Session 2

Group Sequential Stopping Rules

- Need for Monitoring a Trial
- Criteria for Early Stopping
- Inadequacy of Fixed Sample Methods
- Stopping Rules

Families of Designs

- Boundary Scales
- Unified Family (Sample Mean Scale)
- Error Spending Family
- Comparison of Parameterizations

Evaluation of Group Sequential Designs

Need for Monitoring a Trial

Need for Monitoring a Trial

Fixed sample two-sided tests

Test of a two-sided alternative ($\theta_+ > \theta_0 > \theta_-$)

- Upper Alternative: $H_+: \theta \geq \theta_+$ (superiority)
- Null: $H_0: \theta = \theta_0$ (equivalence)
- Lower Alternative: $H_-: \theta \leq \theta_-$ (inferiority)

Data analyzed once at the end of all data accrual

Decisions:

- Reject H_0, H_- (for H_+) $\Longleftrightarrow$ $T \geq c_U$
- Reject H_+, H_- (for H_0) $\Longleftrightarrow$ $c_L \leq T \leq c_U$
- Reject H_+, H_0 (for H_-) $\Longleftrightarrow$ $T \leq c_L$

Need for Monitoring a Trial

Ethical concerns

Patients already on trial

- Avoid continued administration of harmful treatments
- Maintain validity of informed consent

Need for Monitoring a Trial

Ethical concerns (cont.)

Patients not yet on trial

- Start treatment with best therapy
- Ensure informed consent valid

Need for Monitoring a Trial

Ethical concerns (cont.)

Patients never on trial

- Facilitate rapid introduction of beneficial treatments
- Warn about risks of existing treatments

Need for Monitoring a Trial

Efficiency considerations

Fewer patients may be needed on average

- Decreases costs associated with number of patients

Time savings

- Decreases costs associated with monitoring patients

Need for Monitoring a Trial

Futility considerations: Efficiency and Ethics

Efficiency

- Stop a study when it is known (or reasonably certain) that no effect will be demonstrated
- Can perform more studies with limited resources

Ethics

- Is it ever ethical to expose patients to experimental treatments when no meaningful information will be gained?
- Can devote resources to study of more promising agents

Criteria for Stopping a Trial

Criteria for Stopping a Trial

Sufficient evidence available to be confident of rejecting specific hypotheses

Stopping early for

- Efficacy (superiority)
- Harm (inferiority)
- Equivalence

Criteria for Stopping a Trial

Futility of demonstrating effect that would change behavior

Stopping early for futility

- Not sufficiently superior
- Not dangerously harmful

Criteria for Stopping a Trial

And there is no advantage in continuing

Even if confident of ultimate decision about primary endpoint, may want to continue trial to gain more information on

- Safety
- Longer term follow-up
- Gather additional data on secondary outcomes

Criteria for Stopping a Trial

Statistical basis for stopping criteria

Curtailment

- Boundary has been reached early
- E.g., one arm study with binary endpoint
  - Critical value for rejection of null might be observation of K events
  - Kth event may occur well before all subjects accrued

Criteria for Stopping a Trial

Statistical basis for stopping criteria (cont.)

Stochastic Curtailment

- High probability that a particular decision will be made at final analysis
- Calculate probability of exceeding some critical value conditional on data observed so far
- Probability calculated based on hypothesized treatment effect (which hypothesis?) or current estimate

Criteria for Stopping a Trial

Statistical basis for stopping criteria (cont.)

Predictive probability of final statistic

- A special form of stochastic curtailment
- Uses a Bayesian prior distribution on the treatment effect

Criteria for Stopping a Trial

Statistical basis for stopping criteria (cont.)

Group sequential test

- Sufficient evidence to make decision in classical frequentist framework
- Type I and II errors controlled at desired levels

Criteria for Stopping a Trial

Statistical basis for stopping criteria (cont.)

Bayesian analysis

- Compute the probability that the treatment effect is in some specified range
- Calculations based on a user specified prior distribution for the treatment effect (which is treated as a random variable)

Inadequacy of Fixed Sample Methods

Inadequacy of Fixed Sample Methods

Sequential monitoring of a trial

Data are analyzed after accrual of each observation

- (Group sequential monitoring: analysis after groups of observations accrued)
- Analyses must take into account the repeated analyses of the same data
  - Sampling distribution of the test statistic is altered
  - Frequentist properties are altered

Inadequacy of Fixed Sample Methods

Setting for demonstration of the problem

Observations:

$$X_1, X_2, X_3, \ldots, X_N$$

$$X_i \sim N(\mu, \sigma^2)$$

Hypothesis:

$$H_0 : \mu = \mu_0$$

Inadequacy of Fixed Sample Methods

Test Statistic

Sample mean computed after each observation:

$$\bar{X}_j = \frac{1}{j}\sum_{i=1}^{j} X_j, \quad j = 1, \ldots, N$$

Inadequacy of Fixed Sample Methods

Fixed sample decision rule

Hypothesis test when all data accrued:

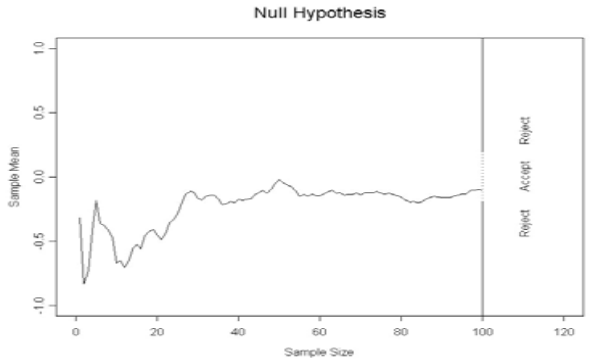
- Reject H_0 when

$$\bar{X}_N > \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{N}}$$

$$\bar{X}_N < \mu_0 - z_{1-\alpha}\frac{\sigma}{\sqrt{N}}$$

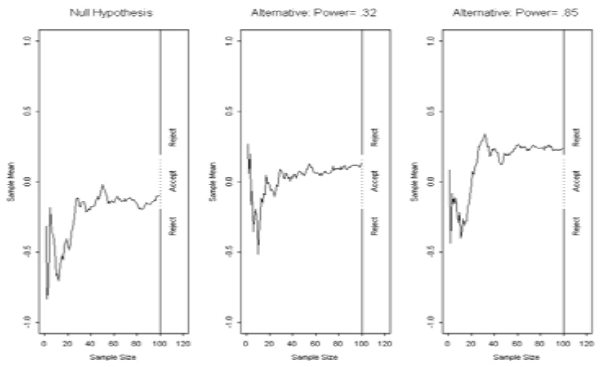
Inadequacy of Fixed Sample Methods

Sample path for sample mean



Inadequacy of Fixed Sample Methods

Sample path for sample mean



Inadequacy of Fixed Sample Methods

Repeated significance testing

Continuous monitoring:

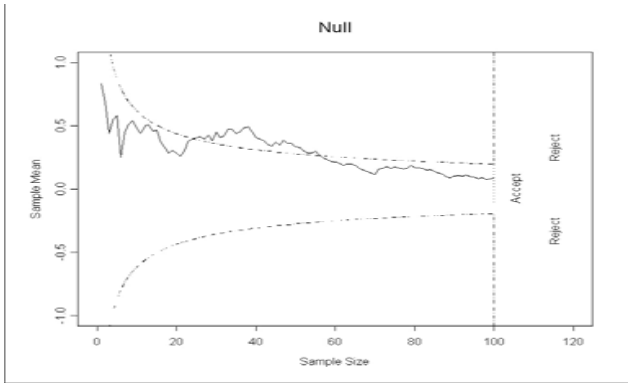
- Reject H_0 the first time

$$\bar{X}_j > \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{j}}$$

$$\bar{X}_j < \mu_0 - z_{1-\alpha}\frac{\sigma}{\sqrt{j}}$$

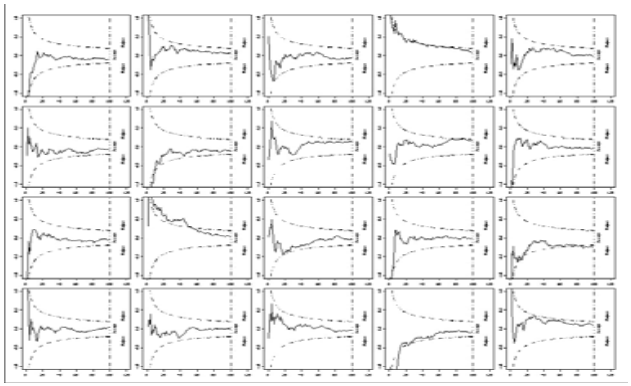
Inadequacy of Fixed Sample Methods

Simulated trials when H_0 is true:



Inadequacy of Fixed Sample Methods

Simulated trials when H_0 is true:



Inadequacy of Fixed Sample Methods

Repeated significance testing

Monitoring after each of J groups of observations:

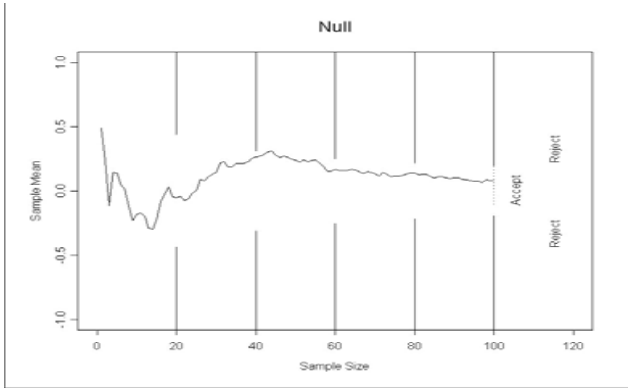
- Analyses at $N_1, N_2, \ldots, N_J$
- Reject H_0 the first time

$$\bar{X}_{N_j} > \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{N_j}}$$

$$\bar{X}_{N_j} < \mu_0 - z_{1-\alpha}\frac{\sigma}{\sqrt{N_j}}$$

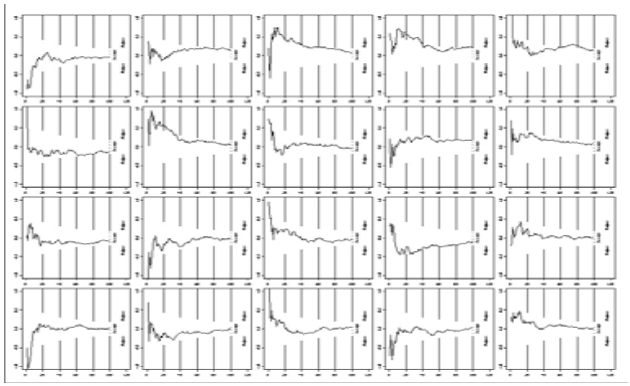
Inadequacy of Fixed Sample Methods

Simulated trials when H_0 is true:



Inadequacy of Fixed Sample Methods

Simulated trials when H_0 is true:



Inadequacy of Fixed Sample Methods

Simulate 100,000 Trials under the Null Hypothesis

Three equally spaced level .05 analyses

Proportion Significant	
1st	
.05038	

Inadequacy of Fixed Sample Methods

Simulate 100,000 Trials under the Null Hypothesis

Three equally spaced level .05 analyses

Proportion Significant	
1st	2nd
.05038	.05022

Inadequacy of Fixed Sample Methods

Simulate 100,000 Trials under the Null Hypothesis

Three equally spaced level .05 analyses

Proportion Significant		
1st	2nd	3rd
.05038	.05022	.05056

Inadequacy of Fixed Sample Methods

Simulate 100,000 Trials under the Null Hypothesis

Three equally spaced level .05 analyses

Pattern of Significance	Proportion Significant 1st
1st only	*.03046*
1st, 2nd	*.00807*
1st, 3rd	*.00317*
1st, 2nd, 3rd	*.00868*
Any pattern	.05038

Inadequacy of Fixed Sample Methods

Simulate 100,000 Trials under the Null Hypothesis

Three equally spaced level .05 analyses

Pattern of Significance	Proportion Significant 1st	2nd
1st only	*.03046*	
1st, 2nd	*.00807*	*.00807*
1st, 3rd	*.00317*	
1st, 2nd, 3rd	*.00868*	*.00868*
2nd only		*.01921*
2nd, 3rd		*.01426*
Any pattern	.05038	.05022

Inadequacy of Fixed Sample Methods

Simulate 100,000 Trials under the Null Hypothesis

Three equally spaced level .05 analyses

Pattern of Significance	Proportion Significant 1st	2nd	3rd
1st only	*.03046*		
1st, 2nd	*.00807*	*.00807*	
1st, 3rd	*.00317*		*.00317*
1st, 2nd, 3rd	*.00868*	*.00868*	*.00868*
2nd only		*.01921*	
2nd, 3rd		*.01426*	*.01426*
3rd only			*.02445*
Any pattern	.05038	.05022	.05056

Inadequacy of Fixed Sample Methods

Simulate 100,000 Trials under the Null Hypothesis

Three equally spaced level .05 analyses

Pattern of Significance	Proportion Significant 1st	2nd	3rd	Ever
1st only	*.03046*			.03046
1st, 2nd	*.00807*	*.00807*		.00807
1st, 3rd	*.00317*		*.00317*	.00317
1st, 2nd, 3rd	*.00868*	*.00868*	*.00868*	.00868
2nd only		*.01921*		.01921
2nd, 3rd		*.01426*	*.01426*	.01426
3rd only			*.02445*	.02445
Any pattern	.05038	.05022	.05056	**.10830**

Inadequacy of Fixed Sample Methods

Group sequential test: Pocock (1977) level .05

Three equally spaced level .022 analyses

Pattern of Significance	Proportion Significant 1st	2nd	3rd	Ever
1st only	*.01520*			.01520
1st, 2nd	*.00321*	*.00321*		.00321
1st, 3rd	*.00113*		*.00113*	.00113
1st, 2nd, 3rd	*.00280*	*.00280*	*.00280*	.00280
2nd only		*.01001*		.01001
2nd, 3rd		*.00614*	*.00614*	.00614
3rd only			*.01250*	.01250
Any pattern	.02234	.02216	.02257	**.05099**

Inadequacy of Fixed Sample Methods

Critical values depend on spacing of analyses

Level .022 analyses at 10%, 20%, 100% of data

Pattern of Significance	Proportion Significant 1st	2nd	3rd	Ever
1st only	*.01509*			.01509
1st, 2nd	*.00521*	*.00521*		.00521
1st, 3rd	*.00068*		*.00068*	.00068
1st, 2nd, 3rd	*.00069*	*.00069*	*.00069*	.00069
2nd only		*.01473*		.01473
2nd, 3rd		*.00165*	*.00165*	.00165
3rd only			*.01855*	.01855
Any pattern	.02167	.02228	.02157	**.05660**

Inadequacy of Fixed Sample Methods

The critical values can be varied across analyses

Level 0.10 O'Brien-Fleming (1979); equally spaced tests at .003, .036, .087

Pattern of Significance	Proportion Significant 1st	2nd	3rd	Ever
1st only	*.00082*			.00082
1st, 2nd	*.00036*	*.00036*		.00036
1st, 3rd	*.00037*		*.00037*	.00037
1st, 2nd, 3rd	*.00127*	*.00127*	*.00127*	.00127
2nd only		*.01164*		.01164
2nd, 3rd		*.02306*	*.02306*	.02306
3rd only			*.06223*	.01855
Any pattern	.00282	.03633	.08693	**.09975**

Inadequacy of Fixed Sample Methods

Error spending function: Pocock (1977) level .05

Pattern of Significance	Proportion Significant 1st	2nd	3rd	Ever
1st only	*.01520*			.01520
1st, 2nd	*.00321*	*.00321*		.00321
1st, 3rd	*.00113*		*.00113*	.00113
1st, 2nd, 3rd	*.00280*	*.00280*	*.00280*	.00280
2nd only		*.01001*		.01001
2nd, 3rd		*.00614*	*.00614*	.00614
3rd only			*.01250*	.01250
Any pattern	.02234	.02216	.02257	**.05099**
Incremental error	.02234	.01615	.01250	
Cumulative error	.02234	.03849	.05099	

Stopping Rules

Stopping Rules

Basic Strategy

Find stopping boundaries at each analysis such that desired operating characteristics (e.g., type I and type II statistical errors) are attained

Stopping Rules

Issues

- Conditions under which the trial might be stopped early
- When to perform analyses
- Test statistic to use
- Relative position of boundaries at successive analyses
- Desired operating characteristics

Stopping Rules

Choice of Test Statistic

Let $T_n(X_1, ..., X_n)$ be any test statistic such that T_n tends to be large for larger values of θ

(Later we will consider possible choices for T_n)

Stopping Rules

Conditions for Early Stopping: One-sided tests

Test of a greater alternative ($\theta_+ > \theta_0$)

- Null: $H_0: \theta \leq \theta_0$
- Alternative: $H_1: \theta \geq \theta_+$

Possibilities for early stopping:

- Stop only for the null (when T_n small)
- Stop only for the alternative (when T_n large)
- Stop either for the null or for the alternative

Stopping Rules

Conditions for Early Stopping: One-sided tests

Test of a lesser alternative ($\theta_- < \theta_0$)

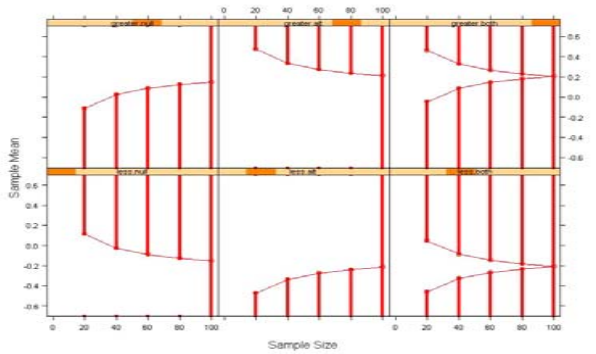
- Null: $H_0: \theta \geq \theta_0$
- Alternative: $H_1: \theta \leq \theta_-$

Possibilities for early stopping:

- Stop only for the null (when T_n large)
- Stop only for the alternative (when T_n small)
- Stop either for the null or for the alternative

Stopping Rules

One-sided Test Boundaries: Sample Mean Statistic



Stopping Rules

Conditions for Early Stopping: Two-sided tests

Test of a two-sided alternative ($\theta_+ > \theta_0 > \theta_-$)

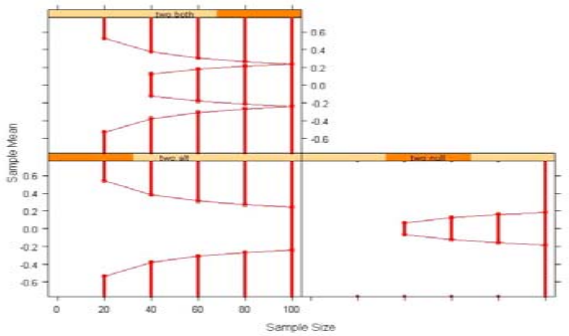
- Upper Alternative: H_+: $\theta \geq \theta_+$
- Null: H_0: $\theta = \theta_0$
- Lower Alternative: H_-: $\theta \leq \theta_-$

Possibilities for early stopping:

- Stop only for the null (when T_n intermediate)
- Stop only for the alternative (when T_n small or large)
- Stop either for the null or for the alternative

Stopping Rules

Two-sided Test Boundaries: Sample Mean Statistic



Stopping Rules

General stopping rule

Maximum of four boundaries

- 'd' boundary: upper outer boundary
- 'c' boundary: upper inner boundary
- 'b' boundary: lower inner boundary
- 'a' boundary: lower outer boundary

Early stopping

- T_n greater than 'd' boundary
- T_n between 'b' and 'c' boundaries
- T_n less than 'a' boundary

Stopping Rules

One-sided tests of greater hypotheses

Always have 'b' and 'c' boundaries are equal

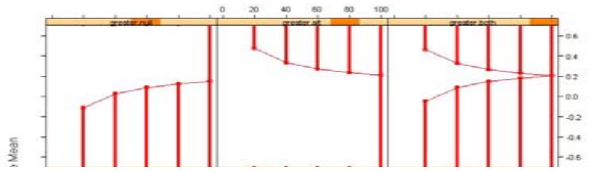
- so no early stopping for intermediate T_n

Early stopping

- If 'a' boundary at $-\infty$: no early stopping for null
- If 'd' boundary at ∞: no early stopping for alternative

Stopping Rules

One-sided Test Boundaries: Sample Mean Statistic



Stopping Rules

One-sided tests of lesser hypotheses

Always have 'b' and 'c' boundaries are equal

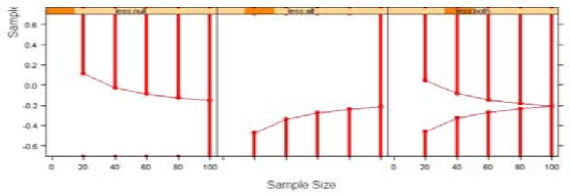
- so no early stopping for intermediate T_n

Early stopping

- If 'a' boundary at $-\infty$: no early stopping for alternative
- If 'd' boundary at ∞: no early stopping for null

Stopping Rules

One-sided Test Boundaries: Sample Mean Statistic



Stopping Rules

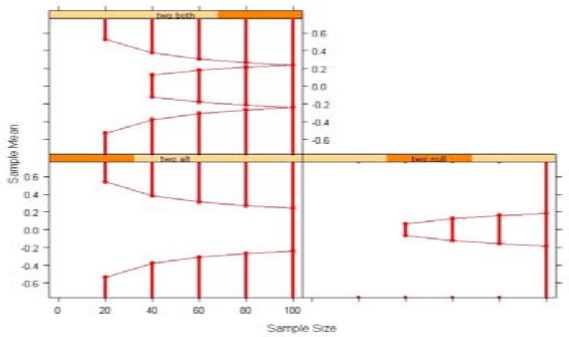
Two-sided tests

Early stopping

- If 'a' boundary at $-\infty$: no early stopping for lower alternative
- If 'b' and 'c' boundaries equal: no early stopping for null
- If 'd' boundary at ∞: no early stopping for upper alternative

Stopping Rules

Two-sided Test Boundaries: Sample Mean Statistic



Stopping Rules

Representation of two-sided hypothesis tests

Two-sided tests take on appearance of two superposed hypothesis tests

- Lower test
  - $H_{0-}: \theta \geq \theta_{0-}$ versus $H_{-}: \theta \leq \theta_{-}$
- Upper test
  - $H_{0+}: \theta \leq \theta_{0+}$ versus $H_{+}: \theta \geq \theta_{+}$
- Classic two-sided test:

  $\theta_{0-} = \theta_{0+} = \theta_0$

  $\theta_{-} = -\theta_{+}$

Stopping Rules

Generalization of hypothesis tests

Require only $\theta_{-} \leq \theta_{0+} \leq \theta_{0-} \leq \theta_{+}$

Correspondence between hypotheses and boundaries

- 'a' boundary rejects $H_{0-}: \theta \geq \theta_{0-}$
- 'b' boundary rejects $H_{-}: \theta \leq \theta_{-}$
- 'c' boundary rejects $H_{+}: \theta \geq \theta_{+}$
- 'd' boundary rejects $H_{0+}: \theta \leq \theta_{0+}$

Stopping Rules

Correspondence to classical tests of $H_0: \theta = \theta_0$

- One-sided tests of greater alternative (upper and lower tests coincident)

  $\theta_{-} < \theta_{0-} = \theta_0$ (define $\theta_{0+} = \theta_{-}$ and $\theta_{+} = \theta_{0-}$)

- One-sided tests of lesser alternative (upper and lower tests coincident)

  $\theta_0 = \theta_{0+} < \theta_{+}$ (define $\theta_{-} = \theta_{0+}$ and $\theta_{0-} = \theta_{+}$)

- Two-sided tests

  $\theta_{-} < \theta_{0-} = \theta_0 = \theta_{0+} < \theta_{+}$ (with $\theta_{-} = -\theta_{+}$)

Stopping Rules

Parameterize hypotheses by shift parameters ε_L, ε_U

- $0 \le \varepsilon_L \le 1$ is shift of θ_{0-} away from θ_+ toward θ_0
  $\theta_{0-} = \theta_+ - \varepsilon_L \cdot (\theta_+ - \theta_0)$
- $0 \le \varepsilon_U \le 1$ is shift of θ_{0+} away from θ_- toward θ_0
  $\theta_{0+} = \theta_- + \varepsilon_U \cdot (\theta_0 - \theta_-)$
- Constraint: $1 \le \varepsilon_L + \varepsilon_U \le 2$
- Test can be thought of as $(\varepsilon_L + \varepsilon_U)$-sided

Stopping Rules

Parameterization special cases

One-sided test of greater alternative:
- $\varepsilon_L = 0 \quad \varepsilon_U = 1$

One-sided test of lesser alternative:
- $\varepsilon_L = 1 \quad \varepsilon_U = 0$

Two-sided test:
- $\varepsilon_L = 1 \quad \varepsilon_U = 1$

One-sided equivalence (noninferiority) test:
- $\varepsilon_L = 0.5 \quad \varepsilon_U = 0.5$

Stopping Rules

Number and timing of analyses

N counts the sampling units accrued to the study

Up to J analyses of the data to be performed

Analyses performed after accruing sample sizes of
$N_1 < N_2 < \cdots < N_J$

(More generally, N measures statistical information)

Stopping Rules

Boundaries at the analyses

$a_j \le b_j \le c_j \le d_j$ are the 'a', 'b', 'c', and 'd' boundaries at the j-th analysis (when N_j observations)

At the final (J-th) analysis $a_J = b_J$ and $c_J = d_J$ to guarantee stopping

Stopping Rules

Boundary shape functions

Π_j measures the proportion of information accrued at the j-th analysis

- often $\Pi_j = N_j / N_J$

Boundary shape function $f(\Pi_j)$ is a monotonic function used to relate the dependence of boundaries at successive analyses on the information accrued to the study at that analysis

Stopping Rules

Formulation of stopping boundaries

At the j-th analysis

- a_j is determined by $\theta_a = \theta_{0-}$ and $f_a(\Pi_j)$
- b_j is determined by $\theta_b = \theta_{-}$ and $f_b(\Pi_j)$
- c_j is determined by $\theta_c = \theta_{+}$ and $f_c(\Pi_j)$
- d_j is determined by $\theta_d = \theta_{0+}$ and $f_d(\Pi_j)$

Stopping Rules

Parameterization of boundary shape functions

$$f_*(\Pi_j) = [A_* + \Pi_j^{-P_*}(1-\Pi_j)^{R_*}] \times G_*$$

Distinct parameters possible for each boundary

Parameters A_*, P_*, R_* typically chosen by user

Critical value G_* usually calculated from search

Boundary Scales

Boundary Scales

Choices for test statistic T_n

- Sum of observations
- Point estimate of treatment effect
- Normalized (Z) statistic
- Fixed sample P value
- Error spending function
- Conditional probability
- Predictive probability
- Bayesian posterior probability

Boundary Scales

Choices for test statistic T_n

All of those choices for test statistics can be shown to be transformations of each other

Hence, a stopping rule for one test statistic is easily transformed to a stopping rule for a different test statistic

We regard these statistics as representing different scales for expressing the boundaries

Boundary Scales: Notation

One sample inference about means

Generalizable to most other commonly used models

Probability model: $X_1, \ldots, X_N \quad iid \quad (\mu, \sigma^2)$

Null hypothesis: $H_0 : \mu = \mu_0$

Analyses after $N_1, \ldots, N_J = N$

Data at jth analysis: $x_1, \ldots, x_{N_j}$

Distributional assumptions:

in absence of a stopping rule $\overline{X}_j \dot{\sim} N\left(\mu, \frac{\sigma^2}{N_j}\right)$

Boundary Scales

Partial Sum Scale:

$$S_j = \sum_{i=1}^{N_j} X_i$$

Uses:

- Cumulative number of events
- Convenient when computing density

Boundary Scales

Sample Mean Scale:

$$\overline{x}_j = \frac{1}{N_j}\sum_{i=1}^{N_j} x_i = \frac{s_j}{N_j}$$

Uses:

Natural estimate of treatment effect

Boundary Scales

Normalized Statistic Scale:

$$z_j = \sqrt{N_j}\,\frac{\left[\overline{x}_j - \mu_0\right]}{\sigma}$$

Uses:

Commonly computed in analysis routines

Boundary Scales

Fixed Sample P value Scale:

$$p_j = 1 - \Phi\left(z_j\right)$$

$$= 1 - \int_{-\infty}^{z_j} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\,du$$

Uses:

Commonly computed in analysis routine

Robust to use with other distributions for estimates of treatment effect

Boundary Scales

Bayesian Posterior Scale:

Prior $\mu \sim N\left(\zeta, \tau^2\right)$

$$B_j\left(\mu_*\right) = \Pr\left(\mu \geq \mu_* \mid \left(X_{1,\ldots,} X_{N_j}\right)\right)$$

$$= 1 - \Phi\left(\frac{\mu_*\left[N_j\tau^2 + \sigma^2\right] - N_j\tau^2\overline{x}_j - \sigma^2\zeta}{\sigma\tau\sqrt{N_j\tau^2 + \sigma^2}}\right)$$

Uses:

Bayesian inference (unaffected by stopping)

Posterior probability of hypotheses

Boundary Scales

Conditional Probability Scale:

Threshold at final analysis $t_{\overline{X}J}$

Hypothesized value of mean μ_*

$$C_j\left(t_{\overline{X}J},\mu_*\right) = \Pr\left(\overline{X}_J \geq t_{\overline{X}J} \mid \overline{X}_j;\mu = \mu_*\right)$$

$$= 1 - \Phi\left(\frac{N_J\left[t_{\overline{X}J} - \mu_*\right] - N_j\left[\overline{x}_j - \mu_*\right]}{\sigma\sqrt{N_J - N_j}}\right)$$

Uses:

Conditional power

Futility of continuing under specific hypothesis

Boundary Scales

Conditional Probability (estimate) Scale:

Threshold at final analysis $t_{\overline{X}J}$

$$C_j\left(t_{\overline{X}J},\mu_*=\overline{X}_j\right) = \Pr\left(\overline{X}_J \geq t_{\overline{X}J} \mid \overline{X}_j;\mu = \overline{x}_j\right)$$

$$= 1 - \Phi\left(\frac{N_J\left[t_{\overline{X}J} - \overline{x}_j\right]}{\sigma\sqrt{N_J - N_j}}\right)$$

Uses:

Futility of continuing using best estimate

Boundary Scales

Predictive Probability Scale:

Prior distribution $\mu \sim N\left(\zeta, \tau^2\right)$

$$H_j\left(t_{\overline{X}j}\right) = \int \Pr\left(\overline{X}_j \geq t_{\overline{X}j} \mid \overline{X}_j, \mu\right) \lambda\left(\mu \mid \overline{X}_j\right) d\mu$$

$$= 1 - \Phi\left(\frac{N_J\left[N_j\tau^2 + \sigma^2\right]\left[t_{\overline{X}j} - \overline{x}_j\right] + \sigma^2\left[N_J - N_j\right]\left[\overline{x}_j - \zeta\right]}{\sigma\sqrt{\left[N_J - N_j\right]\left[N_J\tau^2 + \sigma^2\right]\left[N_j\tau^2 + \sigma^2\right]}}\right)$$

Uses:

Futility of continuing study

Boundary Scales

Predictive Probability Scale:

Noninformative Prior $\mu \sim N\left(\zeta, \tau^2\right), \quad \tau^2 \to \infty$

$$H_j\left(t_{\overline{X}j}\right) = \int \Pr\left(\overline{X}_j \geq t_{\overline{X}j} \mid \overline{X}_j, \mu\right) \lambda\left(\mu \mid \overline{X}_j\right) d\mu$$

$$= 1 - \Phi\left(\frac{N_J\left[t_{\overline{X}j} - \overline{x}_j\right]}{\sigma\sqrt{\frac{N_J}{N_j}\left[N_J - N_j\right]}}\right)$$

Uses:

Futility of continuing study

Boundary Scales

Error Spending (outer lower boundary) Scale:

$$E_{aj} = \frac{1}{\alpha_\ell}\left[\sum_{i=1}^{j-1}\Pr\left(S_i \le a_i, \bigcap_{k=1}^{i-1}\left(S_k \in \left(a_k, b_k\right) \cup \left(c_k, d_k\right)\right) ; \mu_a\right) + \Pr\left(S_j \le s_j ; \mu_a\right)\right]$$

Uses:

Implementation of stopping rules with flexible determination of number and timing of analyses

Boundary Scales

Error Spending (inner lower boundary) Scale:

$$E_{bj} = \frac{1}{1-\beta_\ell}\left[\sum_{i=1}^{j-1}\Pr\left(S_i \ge b_i, \bigcap_{k=1}^{i-1}\left(S_k \in \left(a_k, b_k\right) \cup \left(c_k, d_k\right)\right) ; \mu_b\right) + \Pr\left(S_j \ge s_j ; \mu_{ba}\right)\right]$$

Uses:

Implementation of stopping rules with flexible determination of number and timing of analyses

Boundary Scales

Error Spending (inner upper boundary) Scale:

$$E_{cj} = \frac{1}{1-\beta_u}\left[\sum_{i=1}^{j-1}\Pr\left(S_i \le c_i, \bigcap_{k=1}^{i-1}\left(S_k \in \left(a_k, b_k\right) \cup \left(c_k, d_k\right)\right) ; \mu_c\right) + \Pr\left(S_j \le s_j ; \mu_c\right)\right]$$

Uses:

Implementation of stopping rules with flexible determination of number and timing of analyses

Boundary Scales

Error Spending (outer upper boundary) Scale:

$$E_{dj} = \frac{1}{\alpha_u}\left[\sum_{i=1}^{j-1}\Pr\left(S_i \ge d_i, \bigcap_{k=1}^{i-1}\left(S_k \in \left(a_k, b_k\right) \cup \left(c_k, d_k\right)\right) ; \mu_d\right) + \Pr\left(S_j \ge s_j ; \mu_d\right)\right]$$

Uses:

Implementation of stopping rules with flexible determination of number and timing of analyses

Boundary Scales

Use in evaluating designs

Several of the boundary scales have interpretations that are useful in evaluating the operating characteristics of a design

- Sample Mean Scale
- Conditional Probability Futility Scales
- Predictive Probability Futility Scale
- Bayesian Posterior Probability Scale
- (Error Spending Scale)

Unified Design Family

Unified Design Family

Unifying parameterization for the most commonly used group sequential designs (Kittelson & Emerson, 1999)

Rich parameterization facilitates search for stopping rule appropriate for specific applications

Inclusion of broad spectrum of designs means that comparisons within this family will consider full range of possible designs

(Default family in S+SeqTrial)

Unified Design Family

Stopping Boundaries for Sample Mean Statistic:

$$a_j = \mu_a - f_a(\Pi_j)$$
$$b_j = \mu_b + f_b(\Pi_j)$$
$$c_j = \mu_c - f_c(\Pi_j)$$
$$d_j = \mu_d + f_d(\Pi_j)$$

Unified Design Family

Parameterization of boundary shape functions

$$f_*(\Pi_j) = [A_* + \Pi_j^{-P_*}(1-\Pi_j)^{R_*}] \times G_*$$

Distinct parameters possible for each boundary

Parameters A_*, P_*, R_* typically chosen by user
Critical value G_* usually calculated from search

Unified Design Family

Choice of P parameter

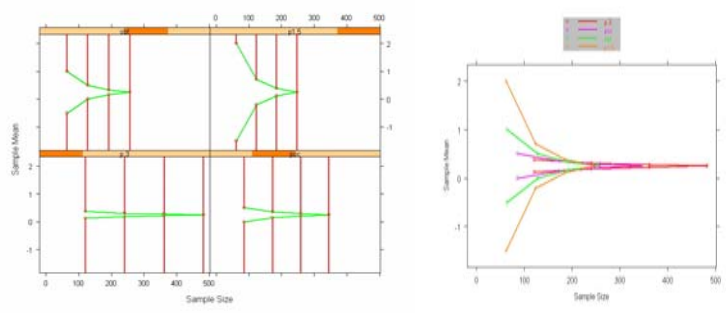
P ≥ 0:

- Larger positive values of P make early stopping more difficult (impossible when P infinite)
- When A=R=0, 0.5 < P < 1 corresponds to power family parameter (Δ) in Wang & Tsiatis (1987): P= 1 - Δ
- Reasonable range of values: 0 < P < 2.5
- P=0 with A=R=0 possible for some (not all) boundaries, but not particularly useful

Unified Design Family

Effect of varying P>0 (when A=0, R=0)

Higher P leads to early conservatism
P > 0 has infinite boundaries when N=0



Unified Design Family

Choice of P parameter

P < 0:

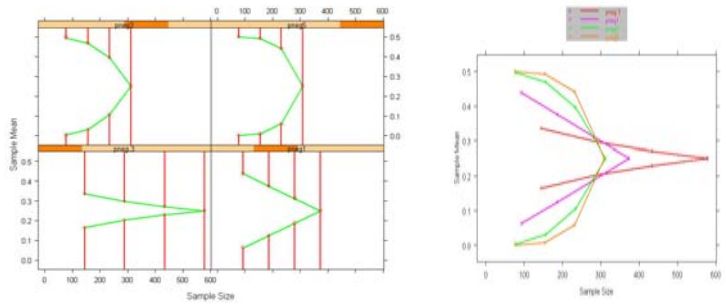
- Must have R = 0 and (typically) A < 0
- More negative values of P make early stopping more difficult

Unified Design Family

Effect of varying P<0 (when A=2, R=0)

More negative P leads to early conservatism

P < 0 has finite boundaries when N=0



Unified Design Family

Choice of R parameter

R > 0:

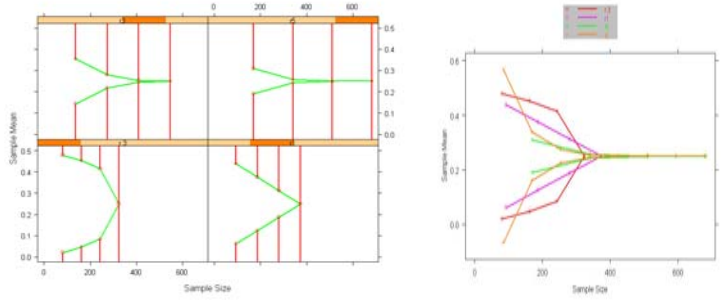
- Larger positive values of R make early stopping easier
- When R>0 and P=0, typically need A>0
- Reasonable range of values: 0.1 < R < 20
- R < 1 is convex outward
- R > 1 is convex inward
- When R>0 and P>0, can get change in convexity of boundaries

Unified Design Family

Effect of varying R (when A=1, P=0)

R < 1 leads to convex outward

R > 1 leads to convex inward

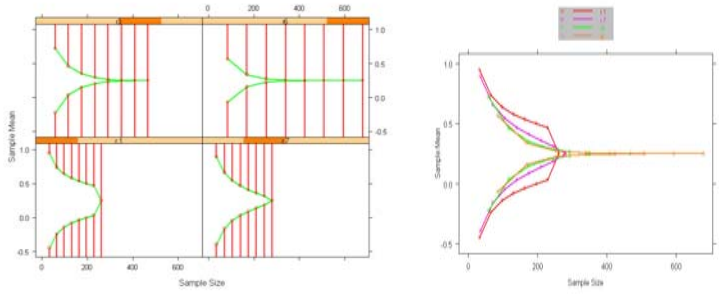


Unified Design Family

Effect of varying R (when A=1, P=0.5)

With P > 0, boundaries infinite when N=0

R < 1 and P > 0 has change in convexity



Unified Design Family

Choice of A parameter

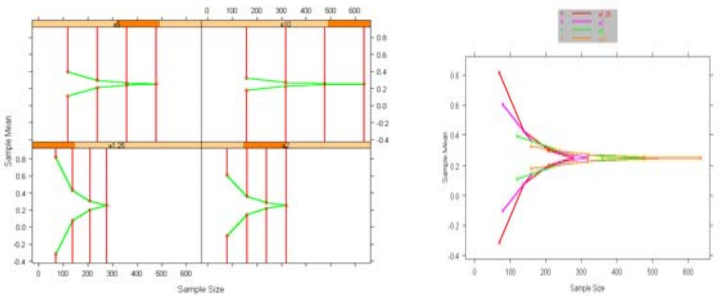
- Lower absolute values of A makes it harder to stop at early analyses
- Valid choices of A depend upon choices of P and R
- Useful ranges for A
  - $P \geq 0, R \geq 0$: $0.2 \leq A \leq 15$
  - $P \leq 0, R = 0$: $-15 \leq A \leq -1.25$

Unified Design Family

Effect of varying A (when P=0, R=1.2)

Values of A closer to 0 make it harder to stop early

Higher absolute value of A makes flatter boundaries



Unified Design Family

Parameterization of boundary shape function includes many previously described approaches

Wang & Tsiatis Boundary Shape Functions:

- $A_* = 0, R_* = 0, P_* > 0$
- P_* measures early conservatism
  - $P_* = 0.5$ Pocock (1977)
  - $P_* = 1.0$ O'Brien-Fleming (1979)
- ($P_* = \infty$ precludes early stopping)

Unified Design Family

Parameterization of boundary shape function includes many previously described approaches

Triangular Test Boundary Shape Functions (Whitehead)

- $A_* = 1, R_* = 0, P_* = 1$

Sequential Conditional Probability Ratio Test (Xiong):

- $R_* = 0.5, P_* = 0.5$

Unified Design Family

Parameterization of hypothesis shifts and boundary shape function unifies what were discrete families

Triangular tests vs Wang and Tsiatis based families

- Choice of A^*

One-sided vs two-sided tests

- Choice of ε_L, ε_U

Early stopping under one hypothesis vs both hypotheses

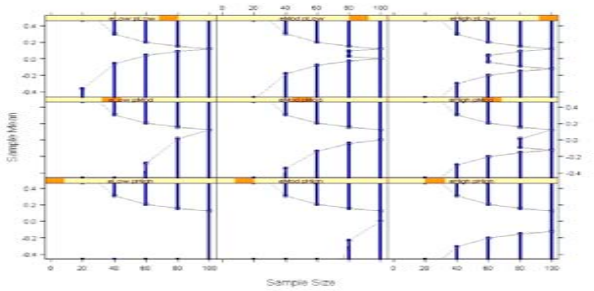
- Choice of P^*

Unified Design Family

Spectrum of designs

ε_L increases across rows

P_a and/or P_c increases down columns



Unified Design Family

Operating characteristics

User specifies size α_U, α_L of upper and lower tests

User specifies power β_U, β_L of upper and lower tests

Computer search for G_a, G_b, G_c, G_d that attains those operating characteristics

(Sample size can be computed using some other power besides β_U, β_L)

Error Spending Family

Error Spending Family

Lan and DeMets (1983) approach

At each analysis, some of the type I error is `used up'

Describe a stopping rule according to the proportion of α_U, α_L used at each analysis

- General case: alpha used by the j-th analysis determined by some function of the proportion of maximal information available

$$\alpha_j = f(\alpha, \Pi_j)$$

Error Spending Family

Lan and DeMets (1983) approach (cont.)

Lan and DeMets (1983) describe error spending functions comparable to O'Brien-Fleming or Pocock designs

- O'Brien-Fleming

$$\alpha_j = 2\left[1 - \Phi\left(z_{1-\alpha/2} / \sqrt{\Pi_j}\right)\right]$$

- Pocock

$$\alpha_j = \alpha \log\left[1 + (e-1)\Pi_j\right]$$

Error Spending Family

Lan and DeMets (1983) approach (cont.)

Lan and DeMets (1983) describe error spending functions comparable to O'Brien-Fleming or Pocock designs <u>for specific type I errors</u>

Error Spending Family

Lan and DeMets (1983) approach (cont.)

More recently authors have focussed on error spending functions of the form

$$\alpha_j = f(\alpha, \Pi_j) = \alpha\, f(\Pi_j)$$

(Kim and DeMets, 1987; Jennison and Turnbull, 1989; Hwang, Shih, and DeCani, 1990)

Error Spending Family

Lan and DeMets (1983) approach (cont.)

Kim and DeMets (1987) and Jennison and Turnbull (1989) consider an error spending family corresponding to

$$\alpha_j = \alpha \ \Pi_j^{-P}$$

Useful special cases identified by those authors:

- P = 1 is similar to Pocock (1977)
- P = 3 is similar to O'Brien and Fleming (1979)

Error Spending Family

Pampallona, Tsiatis, and Kim (1995) extension

Defines type II error spending functions

At each analysis, recompute maximal sample size which will maintain planned level of significance and power

Error Spending Family

Implementation of an Error Spending Family

Define stopping rule on error spending function scale by defining E_{aj}, E_{bj}, E_{cj}, E_{dj}

Use framework of superposed one-sided hypothesis tests described by Kittelson and Emerson (1999) to define relationships among hypotheses rejected by each of the four possible stopping boundaries

Error Spending Family

Correspondence with type I and II error spending

For user specified size α_U, α_L of upper and lower tests and power β_U, β_L of upper and lower tests, error spent at the j-th analysis specified as:

$$\alpha_{Lj} = \alpha_L \ E_{aj}$$
$$1 - \beta_{Lj} = (1 - \beta_L) \ E_{bj}$$
$$1 - \beta_{Uj} = (1 - \beta_U) \ E_{cj}$$
$$\alpha_{Uj} = \alpha_U \ E_{dj}$$

Error Spending Family

Boundary shape functions

Boundary shape function can be defined separately for each of the four boundaries

$$E_{*j} = f_*(\Pi_j)$$

$$f_*(\Pi_j) = \left[A_* + \Pi_j^{-P_*}(1-\Pi_j)^R\right]G_*$$

Error Spending Family

Constraints on parameters

f(0) = 0 and f(1) = 1

If P < 0

- R = 0, A = 1, G = 1

If R > 0

- P = 0, A = -1, G = -1

If P = 0 and R = 0, no early stopping

Error Spending Family

Computer search for stopping boundaries

Error spending family defines E_{aj}, E_{bj}, E_{cj}, E_{dj}

Appendix of Kittelson and Emerson (1999) describes general algorithm for finding design when hypotheses known

At design stage, must search for standardized hypotheses that result in a valid design, and then compute sample size to map standardized design to specified alternative hypotheses.

Error Spending Family

Computer search for stopping boundaries (cont.)

In order to more easily obtain more efficient designs, when designing a study using error spending functions, the specified type II error spending functions are only used as upper bounds on the true type II error spending function.

Comparison of Parameterizations

Comparison of Parameterizations

General comments

Families also defined for other boundary scales

- Partial sum and Z statistic scale families implemented in S+SeqTrial
- Bayesian and Futility scale families under construction

If stopping rules are carefully evaluated, it does not matter too much which scale (and therefore family) is used to derive the stopping rule.

Comparison of Parameterizations

General comments (cont.)

The best design family to use will be the one which allows a user to most quickly find a stopping rule having desirable operating characteristics

The ease of use will therefore depend in part on

- Interpretability of boundary scale
- Interpretability of parameters

Comparison of Parameterizations

General comments (cont.)

My view:

- Sample mean scale (unified family) has easier scientific interpretation than the error spending scale which has a purely statistical interpretation that, in my experience, is poorly understood by both users and researchers
- The parameterization of the unified family produces a more useful grouping of designs on some level than does the parameterization of the error spending family

Comparison of Parameterizations

ASSERTION: Interpretability of boundary scales

The concept of an error spending scale is less relevant to clinical researchers

- Type I error reflects only statistical evidence
- May conflict with scientific importance
  - Underpowered studies: Failure to reject the null in the face of large estimates of treatment effect
  - Overpowered studies: Rejection of the null hypothesis when differences are scientifically unimportant

Comparison of Parameterizations

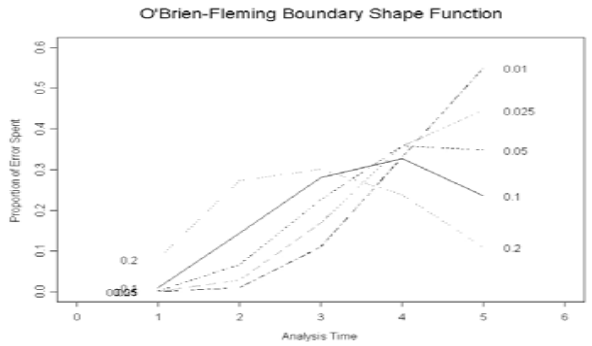
ASSERTION: Interpretability of boundary scales

The formulation of error spending scales is not well understood by the researchers developing such methods

- Lan & DeMets (1983), Kim & DeMets (1987), Jennison & Turnbull (1989 and 2000) all describe error spending functions which mimic O'Brien-Fleming (1979) or Pocock (1977) group sequential designs
- In fact, for different levels of type I (or type II) error, the error spending functions are different within those families of designs

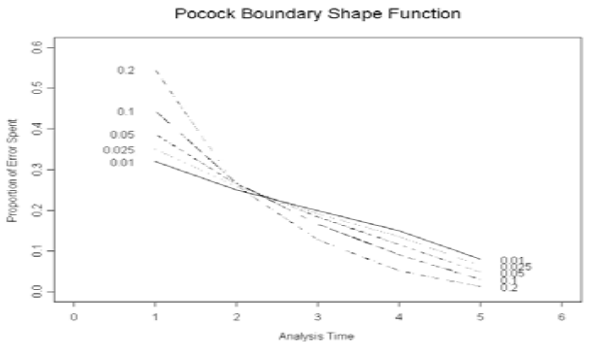
Comparison of Parameterizations

Error spent at each analysis for O'Brien and Fleming (1979) designs depends on Type I or Type II errors



Comparison of Parameterizations

Error spent at each analysis for Pocock (1977) designs depends on Type I or Type II errors



Comparison of Parameterizations

Is there a problem?

Parameterization of stopping rule families induces a grouping of designs:

- Unified family: Pocock (1977) designs, O'Brien-Fleming (1979) designs, Triangular designs (Whitehead & Stratton, 1983)
- Error spending families: All designs that spend the same proportion of type I or II error at each analysis

Comparison of Parameterizations

Is there a problem? (cont.)

Best parameterization might be defined according to whether such groupings correspond to similar operating characteristics

- efficiency
- Bayesian properties
- futility properties
- others

Comparison of Parameterizations

Efficiency

Consider ability of choice of boundary shape parameter to predict efficiency of design

- No uniformly most powerful design
- Efficiency measured in terms of smallest average sample size for specific hypothesis
  - Measure alternative hypothesis according to the power of the test to detect it

Comparison of Parameterizations

Methods for comparison

Find optimal designs in terms of average sample size (ASN) within family of Wang and Tsiatis (1987) boundary shape functions for one-sided symmetric designs (Emerson and Fleming, 1989)

- Family found to be approximately optimal

Find optimal designs for various choices of type I error and statistical power

Comparison of Parameterizations

Methods for comparison (cont.)

For each optimal design, examine the boundary shape function on

- Sample mean scale
- Error spending scale
- Futility scales

Comparison of Parameterizations

Criteria for "good" parameterizations

If the boundary shape function on a given scale is not independent of choice of type I and II errors, then that would argue that grouping of designs according to parameterization of that scale will not correspond to similar efficiency properties

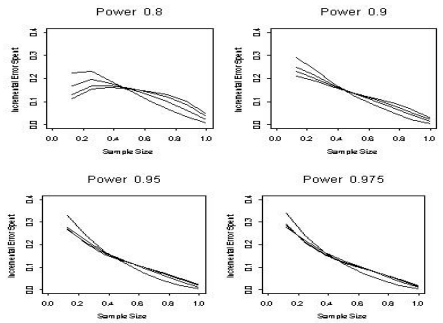
As it is unlikely that boundary shape parameters for efficient designs will be constant across all choices of type I and type II errors, we can also compare the degree that boundary shape parameters change for each boundary scale

Comparison of Parameterizations

Proportion of error spent at each analysis for approximately efficient designs

Power varies across panels

Type I error varies across lines within each panel

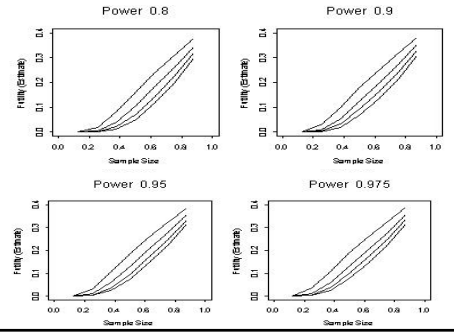


Comparison of Parameterizations

Conditional power (using MLE) at the boundary for each analysis for approximately efficient designs

Power varies across panels

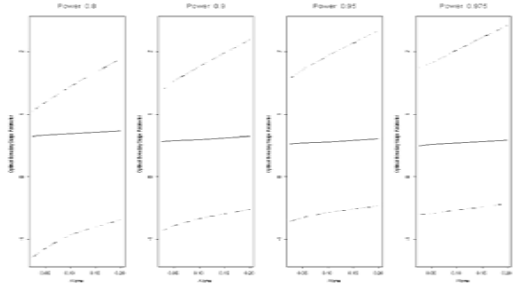
Type I error varies across lines within each panel



Comparison of Parameterizations

Comparison of optimal unified family P parameter as a function of type I errors

Compared to best fitting P or R parameter in error spending family



Comparison of Parameterizations

Search for stopping rule is generally iterative

- An initial design is specified
- Operating characteristics are examined
- Modifications are made to the design

Availability of tools for evaluation of operating characteristics lessens impact of family used to define a stopping rule

- Appropriate designs can be found from almost any starting point

Comparison of Parameterizations

To the extent that parameterization of sample mean family predicts efficiency behavior, use of that family may allow more intuitive search for suitable stopping rules

However, efficiency is not always of paramount concern

Comparison of Parameterizations

Interpretation of unified family boundaries as estimate of treatment effect is meaningful to clinical researcher

Error spending functions are less interpretable, and thus seem less useful when designing a clinical trial or evaluating its operating characteristics

However, error spending scale can be useful in implementing a stopping rule

Comparison of Parameterizations

It is not clear that conditional probabilities are particularly useful in the definition of a stopping rule

- Design family does not have a particularly intuitive parameterization
- Unconditional power considerations would seem more straightforward

Evaluation of Designs

Evaluation of Designs

Process of choosing a trial design

Define candidate design

Evaluate operating characteristics

Modify design

Iterate

Evaluation of Designs

Operating characteristics for fixed sample studies

Level of Significance (often pre-specified)
Sample size requirements
Power Curve
Decision Boundary
Frequentist inference on the Boundary
Bayesian posterior probabilities

Evaluation of Designs

Additional operating characteristics for group sequential studies

Probability distribution for sample size
Stopping probabilities
Boundaries at each analysis
Frequentist inference at each analysis
Bayesian inference at each analysis
Futility measures at each analysis

Evaluation of Designs

Sample size requirements

Number of subjects needed is a random variable

Quantify summary measures of sample size distribution

- maximum (feasibility of accrual)
- mean (Average Sample N- ASN)
- median, quartiles

(Particularly consider tradeoffs between power and sample size distribution)

Evaluation of Designs

Stopping probabilities

Consider probability of stopping at each analysis for arbitrary alternatives

Consider probability of each decision (for null or alternative) at each analysis

Evaluation of Designs

Power curve

Probability of rejecting null for arbitrary alternatives

- Power under null: level of significance
- Power for specified alternative

Alternative rejected by design

- Alternative for which study has high power

S+SeqTrial defines

- Power curves for upper and lower boundaries
- Alternatives having specified power for each boundary

Evaluation of Designs

Decision boundary at each analysis

Value of test statistic leading to rejection of null

- Variety of boundary scales possible

Often has meaning for applied researchers (especially on scale of estimated treatment effect)

- Estimated treatment effects may be viewed as unacceptable for ethical reasons based on prior notions
- Estimated treatment effect may be of little interest due to lack of clinical importance or futility of marketing

Evaluation of Designs

Frequentist inference on the boundary at each analysis

Consider P values, confidence intervals when observation corresponds to decision boundary at each analysis

Ensure desirable precision for negative studies

- Confidence interval identifies hypotheses not rejected by analysis
- Have all scientifically meaningful hypotheses been rejected?

Evaluation of Designs

Bayesian posterior probabilities at each analysis

Examine the degree to which the frequentist inference leads to sensible decisions under a range of prior distributions for the treatment effect

- Posterior probability of hypotheses

Bayesian estimates of treatment effect

- Median (mode) of posterior distribution
- Credible interval (quantiles of posterior distribution

Evaluation of Designs

Futility measures

Consider the probability that a different decision would result if trial continued

Can be based on particular hypotheses, current best estimate, or predictive probabilities

(Perhaps best measure of futility is whether the stopping rule has changed the power curve substantially)

S+SeqTrial Implementation

Evaluation of Designs

Forms of output from S+SeqTrial

- Printed output in report window or command line window
- Plots
- Named seqDesign object

S+SeqTrial Implementation

Evaluation of Designs (cont.)

Sample size requirements

- Printed with boundaries
- X axis with plots of boundaries
- Plots of average sample size, quantiles of sample size distribution

Stopping probabilities

- Printed with operating characteristics
- Plots with color coded decisions

S+SeqTrial Implementation

Evaluation of Designs (cont.)

Power Curve

- Hypotheses, size, power printed with boundaries
- Tabled power with summaries
- Plots of power curve
- Plots versus reference power curve

S+SeqTrial Implementation

Evaluation of Fixed Sample Designs (cont.)

Decision Boundary

- Printed on specified boundary scale
- Plots

Frequentist inference on the boundary

- Printed with summaries
- Plots

S+SeqTrial Implementation

Evaluation of Fixed Sample Designs (cont.)

Bayesian inference

- Posterior probabilities implemented as a boundary scale
- Median (mode) of posterior distribution
- Credible intervals

Futility measures

- Implemented as boundary scale
- Conditional and predictive approaches