Interim Analysis

Michael Martens, PhD

November 10, 2022

Stopping a Clinical Trial Early

- Why stop a trial early?
  - Strong evidence of benefit or futility of novel treatment
  - Safety concerns found through toxicity monitoring
  - Logistical issues: Interruption of drug supply, lack of funding
  - Accrual problems
- Stopping early for established benefit or for futility needs to be built into the study design to support trial's validity
- One common approach is group sequential design: efficacy compared between treatments after each group of patients are enrolled

Early Treatment Comparisons

- How do we judge whether an early treatment difference is sufficiently large or small to justify early termination?
- General approach:
  - Primary research hypothesis is specified, e.g. null hypothesis of no treatment difference
  - Efficacy data are collected over time
  - Primary outcome assessed at interim analyses
  - At each interim analysis, a test statistic is constructed and compared to a stopping boundary
  - Stopping criteria (boundaries) are chosen to preserve operating characteristics of trial (type I error rate, power)

Early Treatment Comparisons

- This general structure applies to
  - t-tests comparing mean between two treatments
  - Z tests comparing two proportions
  - Log-rank tests for comparing two survival curves
  - Common regression models: ANCOVA/linear regression, logistic, generalized linear, Cox proportional hazards
  - Other likelihood-based tests
- Reference: Jennison and Turnbull, Group Sequential Methods with Applications to Clinical Trials

Comparing Two Means of Normal Outcomes

- Assume we want to compare a normal outcome between two treatment groups
- Let X_{ij} denote the outcome for patient j in treatment group i
- Assume $X_{ij} \sim N(\mu_i, \sigma^2)$
- Treatment effect: $\delta = \mu_1 - \mu_2$
- Null hypothesis: $H_0 : \delta = 0$
- Sample size in treatment group i at time t: n_{it}
- Sample mean in treatment group i at time t:

$$\bar{X}_{it} = \sum_{j=1}^{n_{it}} X_{ij}/n_{it}$$

Comparing Two Means

- Test statistic:

$$Z(t) = \frac{\bar{X}_{1t} - \bar{X}_{2t}}{\sqrt{\sigma^2/n_{1t} + \sigma^2/n_{2t}}}$$

- For large sample sizes, $Z(t) \sim N(\eta_t(\delta), 1)$, where

$$\eta_t(\delta) = \frac{\delta}{\sqrt{\sigma^2/n_{1t} + \sigma^2/n_{2t}}}$$

- Under $H_0 : \delta = 0$, $Z(t) \sim N(0, 1)$
- Decision rule: Reject H_0 if $|Z(t)| \geq b(t)$, where $b(t)$ is some critical value / boundary value which depends on t

Fixed Sample vs. Group Sequential Design

- **Fixed Sample Design**:
  - Collect observations from all patients
  - Analyze results once using all observations
  - Critical value of $z_{1-\alpha/2}$ used to produce type I error rate of α
- **Group Sequential Design**:
  - Collect $K > 1$ groups of observations during trial duration
  - After each new group of observations is obtained, conduct analysis and stop trial if stopping criteria are reached
  - Final sample size is a random variable; depends on when the trial stops

Group Sequential Design

- Reject H_0 the first time the boundary is crossed, i.e. first j where $|Z(t_j)| \geq b(t_j)$
- Using this strategy, we reject H_0 if

$$\{|Z(t_1)| \geq b(t_1)\} \text{ or } \{|Z(t_2)| \geq b(t_2)\} \text{ or } \cdots \text{ or } \{|Z(t_K)| \geq b(t_K)\}$$

- This is equivalent to the event

$$\cup_{j=1}^{K} \{|Z(t_j)| \geq b(t_j)\}$$

- Alternatively, we accept the null hypothesis whenever

$$\{|Z(t_1)| < b(t_1), \cdots, |Z(t_K)| < b(t_K)\}, \text{ or } \cap_{j=1}^{K} \{|Z(t_j)| < b(t_j)\}$$

- How to choose the boundary values $b(t_j)$?

Naive Analysis

- Suppose we reject H_0 if $|Z(t_j)| \geq z_{1-\alpha/2}$ at any interim analysis, i.e. use rejection boundaries from fixed sample design
- If $\alpha = 0.05$, the overall type I error rate would be

$$P_{H_0}(\text{reject } H_0) = P_{H_0}\{\cup_{j=1}^{K}[|Z(t_j)| \geq 1.96]\} > 0.05 \text{ for } K \geq 2$$

- Multiple testing issue is at play
- Effect of multiple (equally spaced) looks on type I error rate:

K	1	2	3	4	5	10	20
α	0.050	0.083	0.107	0.126	0.142	0.193	0.246

Boundary Selection for Proper Analysis

- Need to choose boundary values carefully to protect overall type I error rate
- Must satisfy $P\left(\cup_{j=1}^{K} \{|Z(t_j)| \geq b(t_j)\} \right) \leq \alpha$
- Two common boundary choices for equally spaced interim analyses (presented on Z scale):
  - **Pocock Design**
  - **O'Brien-Fleming Design**

Pocock Design

- Same boundary on Z statistic at each interim analysis: $b(t_1) = \cdots = b(t_K) = C_P(K, \alpha)$
- Pocock Design Decision Rule:
  - After group $j = 1, \cdots, K-1$, stop early and reject H_0 if

$$|Z(t_j)| \geq C_P(K, \alpha);$$

    else, enroll the next cohort of patients
  - After group K, reject H_0 if $|Z(t_K)| \geq C_P(K, \alpha)$; else, accept H_0
- Note that critical value C_P is higher than $z_{1-\alpha/2}$ and depends on K (increases as K does)

O'Brien-Fleming Design

- O'Brien-Fleming Decision Rule:
  - After group $j = 1, \cdots, K-1$, stop early and reject H_0 if

$$|Z(t_j)| \geq C_B(K, \alpha)\sqrt{K/j};$$

    else, enroll the next cohort of patients
  - After group K, reject H_0 if $|Z(t_K)| \geq C_B(K, \alpha)$;
    else, accept H_0
- Note that critical value C_B is higher than $z_{\alpha/2}$ and depends on K (increases as K does)
- Critical values are higher at early analyses, lower at later interim analyses

Boundaries for Pocock and O'Brien-Fleming Designs

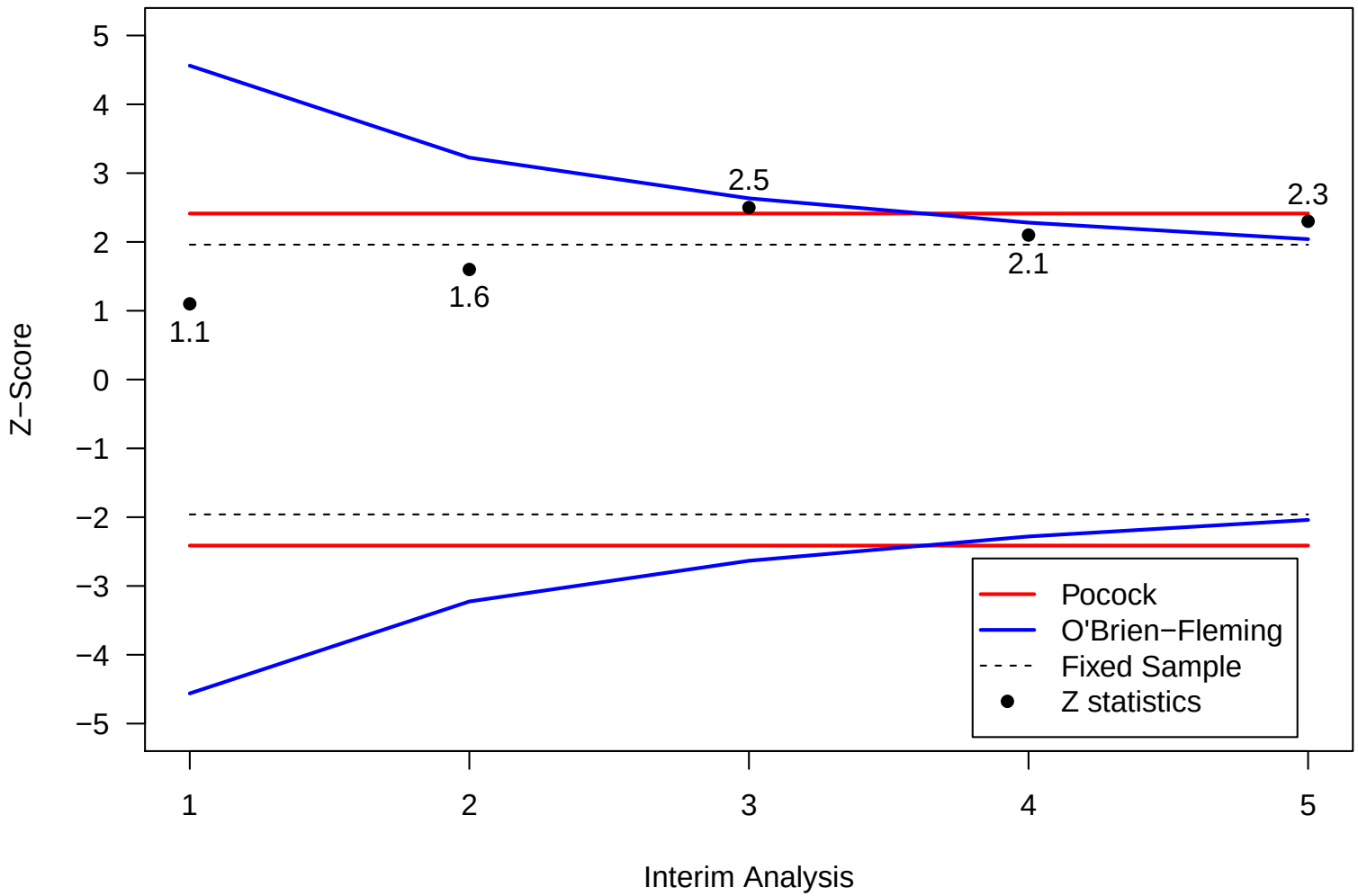
- Values for C_P and C_B can be found in Tables 2.1 and 2.3 of Jennison and Turnbull, Table 14.2 of Piantadosi
- Software can compute these for any specified α and K values:
  - SAS PROC SEQDESIGN
  - R function gsDesign in the "gsDesign" package

Example

- Group sequential analysis with 5 interim looks at the data, 5% type I error rate

Look	Pocock	O'Brien-Fleming
1	2.413	$2.040\sqrt{5/1} = 4.562$
2	2.413	$2.040\sqrt{5/2} = 3.226$
3	2.413	$2.040\sqrt{5/3} = 2.634$
4	2.413	$2.040\sqrt{5/4} = 2.281$
5	2.413	2.040

Example: Stopping boundaries with 5 interim looks, $\alpha = 5\%$



Sample Size Considerations

- Group sequential trials require larger maximum sample sizes than fixed sample-designed trials
- Max sample size for group sequential design, like a fixed sample design, is proportional to $(z_{1-\alpha/2} + z_{1-\beta})^2\sigma^2/\delta^2$
- Max sample size can be written in terms of a sample size inflation factor R and fixed sample design size n_F:

$$n_{GS} = R \cdot n_F$$

- Inflation Factors depend on
  - α
  - β (equivalently, power)
  - Stopping boundary shape (e.g. Pocock, O'Brien-Fleming)
  - Number of planned looks, K

Sample Size Considerations

- Inflation factors for Pocock and O'Brien-Fleming designs can be found in Tables 2.2 and 2.4 of Jennison and Turnbull
- Software can compute these for specified α, β, K values and boundary shapes:
  - SAS PROC SEQDESIGN
  - R function gsDesign in the "gsDesign" package

Example

- Suppose we want to test $H_0 : \mu_1 = \mu_2$ with type I error $\alpha = 0.05$ and power $1 - \beta = 0.9$ to detect $\delta = \mu_1 - \mu_2 = 1$, when the observations have a variance $\sigma^2 = 4$
- Fixed sample size:

$$n_F = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2\sigma^2}{\delta^2} = \frac{2(1.96 + 1.28)^2 4}{1} = 83.98$$

$$\approx 84 \text{ patients per arm}$$

Example

- Pocock design with 5 looks (equally spaced):
  - Inflation factor: $R_P(5, 0.05, 0.1) = 1.207$
  - Sample size needed:

$$n_{GS} = 84 \cdot 1.207 = 101.5 \approx 102 \text{ per arm}$$

  - Group size to enroll for each stage is $m = 102/5 = 20.4 \approx 21$ per treatment
  - Stop and reject H_0 at analysis j if $|Z(t_j)| \geq 2.413$ for $j = 1, \ldots, 5$

Example

- O'Brien-Fleming design with 5 looks (equally spaced)
  - Inflation factor: $R_{OBF}(5, 0.05, 0.1) = 1.026$
  - Sample size needed:

$$n_{GS} = 84 \cdot 1.026 = 86.2 \approx 87 \text{ per arm}$$

  - Group size to enroll for each stage is $m = 87/5 = 17.4 \approx 18$ per treatment
  - Stop and reject H_0 at analysis j if $|Z(t_j)| \geq 2.040\sqrt{5/j}$ for $j = 1, \ldots, 5$

Expected Sample Size

- Let V be the interim analysis number at the time that the trial stops
- Let N be the trial's final sample size; note that N depends on V
- Probability of stopping at analysis j is $P(V = j)$, given by

$$P\left[|Z(t_1)| < b(t_1), \ldots, |Z(t_{j-1})| < b(t_{j-1}), |Z(t_j)| \geq b(t_j)\right]$$

- Let $E(V)$ be the expected number of interim analyses conducted assuming H_1 is true ($\delta \neq 0$)
- Expected sample size if H_1 is true:

$$E(N) = n_F \cdot R \cdot \frac{E(V)}{K}$$

Expected Sample Size

- For a design with $\alpha = 0.05$, $1 - \beta = 0.9$, and $K = 5$:

Boundary	$E(V)$	Max N	$E(N)$
Pocock	2.83	$R_P \cdot n_F = 1.21 n_F$	$0.68 n_F$
OBF	3.65	$R_{OBF} \cdot n_F = 1.03 n_F$	$0.75 n_F$
Fixed		n_F	n_F

- O'Brien-Fleming designs have a lower maximum sample size than Pocock, but higher average sample size if H_1 is true

Choosing a Stopping Boundary

- Pocock Design:
  - Higher chance of early stopping under alternative
  - Higher inflation factor and maximum N
  - Lower expected sample size under alternative
  - Higher expected sample size under H_0
- O'Brien-Fleming Design:
  - Lower chance of stopping early under alternative
  - Lower inflation factor and maximum N
  - Higher expected sample size under alternative
  - Lower expected sample size under H_0

Choosing a Stopping Boundary

- O'Brien-Fleming (OBF) design is much more common than Pocock; reasons include:
  - Many trials fail to find efficacy, in which case N is smaller and trial is less costly with OBF design
  - In early stages of trial, data may be less reliable and possibly un-representative; better to make it more difficult to stop early
  - Similarly, may prefer to make it easier to reject later in the trial when more data are available
  - OBF still offers big reduction in expected sample size under H_1 with very slight increase in maximum N compared to fixed sample design

Choosing a Stopping Boundary

- For OBF test, boundary value at the end of the study is very close to usual fixed sample boundary
- Example:
  - Suppose $Z = 2.30$ at end of study
  - For fixed sample design, $|Z| \geq 1.96$, so reject H_0
  - For Pocock design with 5 looks, $|Z| \geq C_P = 2.413$, so accept H_0
  - For OBF design with 5 looks, $|Z| \geq C_B = 2.04$, so reject H_0;
  - OBF final boundary of 2.04 is very close to fixed sample design boundary of 1.96

Theory of Boundary Selection

- Boundary values must satisfy

$$P_{H_0}[|Z(t_1)| < b(t_1), \ldots, |Z(t_K)| < b(t_K)] \geq 1 - \alpha$$

- Need to known the joint distribution of $(Z(t_1), \ldots, Z(t_K))$ to pick boundaries
- Recall that the variance of $\bar{X}_{1j} - \bar{X}_{2j}$ is

$$V_j = \sigma^2/n_{1j} + \sigma^2/n_{2j}$$

  where n_{ij} is the sample size in group i at time t_j
- Define the information at t_j as $I_j = V_j^{-1}$; then

$$Z_j = Z(t_j) = (\bar{X}_{1j} - \bar{X}_{2j})\sqrt{I_j}$$

Theory of Boundary Selection

- The vector $\mathbf{Z} = (Z_1, \ldots, Z_K)$ is multivariate normally distributed
- It can be shown that:
  - $Z_j \sim N(\delta\sqrt{I_j}, 1)$ for $j = 1, ..., K$
  - $Cor(Z_i, Z_j) = Cov(Z_i, Z_j) = \sqrt{I_i/I_j}$ for $i, j = 1, ..., K, i \leq j$, since $Var(Z_j) = 1$ for all j
- This multivariate normal structure is sometimes referred to as the **Canonical Distribution**; also called the **Independent Increments** property

Theory of Boundary Selection

- Evaluation of type I error rate requires integration of the multivariate normal distribution over subsets of $\mathbb{R}^K$
- This can be done efficiently by numerical methods if the independent increments property holds
- Test statistics for many common tests approximately follow the canonical distribution, so boundaries can be computed by standard methods

Boundary Selection for Other Test Statistics

- Comparison of two proportions: Z test at interim look j has test statistic

$$Z = (\hat{p}_{1j} - \hat{p}_{2j})\sqrt{I_j}, \text{ where}$$
$$I_j = [\bar{p}_j(1-\bar{p}_j)(1/n_{1j} + 1/n_{2j})]^{-1};$$

  information is approximately proportional to sample size
- Comparison of two survival curves: Log-rank test
  - Information is approximately proportional to the number of deaths

Timing of Interim Analyses

- O'Brien-Fleming and Pocock designs assume that interim analyses occur after equal amounts of information (or sample sizes) are collected
- It may be impractical to monitor the data at equal increments of information:
  - Enrollment pattern may be unpredictable
  - Scheduling interim analyses on a calendar basis (e.g. yearly) is often more convienent; likely won't align with equal information increments
  - Nuisance parameters may be misspecified (e.g. we assumed $\sigma^2 = 4$, but really $\sigma^2 = 9$)
- The error rate α can be controlled while allowing flexibility in analyses timings using an **Error Spending Function**

Error Spending Approach for Group Sequential Trials

Time	Rejection Region	Stopping Prob (H_0)	Cumulative Stopping Prob (H_0)
t_1	$\{\|Z(t_1)\| \geq b_1\}$	θ_1	$\alpha_1 = \theta_1$
t_2	$\{\|Z(t_1)\| < b_1, \|Z(t_2)\| \geq b_2\}$	θ_2	$\alpha_2 = \theta_1 + \theta_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_K	$\{\|Z(t_1)\| < b_1, \ldots,$ $\|Z(t_{K-1})\| < b_{K-1}, \|Z(t_K)\| \geq b_K\}$	θ_K	$\alpha_K = \theta_1 + \ldots + \theta_K$

- θ_j is the probability of stopping study at analysis j under H_0
- α_j is the probability of stopping study at or before analysis j under H_0
- α_j is the amount of type I error "spent" through time t_j
- Set overall significance level $\alpha_K = \alpha$

Error Spending Approach

- Lan & DeMets suggested parametrizing group sequential tests through α_j rather than boundaries b_j
- A 1-1 correspondence exists between the b_j and α_j
- To determine the α_j, specify an error spending function $\alpha(\tau)$, an increasing function with $\alpha(0) = 0$, $\alpha(1) = \alpha$
- $\alpha(\tau)$ is associated with an information time scale, τ
- $\tau_j = I(t_j)/I(t_K)$ is the information fraction at analysis j
- Set $\alpha_j = \alpha(\tau_j)$ for jth interim analysis
- $\tau_K = 1$ represents the final analysis, with $\alpha(\tau_k) = \alpha$
- Error spending approach is equivalent to specifying a group sequential boundary, while allowing flexibility in analysis timing

Commonly Used Error Spending Functions

- "Pocock type" spending function (Lan & DeMets):

$$\alpha(t) = \min\left(\alpha \log(1 + (e-1)t), \alpha\right)$$

- "O'Brien-Fleming type" spending function (Lan & DeMets):

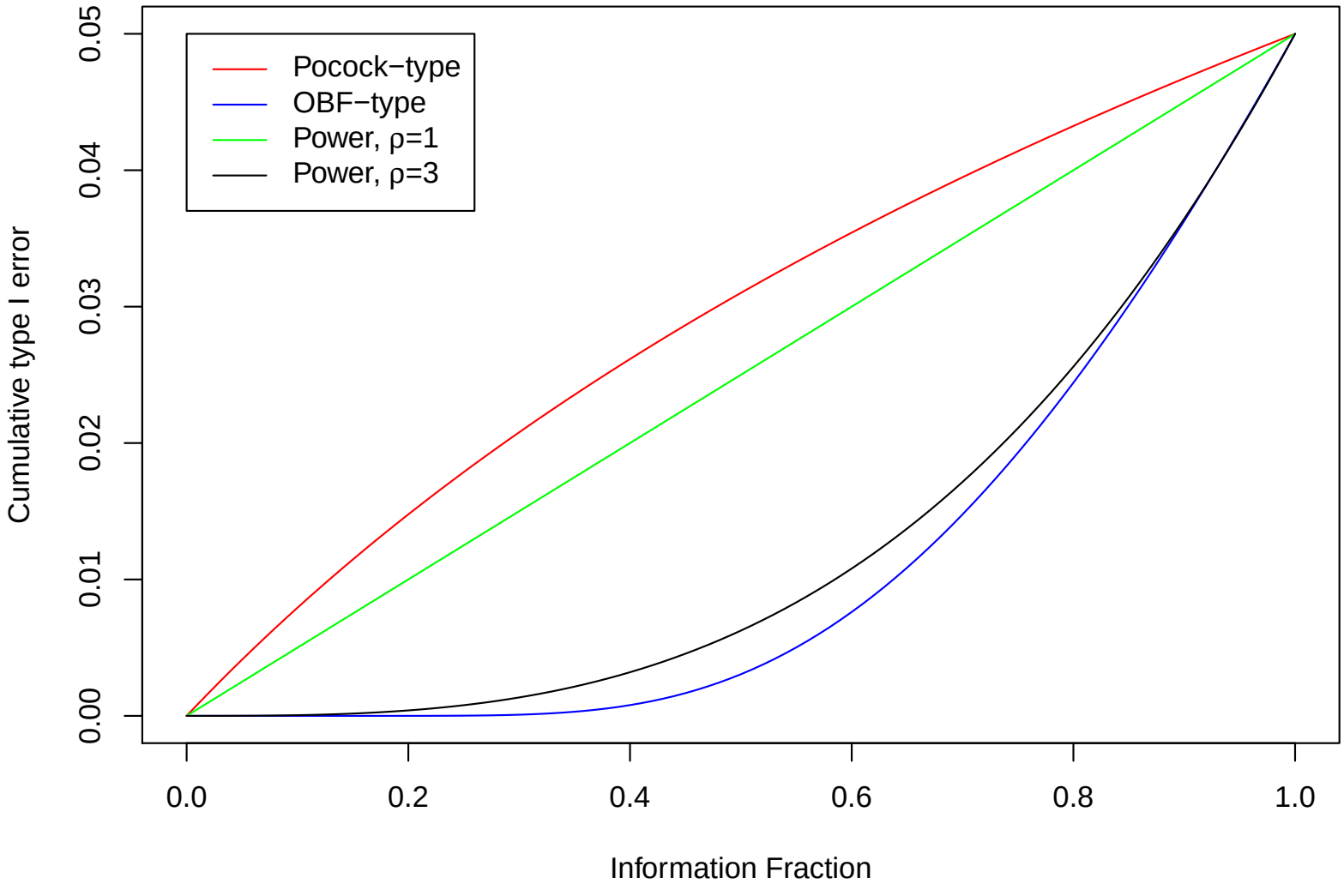
$$\alpha(t) = \min\left(4 - 4\Phi(z_{\alpha/4}/\sqrt{t}), \alpha\right)$$

- Power family (Jennison & Turnbull):

$$\alpha(t) = \min(\alpha t^{\rho}, \alpha), \text{ where } \rho > 0$$

$\rho = 1$ is similar to Pocock, $\rho = 3$ similar to OBF

Commonly Used Error Spending Functions



Calculation of Boundaries

- Suppose analyses will be conducted at times $(t_1, \ldots, t_K)$
- K and times need not be specified in advance
- Boundary values are computed iteratively according to

$$P_{H_0}\{|Z(t_1)| \geq b_1\} = \theta_1 = \alpha(\tau_1),$$

$$P_{H_0}\{|Z(t_1)| < b_1, |Z(t_2)| \geq b_2\} = \theta_2 = \alpha(\tau_2) - \alpha(\tau_1),$$

$$\vdots$$

$$P_{H_0}\{|Z(t_1)| < b_1, \ldots, |Z(t_{K-1})| < b_{K-1}, |Z(t_K)| \geq b_K\} = \theta_K = \alpha(\tau_K) - \alpha(\tau_{K-1})$$

- Can solve these recursively for b_1, then b_2, etc. using the multivariate normal distribution of $\mathbf{Z}$

Considerations for Error Spending Approach

- Type I error rate is not controlled if the choice of monitoring times is dependent on the observed treatment differences
  - Interim analysis times must be chosen independent of treatment effect info
- Boundaries cannot be computed in advance unless information fractions are prespecified
- For study design, we need to provide adequate power to detect δ of interest
  - Power depends on the information sequence $\{\tau_j\}$
  - Protocol should specify an information sequence to be targeted for the study (e.g. 3 looks at 50%, 75%, 100% information)
  - An equally spaced sequence is often assumed when designing study

Impact of Misspecification / Change of Design Parameters

- If design parameters are misspecified or changed, the power can be impacted:
  - Number of analyses K: minimal impact
  - Information fractions: minimal, if close to the sequence assumed at the design stage
  - Nuisance parameters: can result in over or underpowered study; same problem happens with fixed sample designs

Information Fractions

- Two potential definitions of information fraction:
  - Ratio of sample size at interim analysis to maximum sample size
  - Ratio of information at interim analysis to maximum information, where information $= Var(\text{trt effect estimate})^{-1}$

Example

- Clinical trial to test $H_0 : \mu_1 = \mu_2$ with $\alpha = 0.05$, power $1 - \beta = 0.9$ when $\mu_1 - \mu_2 = 1$, and variance $\sigma^2 = 4$
- Fixed sample design requires $n = 84$ per group
- Monitor 3 times, equal information, $\rho = 3$ spending function, $\alpha(\tau) = \min(\alpha\tau^3, \alpha)$: Inflation factor $R_{\rho=3} = 1.018$
- Maximum sample size: $n_{\rho=3} = 84 \cdot 1.018 = 85.5 \approx 86$ per group
- Maximum Information $= SE^{-2}$:

$$MI = \left[\sigma^2(1/n_1 + 1/n_2)\right]^{-1} = \frac{86/2}{4} = 10.8$$

Interim Analysis 1

Group	n	$\bar{x}$	s
1	20	7.1	2.5
2	25	8.3	2.9

- Strategy A: Information fraction based on ratio of sample sizes
  - Information fraction: $\tau_1 = 45/172 = 0.26$
  - Type I error spent: $\alpha(\tau_1) = 0.05(0.26)^3 = 0.00088$
  - Boundary: $b_1 = 3.326$
- Test statistic:

$$Z(t_1) = \frac{7.1 - 8.3}{0.8055} = -1.490$$

- Continue study to next analysis

Interim Analysis 1

Group	n	$\bar{x}$	s
1	20	7.1	2.5
2	25	8.3	2.9

- Strategy B: Information fraction based on ratio of current information to maximum information
  - Standard Error: $SE = \sqrt{s_1^2/n_1 + s_2^2/n_2} = 0.8055$
  - Information: $I(t_1) = (0.8055)^{-2} = 1.541$
  - Information fraction: $\tau_1 = 1.541/10.8 = 0.143$
  - Type I error spent: $\alpha(\tau_1) = 0.05(0.143)^3 = 0.00015$
  - Boundary: $b_1 = 3.791$
- Test statistic:

$$Z(t_1) = -1.490$$

- Continue study to next analysis

Interim Analysis 2

Group	n	$\bar{x}$	s
1	55	6.9	2.8
2	60	8.4	3.1

- Strategy A: Information fraction based on ratio of sample sizes
  - Information fraction: $\tau_2 = 115/172 = 0.669$
  - Cumulative Type I error spent: $\alpha(\tau_2) = 0.05(0.669)^3 = 0.015$
  - Type I error spent at this look: $\theta_2 = 0.015 - 0.00088 = 0.0141$
  - Boundary: $b_2 = 2.452$
- Test statistic:

$$Z(t_2) = \frac{6.9 - 8.4}{0.5502} = -2.726$$

- Stop study and reject H_0

Interim Analysis 2

Group	n	$\bar{x}$	s
1	55	6.9	2.8
2	60	8.4	3.1

- Strategy B: Information fraction based on ratio of current information to maximum information
  - Standard Error: $SE = \sqrt{s_1^2/n_1 + s_2^2/n_2} = 0.5502$
  - Information: $I(t_2) = (0.5502)^{-2} = 3.303$
  - Information fraction: $\tau_2 = 3.303/10.8 = 0.306$
  - Cumulative Type I error spent: $\alpha(\tau_2) = 0.05(0.306)^3 = 0.0014$
  - Type I error spent at this look: $\theta_2 = 0.0014 - 0.00015 = 0.00125$
  - Boundary: $b_2 = 3.227$
- Test statistic: $Z(t_2) = -2.726$
- Continue study to next analysis

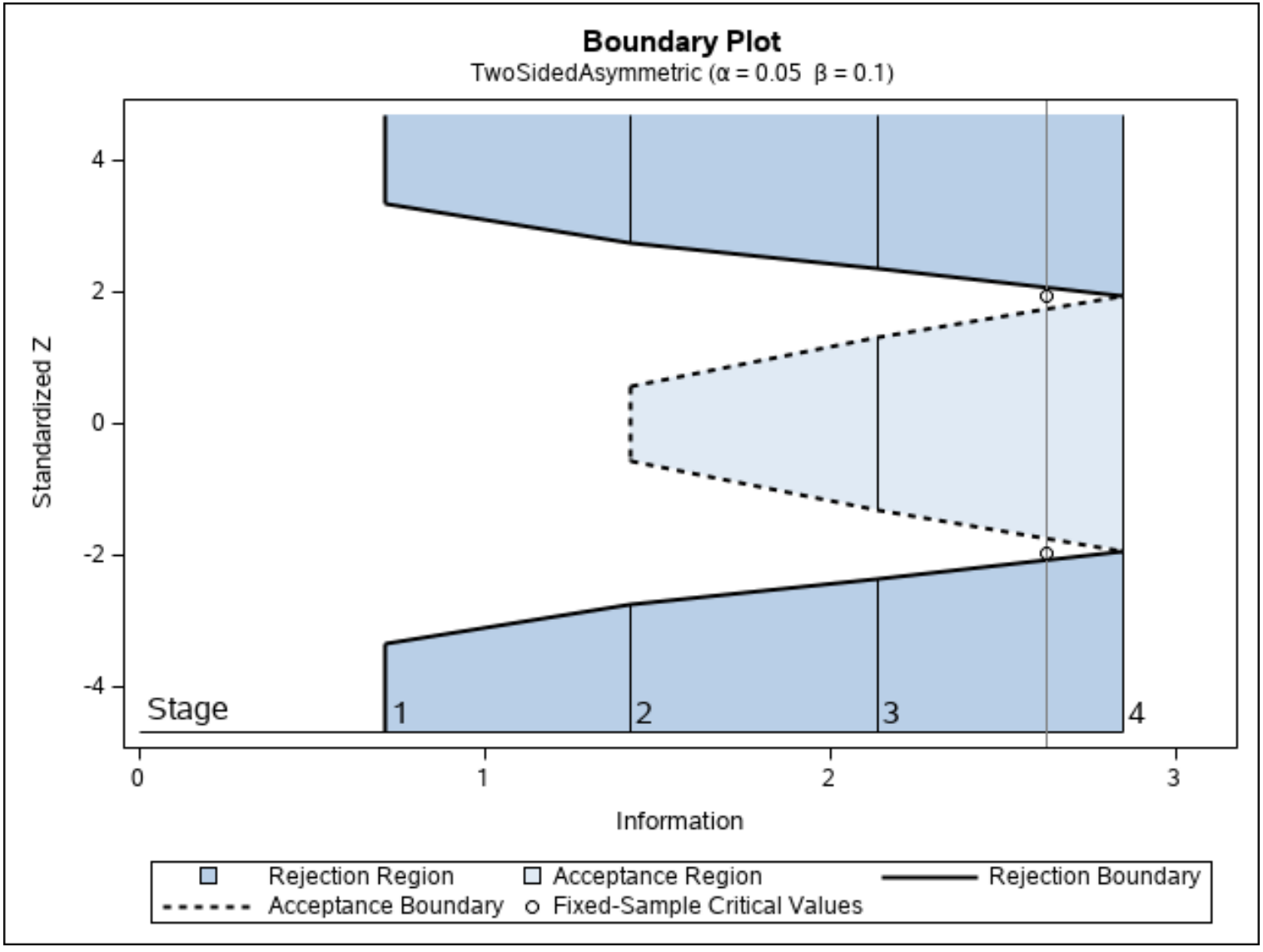
Example

- Some evidence exists that the prespecified variance is wrong: assumed $\sigma^2 = 4$ but $s_1^2 = 7.84, s_2^2 = 9.61$ at look 2
- With $\sigma_1^2 = 7.84$ and $\sigma_2^2 = 9.61$ we would need 186 patients per arm to meet the power requirement
- Strategy B allows for sample size escalation to attain the desired power, since very little type I error is spent early
- Strategy A spends the type I error based on the original assumption about the variance, so will be underpowered

Group Sequential Design: Extensions

- Early futility stopping:
  - Futility boundaries $\{d_j\}$ included such that trial stops for futility at stage j if $|Z(t_j)| \leq d_j$
  - With equal information increments, these can be determined directly similar to Pocock and O'Brien-Fleming designs
  - Error spending approach: Specify function $\beta(\tau)$ such that $\beta(0) = 0$, $\beta(1) = \beta$; use it to determine futility boundaries at interim analyses (type II error spending)
- One-sided testing
- Noninferiority and equivalence testing

Example: Group Sequential Design with Early Efficacy and Futility Stopping



Group Sequential Design: Extensions

- Other types of outcomes (binary, categorical, survival, etc.): Same boundary specification and error-spending methods can be used, relying on asymptotic normal distributions of treatment effects for most common methods, including:
  - Pearson's chi-square and Mantel-Haenszel tests
  - Logistic regression model
  - Log rank test and Cox proportional hazards models
  - Generalized linear models
  - Linear and generalized liear mixed models

Stochastic Curtailment

- Idea: Terminate a study if a particular outcome is highly likely given the observed data
- Often this is done using a concept called conditional power
- Testing $H_0 : \theta = 0$ vs. $H_1 : \theta = \delta$, where θ represents the treatment effect
- Reference test $\mathcal{T}$ used with type I error α under $\theta = 0$ and power $1 - \beta$ at $\theta = \delta$
  - One-sided, fixed-sample test
  - Reject H_0 if $Z \geq z_\alpha$
  - Power of $1 - \beta$ achieved by choosing the sample size to yield an information level I_K

Stochastic Curtailment

- At interim analysis j, let Z_j denote the test statistic so far
- **Conditional Power** to detect θ at stage j:

$$p_j(\theta) = P_\theta(\mathcal{T} \text{ rejects } H_0 | \text{data at stage } j) = P_\theta(\mathcal{T} \text{ rejects } H_0 | Z_j)$$

- If $p_j(0)$ is high, then the reference test is likely to reject H_0 even if H_0 is true
- If $p_j(\delta)$ is low, then the reference test is unlikely to reject H_0 even if H_0 is false
- Formal monitoring rules:
  - Early termination for efficacy: Reject H_0 at stage j if $p_j(0) \geq \gamma$ (Usually 0.8 or 0.9)
  - Early termination for futility: Accept H_0 at stage j if $1 - p_j(\delta) \geq \gamma'$

Stochastic Curtailment: Properties

- Fully sequential monitoring is permitted (can look anytime)
- Can be shown that the type I error rate is no more than α/γ, type II error rate is no more than β/γ'
- Can ensure appropriate α, β error rates for sequential procedure by choosing fixed sample test $\mathcal{T}$ to have type I error rate $\alpha\gamma$ and power $1 - \beta\gamma'$ at $\theta = \delta$

Conditional Power

- Conditional power at interim analysis j, with information I_j

$$p_j(\theta) = P_\theta(Z_K \geq z_{1-\alpha} | Z_j)$$

- The conditional distribution of $Z_K | Z_j$ is

$$Z_K | Z_j \sim N\left(Z_j \sqrt{I_j / I_K} + \frac{\theta(I_K - I_j)}{\sqrt{I_K}}, 1 - \frac{I_j}{I_K} \right)$$

- Then the conditional power is

$$p_j(\theta) = \Phi\left(\frac{Z_j \sqrt{I_j} - z_{1-\alpha}\sqrt{I_K} + \theta(I_K - I_j)}{\sqrt{I_K - I_j}} \right)$$

Conditional Power

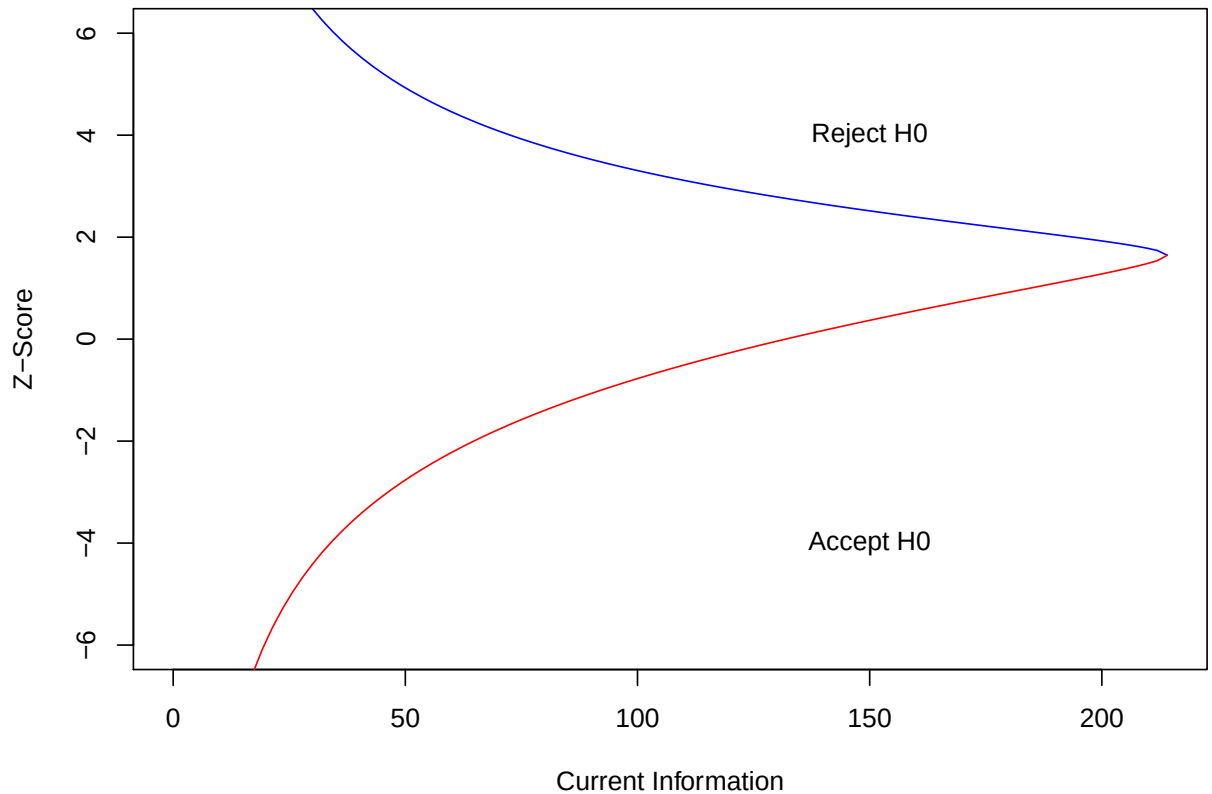
- Stopping rules can be directly translated into stopping boundaries:
  - Stop and reject H_0 if $p_j(0) \geq \gamma$

$$Z_j \geq z_{1-\alpha}\sqrt{I_K/I_j} + z_{\gamma}\sqrt{(I_K - I_j)/I_j}$$

  - Stop and accept H_0 if $p_j(\delta) \leq 1 - \gamma'$

$$Z_j \leq z_{1-\alpha}\sqrt{I_K/I_j} + z_{1-\gamma'}\sqrt{(I_K - I_j)/I_j} - \frac{\delta(I_K - I_j)}{\sqrt{I_j}}$$

Stochastic Curtailment Example: $\alpha = 0.05$, $1 - \beta = 0.9$, $\delta = 0.2$, $I_K = 214.1$, $\gamma = \gamma' = 0.8$



- Actual type I error rate $\leq \alpha/\gamma = 0.0625$
- Actual power is $\geq 1 - 0.1/0.8 = 0.875$

Two-sided Conditional Power

- Reference test: Reject H_0 if $|Z_K| \geq z_{\alpha/2}$
- Conditional power at analysis j:

$$p_j(\theta) = \Phi\left(\frac{Z_j\sqrt{I_j} - z_{\alpha/2}\sqrt{I_K} + \theta(I_K - I_j)}{\sqrt{I_K - I_j}}\right) + \Phi\left(\frac{-Z_j\sqrt{I_j} - z_{\alpha/2}\sqrt{I_K} - \theta(I_K - I_j)}{\sqrt{I_K - I_j}}\right)$$

- Can use this for fully sequential testing, with same impacts on type I/II error rates

Comments on Conditional Power

- Offers a simple rule that is often applied for early futility stopping
- Can be compared directly to group sequential methods as long as we calibrate the error rates to be equal
- Fully sequential, not group sequential monitoring
  - Allows for arbitrary, unplanned analysis times, and any K
  - May be conservative (α too small) compared to a group sequential design with a planned analysis schedule

Repeated Confidence Intervals

- Multiple looks at the data affect confidence intervals, just like they affect significance levels of hypothesis tests
- Unadjusted (naive) 95% confidence intervals for mean differences: After each group j of observations, $j = 1, \ldots, K$, form a 95% confidence interval

$$\bar{X}_{1j} - \bar{X}_{2j} \pm 1.96\sqrt{\sigma_1^2/n_{1j} + \sigma_2^2/n_{2j}}$$

- The probability that all K intervals contain the true value of θ is $< 95\%$

Repeated Confidence Intervals

- **Repeated Confidence Intervals**: Set of random intervals $\{\text{RCI}_j\}$ where the simultaneous coverage probability is maintained at $1-\alpha$:

$$P(\theta \in \text{RCI}_j \text{ for all } j = 1, \ldots, K) \geq 1 - \alpha$$

- RCIs constructed by inverting group sequential tests
- Define $Z_j(\theta_0) = Z_j - \theta_0\sqrt{I_j}$ for $j = 1, \ldots, K$
- Group sequential test: reject $H_0 : \theta = \theta_0$ at stage j if $|Z_j(\theta_0)| \geq c_j(\alpha)$
- Then the RCIs are defined as

$$\begin{aligned}\text{RCI}_j &= \{\theta_0 : |Z_j(\theta_0)| < c_j(\alpha)\} = \left(\frac{Z_j - c_j(\alpha)}{\sqrt{I_j}}, \frac{Z_j + c_j(\alpha)}{\sqrt{I_j}}\right) \\ &= (\hat{\theta}_j - c_j(\alpha)I_j^{-1/2}, \hat{\theta}_j + c_j(\alpha)I_j^{-1/2}),\end{aligned}$$

where $\hat{\theta}_j$ is the estimate of θ at look j

Repeated Confidence Intervals

- For comparing two means:
  - Naive 95% confidence interval:

$$\bar{X}_{1j} - \bar{X}_{2j} \pm z_{1-\alpha/2}\sqrt{\sigma_1^2/n_{1j} + \sigma_2^2/n_{2j}}$$

  - RCIs:

$$\bar{X}_{1j} - \bar{X}_{2j} \pm c_j(\alpha)\sqrt{\sigma_1^2/n_{1j} + \sigma_2^2/n_{2j}}, \quad j = 1, ..., K$$

  - With O'Brien-Fleming design, RCIs are wide at early analyses and get closer to naive CIs at end of study

Example: Group Sequential Design in R

- Designing a group sequential study:
  - $H_0 : \mu_1 = \mu_2$, 2-sided $\alpha = 0.05$, power $1 - \beta = 0.9$ when $\delta = \mu_1 - \mu_2 = 1$, $\sigma^2 = 4$
  - $K = 3$ looks with an O'Brien-Fleming boundary
  - Sample size for fixed design

$$n = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2 \sigma^2}{\delta^2} = 84 \text{ per group, } 168 \text{ total}$$

  - Sample size for O'Brien-Fleming design using R:

```
gsDesign(k=3, test.type=2, sfu="OF",n.fix=168)
```

Example: O'Brien-Fleming Design in R

```
Symmetric two-sided group sequential design with
90 % power and 2.5 % Type I Error.
Spending computations assume trial stops
if a bound is crossed.
```

Analysis	N	Z	Nominal p	Spend
1	57	3.47	0.0003	0.0003
2	114	2.45	0.0071	0.0069
3	171	2.00	0.0225	0.0178
Total				0.0250

Example: O'Brien-Fleming Design in R

```
++ alpha spending:
 O'Brien-Fleming boundary.

Boundary crossing probabilities and expected sample size
assume any cross stops the trial

Upper boundary (power or Type I Error)
          Analysis
   Theta      1      2      3 Total  E{N}
  0.0000 0.0003 0.0069 0.0178 0.025 169.9
  0.2501 0.0565 0.5288 0.3147 0.900 134.2

Lower boundary (futility or Type II Error)
          Analysis
   Theta     1      2      3 Total
  0.0000 3e-04 0.0069 0.0178 0.025
  0.2501 0e+00 0.0000 0.0000 0.000
```

Example: O'Brien-Fleming Design

- Interim analysis 1:

Group	n	$\bar{x}$	s
1	20	7.1	2.5
2	25	8.3	2.9

  - Test statistic $Z(t_1) = -1.49$
  - Boundary: $b_1 = 3.47$
- Continue study

Example: O'Brien-Fleming Design

- Interim analysis 2:

Group	n	$\bar{x}$	s
1	55	6.9	2.8
2	60	8.4	3.1

  - Test statistic: $Z(t_2) = -2.73$
  - Boundary: $b_2 = 2.45$
- Stop study and reject H_0

Repeated Confidence Intervals

- 95% RCIs:
  - Look 1: Boundary value $b_1(\alpha) = 3.47$

$$\text{RCI}_1 = (7.1 - 8.3) \pm 3.47(0.8055) = [-4.00, 1.60]$$

  - Look 2: Boundary value $b_2(\alpha) = 2.45$

$$\text{RCI}_2 = (6.9 - 8.4) \pm 2.45(0.5502) = [-2.85, -0.15]$$