%$ (holding constant type I error, power, maximum N, and # analyses)
- "Observed the sampling rules of optimal adaptive tests to be qualitatively different from rules based on conditional power..."
  - Optimal rules selected smaller maximal Ns when interim statistic close to stopping bounds, larger maximal Ns in middle
  - Others reported similar patterns (Posch, Bauer, Brannath 2003)

Efficiency in Survival Setting

- See Emerson, Rudser, and Emerson 2011 (or ask Sarah!)
- In survival setting, statistical information based on # of events, while cost based on # of patients and length of follow-up
- Evaluate tradeoffs between efficiency (average # events), power, cost
- Possibly greater (but still relatively small) benefits from pre-specified adaptation to sampling plan in time-to-event setting
  - Depends on effect size, accrual rate, per-patient cost, interest rate

Stochastic Curtailment and Conditional Power



Stochastic Curtailment and Conditional Power

- Wide range of conditional power values for each boundary as assumptions and reference design vary
  - Efficient threshold on one scale markedly inefficient on another
- Degree of changes in CP do not accurately reflect changes in unconditional power and ASN
- Efficient choices may not correspond to intuitively desirable changes

1-1 Correspondence Between Scales

- 1-1 correspondence between scales for stopping/adaptation boundaries (see Emerson 2007 for relationships)
  - Sample mean, Z statistic, fixed sample P-value, error-spending function, conditional power under $\hat{\theta}$, conditional power under Δ, Bayesian predictive power under some prior, Bayesian posterior probability of some hypothesis
- Choice of scale relatively unimportant if scientific constraints are met, important operating characteristics evaluated
  - Don't choose "intuitive" rule (e.g., stop early if CP$<30\%$, increase N to achieve CP$=90\%$ if CP$<90\%$) and call it a day!

Collaborate, Evaluate, and Iterate

- Consider scientific/regulatory constraints
  - Maximal feasible sample size, minimal sample size (for adequate safety profile), early conservatism
- Consider important operating characteristics
  - Type I error, power under important alternatives, stopping boundaries on different scales, sample size distribution, stopping probabilities, inference reported at stopping
- Compare candidate designs, modify designs to achieve desired operating characteristics, etc.

Schizophrenia Example (Mehta and Pocock 2010)

- Randomized, phase 3 trial of new drug versus control in patients with negative symptoms schizophrenia
- Primary endpoint: change from baseline in Negative Symptoms Assessment (NSA)
- Desire high power at alternative $\Delta = 2$ with SD ~ 7.5
  - Mean difference as small as 1.6 considered clinically important
- Need complete data on at least 200 patients for adequate safety profile
- Assume overrunning minimal (for ease of illustration)

Schizophrenia Example

- Fixed sample design with $n = 442$ and 80% power at $\Delta = 2$ underpowered at $\Delta = 1.6$
- Fixed sample design with $n = 690$ and 80% power at $\Delta = 1.6$ not feasible
- Also consider group sequential and adaptive designs with up to 2 analyses
  - Compare important operating characteristics

1. Statistical Efficiency of Adaptation
2. Complete Inference after Adaptation
   - Inference for Pre-specified Design
   - Inference after Unplanned Adaptation
3. Evaluating Inferential Methods
4. Additional Issues
5. Adaptive Time-to-event Setting

Motivation for Additional Research

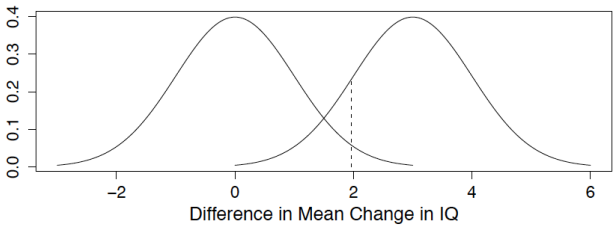
- Despite unclear efficiency gains, adaptive designs implemented in practice, so research needed to propose, evaluate estimation methods
  - Desire for "innovative" designs
  - One sponsor even requires justification if adaptation not included?
- False positive rate and statistical efficiency not only concerns

Case Study: NECAT

- Inhalation of mercury vapor from dental amalgam restorations may have adverse health effects
- Children 6-10 years old randomized to receive dental restoration using either amalgam or resin composite
- Primary outcome: change in full-scale IQ from baseline to 5 years
  - 3 point decline in IQ considered clinically important

Fixed Sample Hypothesis Testing

- Use trial data to decide whether to reject the null hypothesis that amalgam restorations do not lower children's mean IQ
- Design trial to attain low false positive rate (if truly no effect) and high true positive rate (if truly a 3 point average IQ difference)
- Typically 5% false positive rate and 80% or 90% power



Testing versus Estimation

- Testing typically based on *P*-value: probability of obtaining more extreme difference in mean IQ change than what was observed if there were truly no treatment effect
  - If $p < 0.05$, reject null hypothesis of no amalgam effect on IQ
- Four scenarios: What do you conclude?

Study	*P*-value
A	0.263
B	0.263
C	0.025
D	0.025

Testing versus Estimation

- Four scenarios: What do you conclude?

Study	Estimate	Confidence Interval	*P*-value
A	0.5	(-0.4, 1.4)	0.263
B	4.5	(-3.4, 12.3)	0.263
C	0.5	(0.1, 0.9)	0.025
D	4.5	(0.5, 8.4)	0.025

- A: no statistical significance, and ruled out clinical importance
- B: no statistical significance, but consistent with important effect
- C: statistical significance, but ruled out clinical importance
- D: statistical significance, and consistent with important effect

The Need for Good Estimates

- Confirmatory phase III RCTs must produce *interpretable* results
  - Regulatory decisions based on statistical *and clinical* significance
  - Appropriate labeling of newly approved treatment indications
  - Clinicians can effectively practice evidence-based medicine

Complete Inference

Four numbers (with good properties):

- Best point estimate of treatment effect
- Confidence interval providing range of effects consistent with data
- *P*-value reflecting strength of statistical evidence against no effect

Sequential Analyses: Statistical Challenges

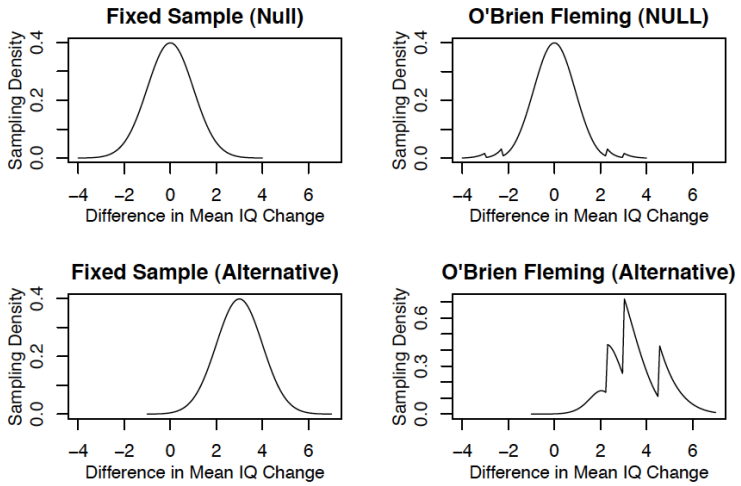
- Sequential testing has implications on estimation of the treatment effect in addition to hypothesis testing
- We stop early only if extreme results are observed
  - Fixed sample estimates such as the sample mean tend to be biased (to the extreme)
  - Confidence intervals do not have correct coverage probabilities (may be conservative or anti-conservative)
- We need point and interval estimates, adjusted for sequential analyses, with desirable "properties"

Connection to Other Types of Studies

- Bottom line: implications of performing *multiple comparisons*
  - Inflated false positive rate
  - Random high bias in estimates of treatment effect for positive results ("winner's curse", "sophomore slump")
- Applies to many other settings
  - Multiple analyses over time
  - Multiple subgroup analyses (e.g. by genetic or other biomarker)
  - Multiple endpoints
  - Publication bias (multiple studies)

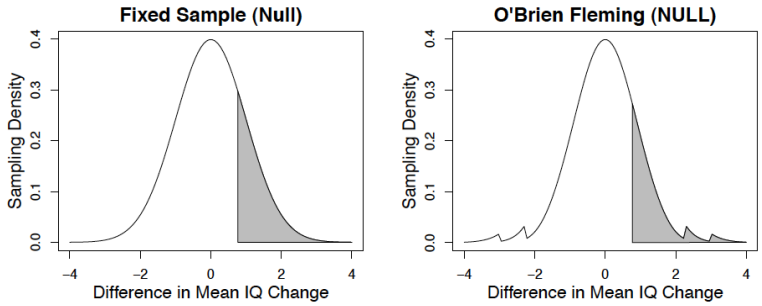
Estimation after Sequential Hypothesis Testing

- Compute estimates, P-values based on true sampling density:



Estimation after Sequential Testing

- Example: *P*-values still probability of observing more "extreme" data under null hypothesis of no treatment effect



Well-understood Methods

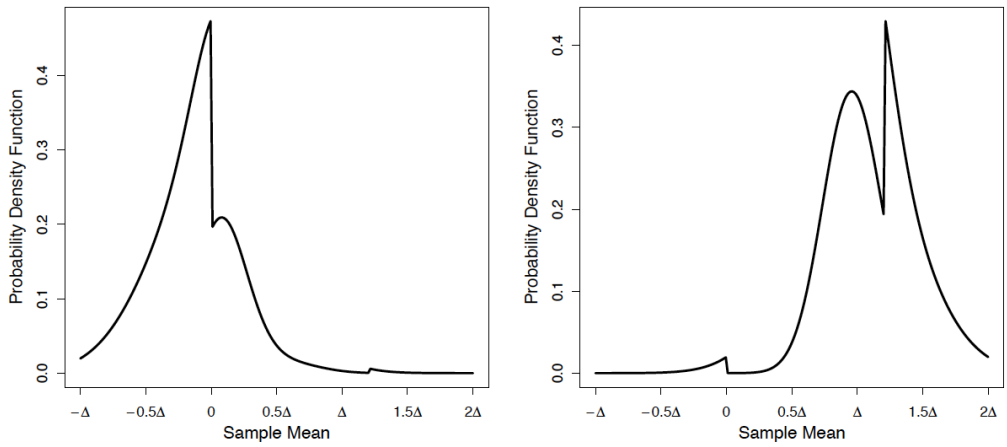
- Extensive literature on estimation after group sequential test
  - Different methods to compute bias-adjusted point estimates, and correct (adjusted) confidence intervals and *P*-values
  - Extensive evaluation of properties assessing the reliability and precision of estimates, CIs, *P*-values
  - Variety of software available for design, conduct, analysis of group sequential designs (PEST, East, SeqTrial, SAS, R)
- Adjusted estimates should be reported, but often are not (even by the best journals)

Extension of Group Sequential Approaches

- Extend orderings of outcome space to adaptive setting
  - Compute p-values
  - Compute confidence regions
  - Compute median-unbiased estimates
- Extend bias-adjusted mean to adaptive setting
- Extend software and evaluate methods

Adaptive Sampling Density of Sample Mean

- Under null (left) and alternative (right)



Duality of Testing and Confidence Sets

- Confidence set: all hypothesized values of θ that would not be rejected by appropriately sized hypothesis test given observed data
- Define acceptance region of "non-extreme" results for each θ:

$$A(\theta, \alpha) = \{(j, t, k) : 1 - \alpha > P[(M, T, K) \succ (j, t, k); \theta] > \alpha\}$$

- Use acceptance region to define equal-tailed $(1 - 2\alpha) \times 100\%$ confidence set:

$$CS^{\alpha}(M, T, K) = \{\theta : (M, T, K) \in A(\theta, \alpha)\}$$

Exact Confidence Sets

To apply, need to define "more extreme" ($\succ$) with outcome ordering:

$$\{(j, t, k) : t \in \mathcal{S}_j^k;\ k = 0, j = 1, \ldots, h \text{ and } k = 1, \ldots, r, j = h + 1, \ldots, J_k\}$$

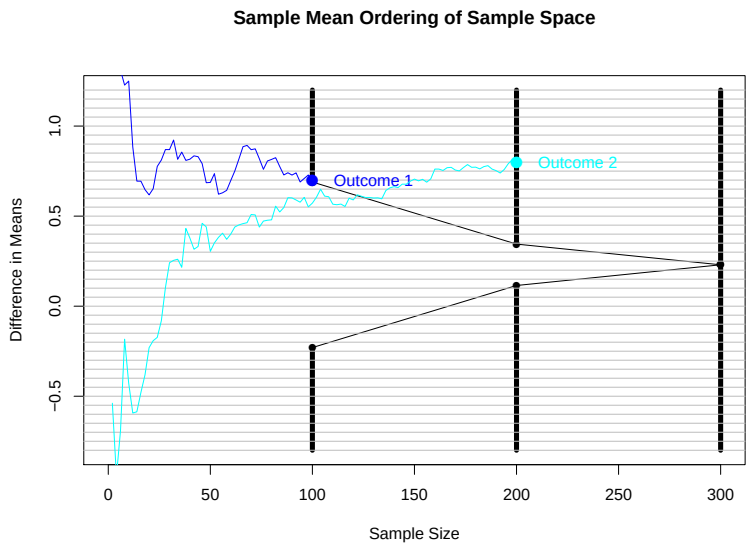
- Neyman-Pearson: likelihood ratio most powerful for simple hypothesis
- Density does not have monotone likelihood ratio, so composite hypothesis theory for optimal tests and CIs does not apply
- Useful to extend straightforward group sequential orderings and evaluate range of properties under variety of designs
  - Relative behavior likely depends on design and treatment effect

Orderings of Outcome Space

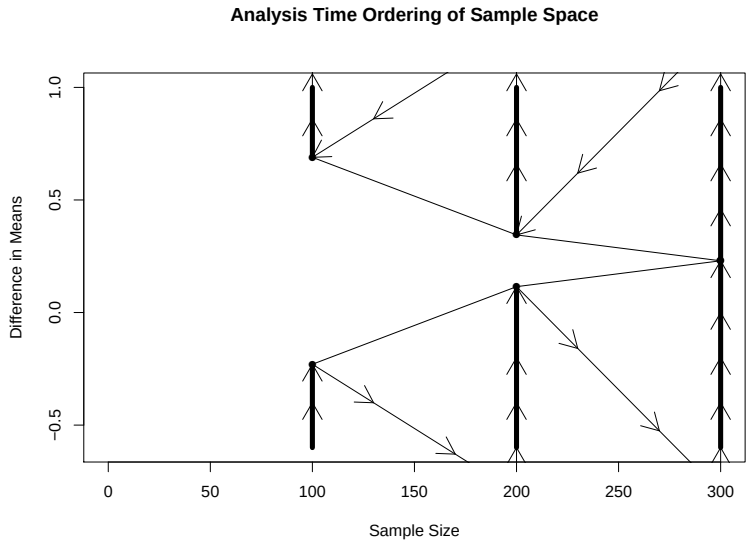
Recall the orderings of the outcome space for standard group sequential designs:

- Sample Mean Ordering
- Analysis Time Ordering
- Signed Likelihood Ratio Ordering

Sample Mean Ordering

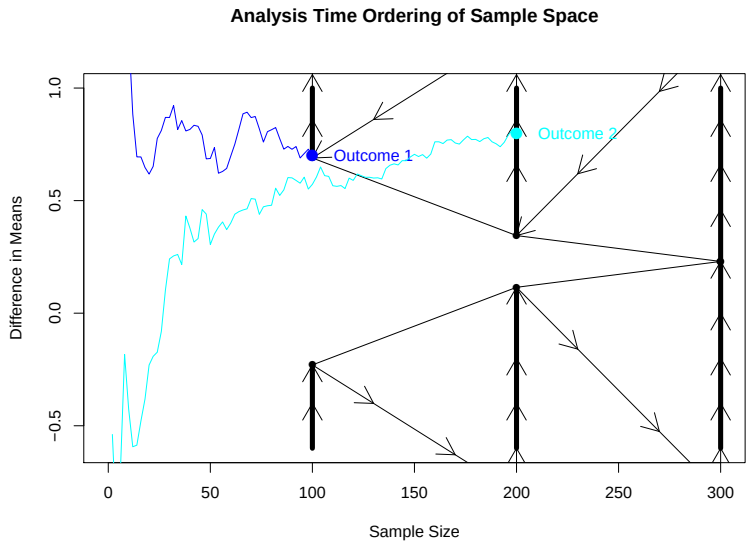


Analysis Time Ordering



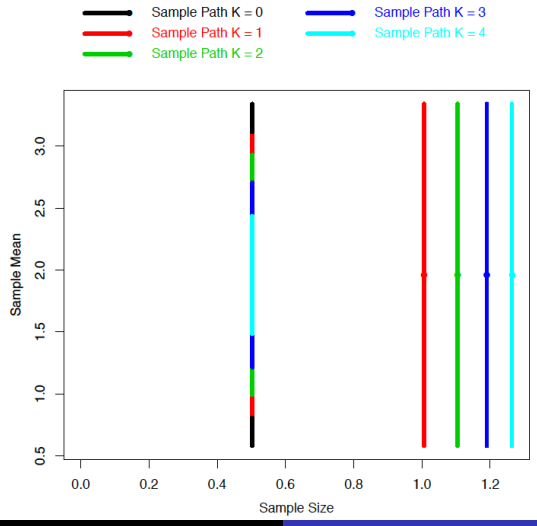
Statistical Efficiency of Adaptation
Complete Inference after Adaptation
Evaluating Inferential Methods
Additional Issues
Adaptive Time-to-event Setting
Inference for Pre-specified Design
Inference after Unplanned Adaptation

Analysis Time Ordering



Orderings of Outcome Space

We can extend these orderings to adaptive sequential trial designs:



Orderings of Outcome Space

- *Sample mean ordering* (SM). Outcomes ordered according to MLE $T \equiv \hat{\theta}$:

$$(j', t', k') \succ (j, t, k) \text{ if } t' > t$$

- *Signed likelihood ratio ordering* (LR). Outcomes ordered according to signed likelihood ratio test statistic against hypothesized θ':

$$(j', t', k') \succ_{\theta'} (j, t, k) \text{ if}$$

$$\text{sign}(t' - \theta') \frac{p_{M,T,K}(j',t',k';\theta=t')}{p_{M,T,K}(j',t',k';\theta=\theta')} > \text{sign}(t - \theta') \frac{p_{M,T,K}(j,t,k;\theta=t)}{p_{M,T,K}(j,t,k;\theta=\theta')}, \text{ i.e., if}$$

$$\sqrt{n^{k'}_{Aj'}}(t' - \theta') > \sqrt{n^{k}_{Aj}}(t - \theta')$$

Orderings of Outcome Space

- *Stage-wise orderings*. Outcomes ordered according to "stage" study stops.
  - Earlier is "more extreme"
  - Unlike GS setting, ranks of analysis times and sample sizes not necessarily equal
  - How to rank statistics observed at same stage through different paths?
  - Several ways to impose this in adaptive setting

Stage-wise Orderings

- *Analysis time + Z statistic ordering (Z)*:

$$(j', t', k') \succ (j, t, k) \text{ if } \begin{cases} j' < j \text{ and } t' \in \mathcal{S}_{j'}^{k'(1)} \\ j' > j \text{ and } t \in \mathcal{S}_{j'}^{k'(0)} \\ j' = j \text{ and } z' > z \end{cases}$$

- *Analysis time + re-weighted Z statistic ordering (Z_w)*:

$$(j', t', k') \succ (j, t, k) \text{ if } \begin{cases} j' < j \text{ and } t' \in \mathcal{S}_{j'}^{k'(1)} \\ j' > j \text{ and } t \in \mathcal{S}_{j'}^{k'(0)} \\ j' = j \text{ and } z'_w > z_w \end{cases}$$

- *Statistical information ordering (N)*:

$$(j', t', k') \succ (j, t, k) \text{ if } \begin{cases} n_{j'}^{k'} < n_j^k \text{ and } t' \in \mathcal{S}_{j'}^{k'(1)} \\ n_{j'}^{k'} > n_j^k \text{ and } t \in \mathcal{S}_{j'}^{k'(0)} \\ n_{j'}^{k'} = n_j^k \text{ and } t' > t \end{cases}$$

Point Estimates and P-values

Define the following point estimates for θ given $(M, T, K) = (j, t, k)$:

- *Sample Mean* (MLE): $\hat{\theta} = \overline{X}_A - \overline{X}_B = t$
- *Bias adjusted mean* (BAM) $\breve{\theta}$: $E_T[\, T; \breve{\theta}\,] = t$
- *Median unbiased estimates* (MUE) $\tilde{\theta}_o$:
  $P[\,(M, T, K) \succ_o (j, t, k); \tilde{\theta}_o\,] = \frac{1}{2}$

For $H_0 : \theta = \theta_0$, define a P-value under imposed ordering $O = o$:

- $p\text{-value}_o = P[\,(M, T, K) \succ_o (j, t, k); \theta_0\,]$

Statistics as Usual

- We frequently use different orderings of the outcome space in order to carry out tests and compute point, interval estimates
  - Wald vs. Score vs. Likelihood Ratio
- Seek as reliable and precise inference as possible
- Desirable properties in sequential setting enumerated by Emerson, Jennison and Turnbull, and others

1. Statistical Efficiency of Adaptation
2. Complete Inference after Adaptation
   - Inference for Pre-specified Design
   - Inference after Unplanned Adaptation
3. Evaluating Inferential Methods
4. Additional Issues
5. Adaptive Time-to-event Setting

Unplanned Adaptation: Motivation

- Motivation
  - Flexibility to modify design (e.g., sample size / power) based on external information
    - ⋆ If truly external (independent), no adjustment to inference needed, but difficult to prove interim data had no role?
  - Flexibility to adapt utilizing information on additional endpoints
- Worth potential losses in reliability, efficiency due to lack of planning?
  - (Plus logistical challenges inherent to all adaptive designs)

Testing versus Estimation

- Many methods to control type I error rate in presence of unplanned adaptation
  - All equivalent to conditional error approach (J+T 2003, Proschan 2009)
- Limited research on estimation after adaptive hypothesis test
  - Exploration of absolute bias of MLE
    - ⋆ As high as 40% of SD of first-stage sample mean in 2-stage setting (Brannath et al. 2006)
  - Extension of repeated confidence intervals
  - Inversion of conditional error testing approach

Brannath, Mehta, and Posch (BMP) 2009

- Outcomes ordered according to smallest level of significance μ for which a conditional-error based adaptive hypothesis test of $H_0 : \theta = \theta'$ would be rejected:

$$(j', t', k', t'_h) \succ_{\theta', GSD} (j, t, k, t_h) \text{ if}$$

$$\mu(j', t', k', t'_h; \theta', GSD) < \mu(j, t, k, t_h; \theta', GSD)$$

- Depends on θ', interim estimate t_h, and original GSD
- But does not depend on what sampling plan we would have chosen had other interim data been observed

Gao, Liu, and Mehta (2012)

- More intuitive derivation of approach to invert conditional error-based tests
- Compute stage-wise ordered p-value of "backward image" of observed test statistic
- Backward image is statistic in outcome space of originally planned design with same stage-wise p-value (conditional on interim estimate) as in adaptively chosen future sampling plan
- Appears to be two-sided generalization of BMP approach

1 Statistical Efficiency of Adaptation
2 Complete Inference after Adaptation
- Inference for Pre-specified Design
- Inference after Unplanned Adaptation

3 Evaluating Inferential Methods
4 Additional Issues
5 Adaptive Time-to-event Setting

Assessing Reliability and Precision of Inference

- Confidence sets
  - True intervals
    - ⋆ If $P[(M, T, K) \succ_o (j, t, k); \theta]$ increases in θ for each (j, t, k) (proof found for sample mean ordering)
    - ⋆ Otherwise, negligible effects on coverage
  - Consistency with hypothesis test
    - ⋆ Requires same ordering for decisions, P-values, intervals
  - Shorter expected length

Assessing Reliability and Precision of Inference

- Point estimates
  - Low bias, variance, mean squared error (MSE)
- P-values
  - High probabilities of falling below important thresholds
    - ⋆ e.g., $0.025^2 = 0.000625$ to potentially approximate statistical strength of evidence of two independent studies

Approach to Evaluating Inferential Methods

- Estimates derived in iterative search by numerically integrating several group sequential densities
  - Densities convolutions of normals and truncated normals
  - Difficult to come up with analytic results on relative behavior
  - Resort to Monte Carlo simulation
- Develop extensive comparison framework to evaluate methods
  - 10,000 simulated trials under a range of treatment effects across a variety of adaptive sampling plans

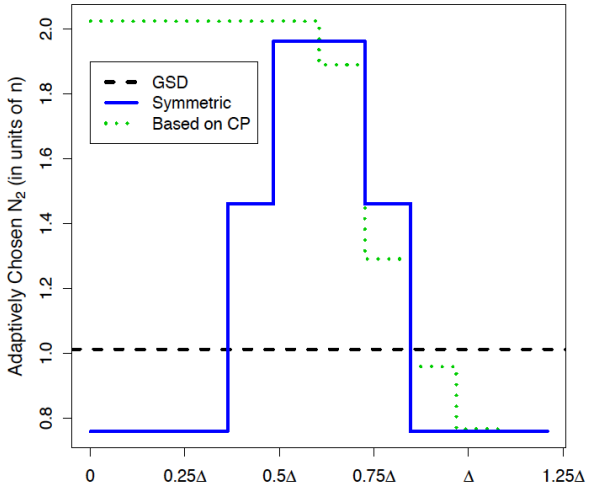
Comparison Framework

Pre-specified adaptive tests of $H_0 : \theta = 0$ against one-sided alternative $\theta > 0$ with $\alpha = 0.025$, power β at $\theta = \Delta$, with varying:

- Degree of early conservatism (reference OF or Pocock GSD)
- Symmetry of early stopping (symmetric or only for superiority)
- Power at Δ (80% to 97.5%)
- Maximum number of analyses (2, 3, or 4)
- Timing of adaptation (25% to 75% of original N_J)
- Maximum allowable sample size (25% to 100% increase)
- Rule for determining final sample size (symmetric or conditional-power based)

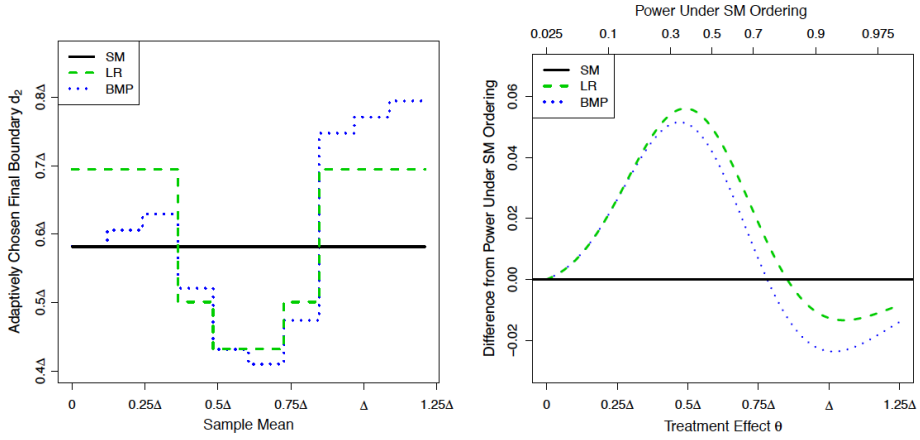
Adaptively Chosen Sample Size

- Example of symmetric and CP-based N_2 functions



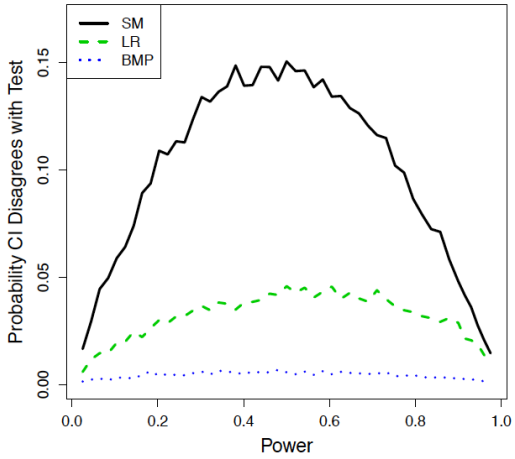
Differences in Boundaries and Power

- Comparing testing based on different orderings of outcome space (OF reference, symmetric N_J rule, 100% maximal increase)...



Avoiding Inconsistent Inference

- Should use same ordering for testing as for estimation



Confidence Intervals: Correct Coverage

- Standard error of CI coverage with 10,000 simulations: 0.0022

	OF Reference GSD				Pocock Reference GSD			
Power	Naive	SM	LR	BMP	Naive	SM	LR	BMP
				Symmetric N_J function, up to 50% Increase				
0.025	0.9442	0.9455	0.9449	0.9462	0.9425	0.9484	0.9485	0.9481
0.500	0.9314	0.9507	0.9488	0.9507	0.9458	0.9507	0.9504	0.9507
0.900	0.9402	0.9493	0.9478	0.9476	0.9350	0.9465	0.9467	0.9466
				CP-based N_J function, up to 100% Increase				
0.025	0.9428	0.9494	0.9497	0.9494	0.9441	0.9502	0.9508	0.9505
0.500	0.9181	0.9462	0.9469	0.9466	0.9355	0.9461	0.9476	0.9462
0.900	0.9291	0.9501	0.9501	0.9501	0.9365	0.9494	0.9489	0.9496

Estimates: Median-unbiased

- SE of probability exceeds MUE with 10,000 simulations: 0.005

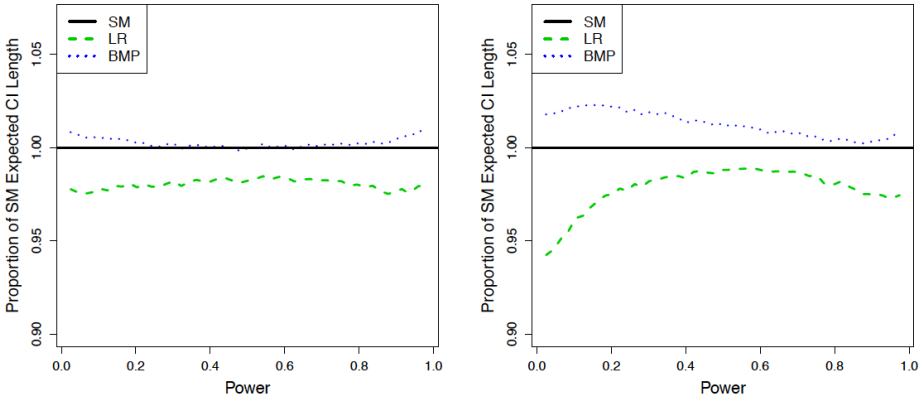
	OF Reference GSD			Pocock Reference GSD		
Power	SM	LR	BMP	SM	LR	BMP
			Symmetric N_J function, up to 100% Increase			
0.0250	0.4956	0.4993	0.4960	0.4983	0.4986	0.4960
0.5000	0.5082	0.5076	0.5081	0.5100	0.5093	0.5095
0.9000	0.5019	0.5006	0.4970	0.5034	0.5028	0.5011
			CP-based N_J function, up to 100% Increase			
0.0250	0.4975	0.4997	0.4958	0.5032	0.5035	0.5025
0.5000	0.5079	0.5075	0.5064	0.5027	0.5027	0.5045
0.9000	0.5001	0.4981	0.5050	0.5105	0.5099	0.5094

Results: Naive Inference

- MLE substantially higher bias than adjusted estimates at all but intermediate effects and higher MSE (up to 40%) across nearly all designs and effects considered
- Naive 95% CIs lack exact coverage, typically 92-93% coverage, occasionally near 90%
- Performance may be worse with more complex multistage designs

Comparing Confidence Intervals: Example

- Reference OF design, symmetric (left) or CP-based (right) N_J function, up to 50% increase, $J = 2$

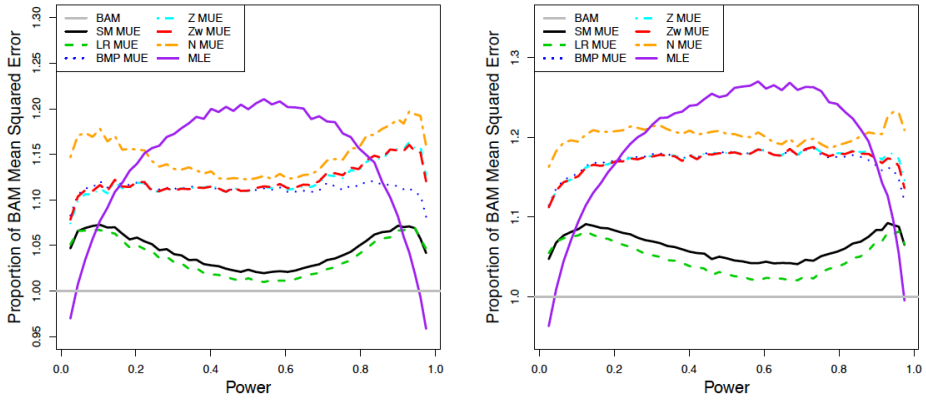


Comparing Confidence Intervals: Trends

- Likelihood ratio ordering shorter expected CI length across nearly all designs and treatment effects studied
  - $\sim 1 - 10\%$ shorter, depending on setting
  - Margin increases with greater potential inflation of N_J
  - Margin slightly larger for CP-based than symmetric N_J function
- Sample mean slightly superior ($\sim 1 - 3\%$) to BMP in some settings, similar in others

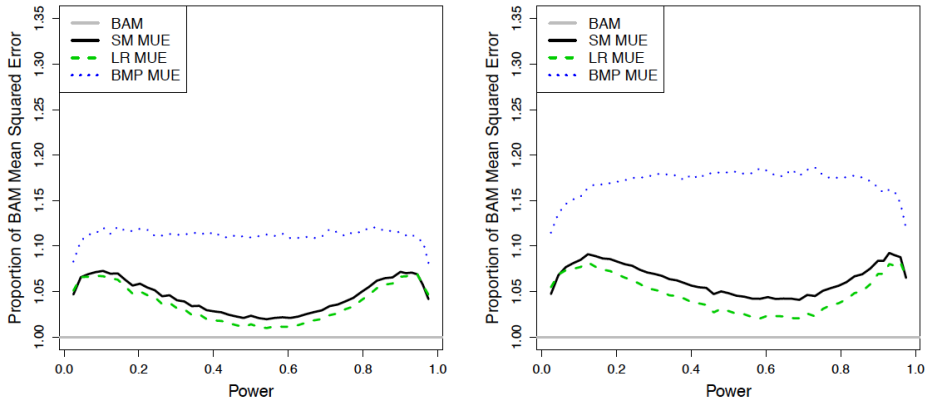
Comparing Point Estimates: Example

- Reference Pocock design, symmetric (left) or CP-based (right) N_J function, up to 100% increase, $J = 2$



Comparing Point Estimates: Example

- Reference Pocock design, symmetric (left) or CP-based (right) N_J function, up to 100% increase, $J = 2$

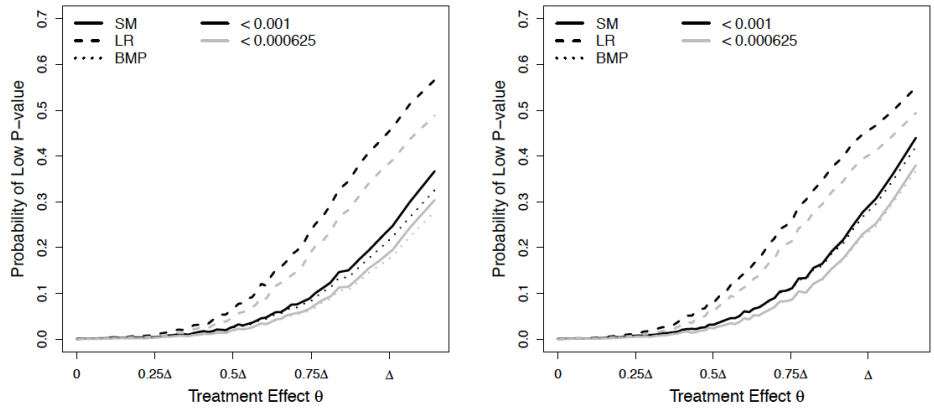


Comparing Point Estimates: Trends

- Bias adjusted mean best MSE across nearly all designs and treatment effects considered
  - $\sim 1 - 20\%$ lower, depending on setting and comparator
  - Margin increases with N_J inflation, CP-based adaptation
  - Lower bias at extreme effects, variance at intermediate effects
  - All CIs observed to always contain BAM
- SM, LR MUEs up to 15% lower MSE than BMP MUE
- LR MUE slightly superior ($\sim 1 - 3\%$) to SM MUE in some settings, similar in others

Comparing *P*-values: Example

- Reference OF (left) or Pocock (right) design, CP-based N_J function, up to 50% increase, $J = 2$



Comparing P-values: Trends

- Likelihood ratio ordering tends to demonstrate greater probabilities of potentially "pivotal" P-values
  - Up to $\sim 20\%$ greater (on absolute scale), depending on setting
  - Margin increases with greater N_J inflation, CP-based adaptation
  - Margin larger for tests derived from OF reference designs
- Sample mean modestly superior (up to $\sim 10\%$ on absolute scale) to BMP in most settings, similar in others

Summary and Conclusions

- Bias adjusted mean most reliable and precise point estimate
- Likelihood ratio ordering CIs and P-values behaved best
- Margins increase with N_J inflation, CP-based N_J function
- Qualitative differences persist varying many design parameters
  - Quantitative differences decrease for early, late adaptations
- MLE and inference using other orderings poor relative behavior

Cost of Planning not to Plan

- Most proposed adaptations could be pre-specified at design stage
- Substantial cost of failing to plan ahead and resorting to conditional error-based (BMP) estimation
  - Large increase (up to 20%) in MSE of point estimate
  - Modest increase (up to 10%) in expected CI length
  - Large decrease (up to 20%) in probability of pivotal P-value
  - Cost is largest for typically proposed adaptation rules
  - Due to inversion of conditional error tests or stage-wise ordering of backward image?
- BMP inference has reasonable behavior if needed

1 Statistical Efficiency of Adaptation

2 Complete Inference after Adaptation
- Inference for Pre-specified Design
- Inference after Unplanned Adaptation

3 Evaluating Inferential Methods

4 Additional Issues

5 Adaptive Time-to-event Setting

Case Study: An Antidepressant in MDD

- Randomized placebo-controlled clinical trial to study safety and effectiveness of novel antidepressant in major depressive disorder
- Primary outcome is 50% improvement at 8 weeks in Hamilton depression rating scale
- 30% response rate expected on placebo
- 10% improvement on treatment considered minimal clinically important difference

Case Study: An Antidepressant in MDD

Candidate designs:

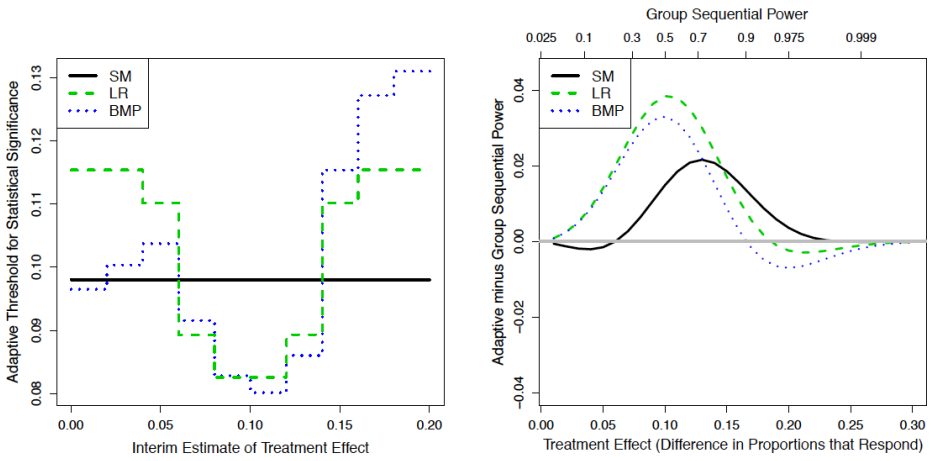
- Fixed sample design with 176 participants per arm
  - $\alpha = 2.5\%$ type I error, $\beta = 90\%$ power at $\theta = 0.165$, threshold for statistical significance of 10%
- Two-analysis O'Brien and Fleming and Pocock group sequential designs with same α, β, significance threshold
- Adaptive designs derived from these GSDs, using symmetric or conditional power-based rules

Statistical *versus* Clinical Significance

- Goal of RCTs not statistical significance but instead "statistically reliable evaluation regarding whether the experimental intervention is safe and provides clinically meaningful benefit." (Fleming 2006)
- Yet adaptation often proposed to increase conditional power presuming treatment effects below the MCID
- Threshold for statistical significance on scale of estimated treatment effect varies greatly under LR, BMP orderings
  - May fall below MCID: ranges from 8.0% to 13.2%

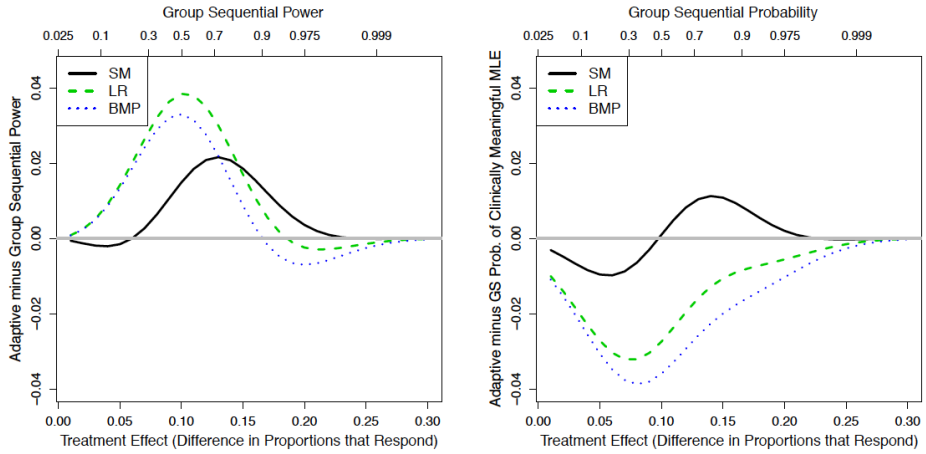
Statistical *versus* Clinical Significance

- Boundary differences result in power differences (OF reference, symmetric N_J rule, 50% maximal increase)...



Statistical *versus* Clinical Significance

- Consider success as statistical *and clinical* ($> 10\%$) significance:



Maintaining Confidentiality

- Maintaining confidentiality protects trial integrity
- Additional challenges in conduct of adaptive trial
  - Sample size may be function of interim estimate:

$$N_2(\hat{\theta}_1) = \left(\frac{\frac{d_2^0 \, n_2^0 - \hat{\theta}_1 \, n_1}{\sqrt{n_2^0 - n_1}} - \sqrt{V} \, \Phi^{-1}(0.1)}{\hat{\theta}_1} \right)^2 + n_1$$

  - Potential unblinding through new recruitment targets
    - ★ Example: New $N_2 = 227$ allows approximation of 13% estimate
  - Less likely with only few possible final sample sizes

Maintaining Confidentiality

Possible approaches if knowledge of adaptively chosen sample size and adaptation rule allows reasonably precise estimate of interim effect?

- Blind trial investigators involved in treatment, outcome assessment to new sample size
- Blind trial investigators to Adaptive Charter (which describes adaptation rule)
- Rely on unplanned adaptation by DMC
  - Too much to ask of DMC? Will require sponsor input/knowledge regardless...

Logistical and Ethical Issues

- Increased effort in planning, protocol development, monitoring
  - FDA Draft Guidance: "added complexities... call for more detailed documentation"
  - SAP must include "summary of each adaptation and its impact upon critical statistical issues"
- Ethics of weighting subjects differently
  - And should weighted or unweighted estimate be reported?
- Allow even greater bias knowing crude estimates will be reported in journals/labeling, interpreted as reliable

Additional Challenges: Summary

- Relative behavior of LR, BMP orderings, adaptive designs in general suffer when considering statistical *and clinical* significance
- Important added logistical and ethical challenges in design and conduct
- In many cases, these considerations alone may render adaptive design inappropriate

Summary and Conclusions

- Pre-specified adaptation attains minor efficiency gain ($< 0.5\%$)
  - Efficient designs differ qualitatively from those in literature
  - Should evaluate important operating characteristics and modifying/comparing candidate designs
- Estimation methods after adaptive test developed and evaluated
  - Avoid using naive CIs and MLE
  - Bias adjusted mean, LR or SM ordering better behavior with respect to important measures of reliability, precision
  - Failing to pre-specify (BMP) comes with meaningful cost

Editorial

- Carefully compare candidate designs before deciding to adapt
- Potential gains in flexibility, efficiency through sample size adaptation likely not worth added interpretability, logistical challenges in most settings
- Possibly more promise with adaptive subgroup selection (e.g., with a pre-specified, clearly defined targeted subset expected to benefit more – see Rosenblum research)

1 Statistical Efficiency of Adaptation
2 Complete Inference after Adaptation
- Inference for Pre-specified Design
- Inference after Unplanned Adaptation

3 Evaluating Inferential Methods
4 Additional Issues
5 Adaptive Time-to-event Setting

Special Topic: Adaptive Time-to-event Setting

Adaptive Sample Size Re-estimation with Time to Event Endpoints

Scott S. Emerson, M.D., Ph.D.
William J.H. Koh
Department of Biostatistics
University of Washington

Summer Institute in Statistics for Clinical Research
July 1, 2015

Special Topic: Adaptive Time-to-event Setting

Abstract

A great many confirmatory phase 3 clinical trials have as their primary endpoint a comparison of the distribution of time to some event (e.g., time to death or progression free survival).

- The most common statistical analysis models include the logrank test and/or the proportional hazards regression model.
- Just as commonly, the true distributions do not satisfy the proportional hazards assumption.

Providing users are aware of the nuances of those methods, such departures need not preclude the use of those analytic techniques any more than violations of the location shift hypothesis precludes the use of the t test.

In this talk I discuss some aspects of the analysis of censored time to event data that must be carefully considered in sequential and adaptive sampling. In particular, I discuss the how the changing censoring distribution during a sequential trial affects the analysis of distributions with crossing hazards and crossing survival curves.

Special Topic: Adaptive Time-to-event Setting

Overall Goal: "Drug Discovery"

- More generally
  - a therapy / preventive strategy or diagnostic / prognostic procedure
  - for some disease
  - in some population of patients

- A ***sequential***, ***adaptive*** series of experiments to establish
  - Safety of investigations / dose (phase 1)
  - Safety of therapy (phase 2)
  - Measures of efficacy (phase 2)
    - Treatment, population, and outcomes
  - Confirmation of efficacy (phase 3)
  - Confirmation of effectiveness (phase 3, post-marketing)

Special Topic: Adaptive Time-to-event Setting

Science and Statistics

- Statistics is about science
  - (Science in the broadest sense of the word)

- Science is about proving things to people
  - (The validity of any proof rests solely on the willingness of the audience to believe it)

- What do we need to consider as we strive to meet the burden of proof with adaptive modification of a RCT design?

- Does time to event data affect those issues?
  - Short answer: No, UNLESS subject to censoring
  - So, true answer: Yes.

Special Topic: Adaptive Time-to-event Setting

Design: Distinctions without Differences

- There is no such thing as a "Bayesian design"
- Every RCT design has a Bayesian interpretation
  - (And each person may have a different such interpretation)
- Every RCT design has a frequentist interpretation
  - (In poorly designed trials, this may not be known exactly)
- In this talk I focus on the use of both interpretations
  - Phase 2: Bayesian probability space
  - Phase 3: Frequentist probability space
  - Entire process: Both Bayesian and frequentist optimality criteria

Special Topic: Adaptive Time-to-event Setting

Application to Drug Discovery

- We consider a population of candidate drugs
- We use RCT to "diagnose" truly beneficial drugs
- Use both frequentist and Bayesian optimality criteria
  - Sponsor:
    - High probability of adopting a beneficial drug (frequentist power)
  - Regulatory:
    - Low probability of adopting ineffective drug (freq type 1 error)
    - High probability that adopted drugs work (posterior probability)
  - Public Health (frequentist sample space, Bayes criteria)
    - Maximize the number of good drugs adopted
    - Minimize the number of ineffective drugs adopted

Special Topic: Adaptive Time-to-event Setting

Frequentist vs Bayesian: Bayes Factor

- Frequentist and Bayesian inference truly complementary
  - Frequentist: Design so the same data not likely from null / alt
  - Bayesian: Explore updated beliefs based on a range of priors
- Bayes rule tells us that we can parameterize the positive predictive value by the type I error and prevalence
  - Maximize new information by maximizing Bayes factor
  - With simple hypotheses:

$$PPV = \frac{power \times prevalence}{power \times prevalence \quad + \quad type\ I\ err \times (1 - prevalence)}$$

$$\frac{PPV}{1 - PPV} = \frac{power}{type\ I\ err} \times \frac{prevalence}{1 - prevalence}$$

$$posterior\ \ odds = Bayes\ Factor \times prior\ \ odds$$

Special Topic: Adaptive Time-to-event Setting

Adaptive Sampling: General Case

- At each interim analysis, possibly modify statistical or scientific aspects of the RCT
- Primarily statistical characteristics
  - Maximal statistical information (UNLESS: impact on MCID)
  - Schedule of analyses (UNLESS: time-varying effects)
  - Conditions for stopping (UNLESS: time-varying effects)
  - Randomization ratios (UNLESS: introduce confounding)
  - Statistical criteria for credible evidence
- Primarily scientific characteristics
  - Target patient population (inclusion, exclusion criteria)
  - Treatment (dose, administration, frequency, duration)
  - Clinical outcome and/or statistical summary measure

Special Topic: Adaptive Time-to-event Setting

FDA Guidance on Adaptive RCT Designs

- Distinctions by role of trial
  - "Adequate and well-controlled" (Kefauver-Harris wording)
  - "Exploratory"
- Distinctions by adaptive methodology
  - "Well understood"
    - Fixed sample design
    - Blinded adaptation
    - Group sequential with pre-specified stopping rule
  - "Less well understood"
    - "Adaptive" designs with a prospectively defined opportunity to modify specific aspects of study designs based on review of unblinded interim data
  - "Not within scope of guidance"
    - Modifications to trial conduct based on unblinded interim data that are not prospectively defined

Special Topic: Adaptive Time-to-event Setting

FDA Concerns

- Statistical errors: Type 1 error; power
- Bias of estimates of treatment effect
  - Definition of treatment effect
  - Bias from multiplicity
- Information available for subgroups, dose response, secondary endpoints
- Operational bias from release of interim results
  - Effect on treatment of ongoing patients
  - Effect on accrual to the study
  - Effect on ascertainment of outcomes

Special Topic: Adaptive Time-to-event Setting

Control of Type 1 Errors

- Proschan and Hunsberger (1995)
  - Adaptive modification of RCT design at a single interim analysis can more than double type 1 error unless carefully controlled
- Those authors describe adaptations to maintain experimentwise type I error and increase conditional power
  - Must prespecify a conditional error function

$$\int_{-\infty}^{\infty} A(z)\,\phi(z)\,dz = \alpha.$$

  - Often choose function from some specified test

$$A(z) = Pr_{\delta=0}(Z_2 \geq \Phi^{-1}(1-\alpha)\,|\,\tilde{Z}_1 = z, \tilde{n}_2 = n_2 - n_1),$$

  - Find critical value to maintain type I error

$$Pr_{\delta=0}\left(Z_2^{\star} \geq c(\tilde{n}_2^{\star}, \tilde{z}_1)\,|\,\tilde{n}_2^{\star}(\tilde{z}_1)\right) = A(\tilde{z}_1).$$

Special Topic: Adaptive Time-to-event Setting

Alternative Approaches

- Combining P values (Bauer & Kohne, 1994)
  - Based on R.A. Fisher's method
  - Extended to weighted combinations
- Cui, Hung, and Wang (1999)
  - Maintain conditional error from pre-specified design
- Self-designing Trial (Fisher, 1998)
  - Combine arbitrary test statistics from sequential groups using weighting of groups pespecified "just in time"

Special Topic: Adaptive Time-to-event Setting

Data at *j*-th Analysis: Immediate Outcome

- Subjects accrued at different stages are independent
- Statistics as weighted average of data accrued between analyses

At kth interim analysis	Incremental	Cumulative
Sample size (stat info)	N_k^*	$N_k = N_1^* + \cdots + N_k^*$
Baseline data	$\vec{X}_k^*$	$\vec{X}_k = (\vec{X}_1^*, \ldots, \vec{X}_k^*)$
1° outcome data	$\vec{Y}_k^*$	$\vec{Y}_k = (\vec{Y}_1^*, \ldots, \vec{Y}_k^*)$
2° outcome data	$\vec{W}_k^*$	$\vec{W}_k = (\vec{W}_1^*, \ldots, \vec{W}_k^*)$

Using $N_k^*, \vec{X}_k^*, \vec{Y}_k^*$:

Estimated treatment effect	$\hat{\theta}_k^* = \hat{\theta}_k^*(N_k^*, \vec{X}_k^*, \vec{Y}_k^*)$	$\hat{\theta}_k = \frac{\sum_{j=1}^{k} N_j^* \hat{\theta}_j^*}{N_k}$
Normalized Z statistic	Z_k^*	$Z_k = \frac{\sum_{j=1}^{k} \sqrt{N_j^*} Z_j^*}{\sqrt{N_k}}$
Fixed sample P value	P_k^*	

Special Topic: Adaptive Time-to-event Setting

Conditional Distn: Immediate Outcomes

- Sample size N_j^* and parameter θ_j can be adaptively chosen based on data from prior stages 1,…,*j-1*
  - (Most often we choose $\theta_j = \theta$ with immediate data)

$$\hat{\theta}_j^* \mid N_j^* \sim N\left(\theta_j, \frac{V(\theta_j)}{N_j^*}\right)$$

$$Z_j^* \mid N_j^* \sim N\left(\frac{\hat{\theta}_j - \theta_{0j}}{\sqrt{V(\theta_j)/N_j^*}}, 1\right)$$

$$P_j^* \mid N_j^* \overset{H_0}{\sim} U(0,1).$$

Conditional distributions are totally independent under the null hypothesis

Special Topic: Adaptive Time-to-event Setting

Estimands by Stage: Time to Event

- Most often we choose $\theta_j = \theta$ with immediate data
- In time to event data, a common treatment effect across stages is reasonable under some assumptions
  - Strong null hypothesis (exact equality of distributions)
  - Strong parametric or semi-parametric assumptions
- The most common methods of analyzing time to event data will often lead to varying treatment effect parameters across stages
  - Proportional hazards regression with non proportional hazards data
  - Weak null hypotheses of equality of summary measures (e.g., medians, average hazard ratio)

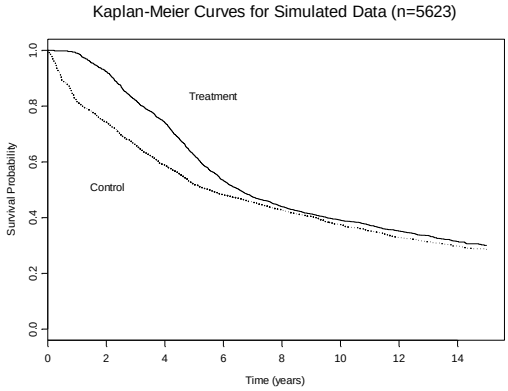
Special Topic: Adaptive Time-to-event Setting

Hypothetical Example: Setting

- Consider survival with a particular treatment used in renal dialysis patients
- Extract data from registry of dialysis patients
- To ensure quality, only use data after 1995
  - Incident cases in 1995: Follow-up 1995 – 2002 (8 years)
  - Prevalent cases in 1995: Data from 1995 - 2002
    - Incident in 1994: Information about 2nd – 9th year
    - Incident in 1993: Information about 3rd – 10th year
    - …
    - Incident in 1988: Information about 8th – 15th year

Special Topic: Adaptive Time-to-event Setting

Hypothetical Example: KM Curves



Special Topic: Adaptive Time-to-event Setting

Who Wants To Be A Millionaire?

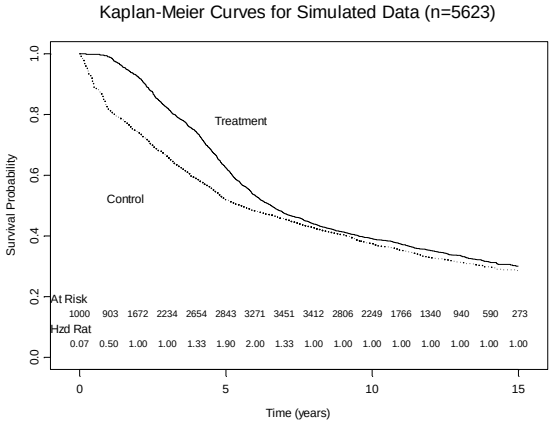
- Proportional hazards analysis estimates a **Treatment : Control** hazard ratio of

  A: 2.07 (logrank P = .0018)
  B: 1.13 (logrank P = .0018)
  C: 0.87 (logrank P = .0018)
  D: 0.48 (logrank P = .0018)

  – Lifelines:
    - 50-50? Ask the audience? Call a friend?

Special Topic: Adaptive Time-to-event Setting

Hypothetical Example: KM Curves



Special Topic: Adaptive Time-to-event Setting

Who Wants To Be A Millionaire?

Proportional hazards analysis estimates a **Treatment : Control** hazard ratio of

B: 1.13 (logrank P = .0018)

The weighting using the risk sets made no scientific sense

- Statistical precision to estimate a meaningless quantity is meaningless

Special Topic: Adaptive Time-to-event Setting

Partial Likelihood Based Score

- Logrank statistic

$$U(\beta)=\frac{\partial}{\partial\beta}\log L(\beta)=\sum_{i=1}^{n}D_i\left[X_i-\frac{\sum_{j:T_j\geq T_i}X_j\exp\{X_j\beta\}}{\sum_{j:T_j\geq T_i}\exp\{X_j\beta\}}\right]$$

$$=\sum_t\left[d_{1t}-\frac{n_{1t}e^{\beta}}{n_{0t}+n_{1t}e^{\beta}}(d_{0t}+d_{1t})\right]$$

$$=\sum_t\frac{n_{0t}n_{1t}}{n_{0t}+n_{1t}}\left[\hat{\lambda}_{1t}-e^{\beta}\hat{\lambda}_{0t}\right]$$

Special Topic: Adaptive Time-to-event Setting

Weighted Logrank Statistics

- Choose additional weights to detect anticipated effects

$$W(\beta)=\sum_t w(t)\frac{n_{0t}n_{1t}}{n_{0t}+n_{1t}}\left[\hat{\lambda}_{1t}-e^{\beta}\hat{\lambda}_{0t}\right]$$

$$n_{kt}=N_k\times\Pr(T\geq t,Cens\geq t)\overset{ind}{=}N_kS_k(t)\times\Pr(Cens\geq t)$$

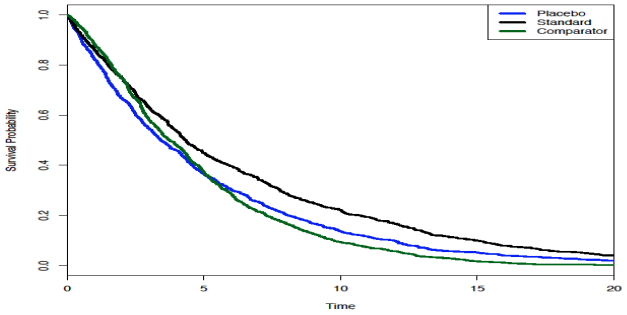
$G^{\rho\gamma}$ Family of weighted logrank statistics :

$$w(t)=\left[\hat{S}_{\bullet}(t)\right]^{\rho}\left[1-\hat{S}_{\bullet}(t)\right]^{\gamma}$$

Special Topic: Adaptive Time-to-event Setting

Impact on Noninferiority Trials

- Weak null hypothesis is of greatest interest
  - Standard superior to placebo
  - Comparator (on average) equivalent to placebo



Special Topic: Adaptive Time-to-event Setting

Conditional Distn: Immediate Outcomes

- Sample size N_j^* and parameter θ_j can be adaptively chosen based on data from prior stages 1,…,*j-1*
  - (Most often we choose $\theta_j = \theta$ with immediate data)

$$\hat{\theta}_j^* \mid N_j^* \sim N\left(\theta_j, \frac{V(\theta_j)}{N_j^*}\right)$$

$$Z_j^* \mid N_j^* \sim N\left(\frac{\hat{\theta}_j - \theta_{0j}}{\sqrt{V(\theta_j)/N_j^*}}, 1\right)$$

$$P_j^* \mid N_j^* \overset{H_0}{\sim} U(0,1).$$

Conditional distributions are totally independent under the null hypothesis

Special Topic: Adaptive Time-to-event Setting

Protecting Type I Error

- Test based on weighted averages of incremental test statistics
  - Allow arbitrary weights W_j specified by stage *j-1*

$$Z = \frac{\sum_{k=1}^{J} \sqrt{W_k}\ Z_k^*}{\sqrt{\sum_{k=1}^{J} W_j}} \quad \overset{\bigcap_{k=1}^{J} H_{0j}}{\sim} \quad N(0,1)$$

$$Z = \frac{\sum_{k=1}^{J} \sqrt{W_k}\ \Phi^{-1}\left(1-P_k^*\right)}{\sqrt{\sum_{k=1}^{J} W_j}} \quad \overset{\bigcap_{k=1}^{J} H_{0j}}{\sim} \quad N(0,1)$$

Special Topic: Adaptive Time-to-event Setting

Complications: Longitudinal Outcomes

- Bauer and Posch (2004) noted that in the presence of incomplete data, partially observed outcome data may be informative of the later contributions to test statistics
- We need to make distinctions between
  - Independent subjects accrued at different stages
  - Statistical information about the primary outcome available at different analyses
- Owing to delayed observations, contributions to the primary test statistic at the *k*-th stage may come from subjects accrued at prior stages
  - Baseline and secondary outcome data available at prior analyses on those subject may inform the value of future data

Special Topic: Adaptive Time-to-event Setting

Data at *j*-th Analysis: Delayed Outcome

- Subjects accrued at different stages are independent
- Some data is "missing"

At *k*th interim analysis	Incremental	Cumulative
Sample size (stat info)	N_k^*	$N_k = N_1^* + \cdots + N_k^*$
Baseline data	$\vec{X}_k^*$	$\vec{X}_k = (\vec{X}_1^*, \ldots, \vec{X}_k^*)$
1° outcome data (msng, observed)	$\vec{Y}_k^{*M}, \vec{Y}_k^{*O}$	$\vec{Y}_k^M, \vec{Y}_k^O$
2° outcome data	$\vec{W}_k^*$	$\vec{W}_k = (\vec{W}_1^*, \ldots, \vec{W}_k^*)$
Estimated treatment effect	$\hat{\theta}_k^* = \hat{\theta}_k^*(N_k^*, \vec{X}_k^*, \vec{Y}_k^{*O}, \vec{Y}_{k-1}^M)$	$\hat{\theta}_k = \frac{\sum_{j=1}^k N_j^* \hat{\theta}_j^*}{N_k}$
Normalized Z statistic	Z_k^*	$Z_k = \frac{\sum_{j=1}^k \sqrt{N_j^*} Z_j^*}{\sqrt{N_k}}$
Fixed sample P value	P_k^*	

Special Topic: Adaptive Time-to-event Setting

Major Problem: Delayed Outcome

- When sample size N_j^* and parameter θ_j adaptively chosen based on data from prior stages 1,…,*j-1*, some aspect of the "future" contributions may already be known

At *k*th interim analysis	Incremental	Cumulative
Sample size	$N_k^* = N_k^*(N_{k-1}, \vec{X}_{k-1}, \vec{W}_{k-1}, \vec{Y}_{k-1}^{*O}, \vec{Y}_{k-2?}^M)$	N_k
Estimated treatment effect	$\hat{\theta}_k^* = \hat{\theta}_k^*(N_k^*, \vec{X}_k^*, \vec{Y}_k^{*O}, \vec{Y}_{k-1}^M)$	$\hat{\theta}_k = \frac{\sum_{j=1}^k N_j^* \hat{\theta}_j^*}{N_k}$

Impact : (One statistici an's mean is another statistici an's variance)

$corr(\vec{Y}_k^{*M}, \vec{W}_k^*) \neq 0$ or $corr(\vec{Y}_k^{*M}, \vec{X}_k^*) \neq 0 \quad \Rightarrow \quad \hat{\theta}_k^* \mid N_k^*$ not indep of $\hat{\theta}_{k+1}^* \mid N_{k+1}^*$

$\hat{\theta}_k^* \mid N_k^*$ is potentiall y biased for θ_k and not approximat ely normal

Special Topic: Adaptive Time-to-event Setting

Impact of Adjusted Analyses

- Clearly, the assumptions of such approaches as CHW do not hold
  - The test statistic at the *k*-th analysis does not capture all of the information present in the data
- If we take the worst case assumption that the interim data has perfect information about the future we can "cherry-pick" the best analysis time
  - This can inflate the type 1 error substantially, depending upon how many censored subjects are present
- If one imagines that we would use the CHW adjustment that re-weights the data, even more damage can be done
  - We can upweight the highly random fluctuations in small amounts of information

Special Topic: Adaptive Time-to-event Setting

Potential Solutions

- Jenkins, Stone & Jennison (2010)
  - Only use data available at the *k*-th stage analysis
- Irle & Schaefer (2012)
  - Prespecify how the full *k*-th stage data will eventually contribute to the estimate of θ_k
- Magirr, Jaki, Koenig & Posch (2014, arXiv.org)
  - Assume worst case of full knowledge of future data and sponsor selection of most favorable P value

Special Topic: Adaptive Time-to-event Setting

Comments: Burden of Proof Dilemma

- There is a contradiction of standard practices when viewing the incomplete data
  - We would never accept the secondary outcomes as validated surrogates
  - But we feel that we must allow for the possibility that the secondary outcomes were perfectly predictive of the eventual data
- We are in some sense preferring mini-max optimality criteria over a Bayes estimator

Special Topic: Adaptive Time-to-event Setting

Comments: Impact on RCT Design

- The candidate approaches will protect the type 1 error, but the impact on power (and PPV) is as yet unclear
- Weighted statistics are not based on minimal sufficient statistics
  - But greatest loss in efficiency comes from late occurring adaptive analyses with large increases in maximal statistical information
  - Time to event will not generally have this
- The adaptation is based on imprecise estimates of the estimates that will eventually contribute to inference
- We may have to eventually either
  - Ignore some observed data (JS&S, I&S), or
  - Adjust for worst case multiple comparisons

Special Topic: Adaptive Time-to-event Setting

What if No Adjustment?

- Many methods for adaptive designs seem to suggest that there is no need to adjust for the adaptive analysis if there were no changes to the study design
- However, changes to the censoring distribution definitely affect
  - Distribution-free interpretation of the treatment effect parameter
  - Statistical precision of the estimated treatment effect
  - Type 1 error when testing a weak null (e.g., noninferiority)
- Furthermore, "less understood" analysis models prone to inflation of type 1 error when testing a strong null
  - Information growth with weighted log rank tests is not always proportional to the number of events

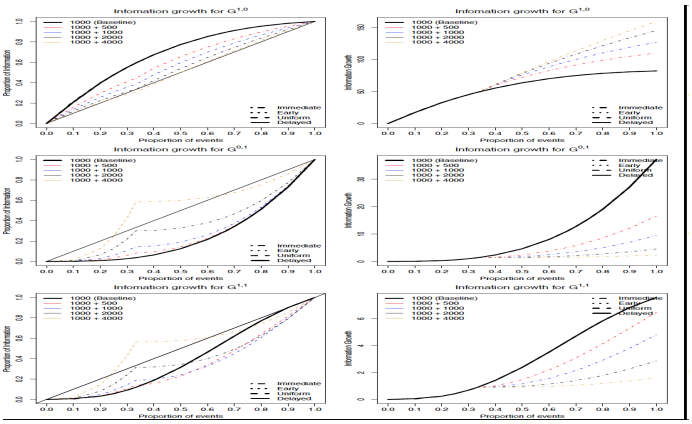
Special Topic: Adaptive Time-to-event Setting

"Intent to Cheat" Zone

- At interim analysis, choose range of interim estimates that lead to increased accrual of patients
- How bad can we inflate type 1 error when holding number of events constant?
- Logrank test under strong null: Not at all
- Weighted logrank tests: Up to relative increase of 20%
  - Sequela of true information growth depends on more than number of events
  - Power largely unaffected, so PPV decreases

Special Topic: Adaptive Time-to-event Setting

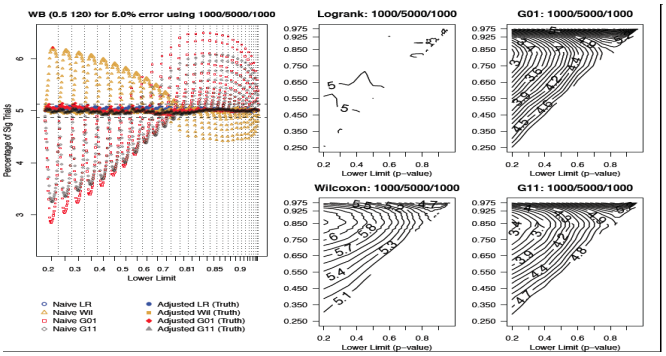
Information Growth with Adaptation



Special Topic: Adaptive Time-to-event Setting

Inflation of Type 1 Error

- Function of definition of the adaptation zone
  - Varies according to weighted log rank test



Special Topic: Adaptive Time-to-event Setting

Final Comments

- There is still much for us to understand about the implementation of adaptive designs
- Most often the "less well understood" part is how they interact with particular data analysis methods
  - In particular, the analysis of censored time to event data has many scientific and statistical issues
- How much detail about accrual patterns, etc. do we want to have to examine for each RCT?
- How much do we truly gain from the adaptive designs?
  - (Wouldn't it be nice if statistical researchers started evaluating their new methods in a manner similar to evaluation of new drugs?)

Special Topic: Adaptive Time-to-event Setting

Bottom Line

- There is no substitute for planning a study in advance
  - At Phase 2, adaptive designs may be useful to better control parameters leading to Phase 3
    - Most importantly, learn to take "NO" for an answer
  - At Phase 3, there seems little to be gained from adaptive trials
    - We need to be able to do inference, and poorly designed adaptive trials can lead to some very perplexing estimation methods
- **"Opportunity is missed by most people because it is dressed in overalls and looks like work."** -- Thomas Edison
- In clinical science, it is the steady, incremental steps that are likely to have the greatest impact.

Special Topic: Adaptive Time-to-event Setting

Really Bottom Line

"You better think (think)
about what you're
trying to do…"

-Aretha Franklin, "Think"