Adaptive Methods for Clinical Trials Applications

Part I: Overview of Literature and a New Approach

Tze Leung Lai

Department of Statistics, Stanford University

October 18, 2011

- Selected topics from a monograph (2012, Springer):

  Sequential Experimentation in Clinical Trials: Design and Analysis

  Bartroff, Lai and Shih

- **Outline:**

  1. Brief survey of adaptive design
  2. Theory of sequential testing
     - ⋆ Fully sequential design
     - ⋆ Group sequential design
  3. A flexible and efficient approach to adaptive design
  4. Comparative studies

1. Brief Survey

- **Sequential learning and adaptation**
  - To address statistical problems for which there are no solutions with fixed sample size
    - Example: testing a normal mean $H_0 : \mu = \mu_0$ with unknown variance σ^2 (Dantzig, 1940)
    - Stein (1945) showed that a two-stage procedure can have power independent of σ^2
  - Adaptive designs
    - Use data during the course of a trial to learn about unknown parameters and thereby modify the design
    - Beyond nuisance parameters and sample size re-estimation

- **Examples of adaptation:**
  - Sample size re-estimation based on observed effect size
  - Drop arms, select dose
  - Change objective (eg, superiority vs. non-inferiority)
  - Choose primary endpoint
  - Enrich study population
  - Outcome-adaptive randomization

I-SPY 2: An Adaptive Breast Cancer Trial Design in the Setting of Neoadjuvant Chemotherapy

AD Barker[1], CC Sigman[2], GJ Kelloff[1], NM Hylton[3], DA Berry[4] and LJ Esserman[3]

I-SPY 2 (investigation of serial studies to predict your therapeutic response with imaging and molecular analysis 2) is a process targeting the rapid, focused clinical development of paired oncologic therapies and biomarkers. The framework is an adaptive phase II clinical trial design in the neoadjuvant setting for women with locally advanced breast cancer. I-SPY 2 is a collaborative effort among academic investigators, the National Cancer Institute, the US Food and Drug Administration, and the pharmaceutical and biotechnology industries under the auspices of the Foundation for the National Institutes of Health Biomarkers Consortium.

treatment options remain limited. These patients continue to represent a disproportionately large fraction of those who die of their disease. Given that the standard of care for these women increasingly includes neoadjuvant therapy prior to surgical resection, this combination of group and setting represents a unique opportunity to learn how to tailor the treatment to patients with high-risk breast cancers.

Cancer research from the past decade has shown that breast cancer is a number of heterogeneous diseases; this finding suggests that directing drugs to molecular pathways that characterize the disease in subsets of patients will improve treatment

Barker et al (2009)

For the assignment of drugs to patients, Bayesian methods of adaptive randomization[10] will be used to achieve a higher probability of efficacy. Drugs that do well within a specific molecular signature will be preferentially assigned within that signature and will progress through the trial more rapidly. Each drug's Bayesian predictive probability[10] of being successful in a phase III confirmatory trial will be calculated for each possible signature. Drugs will be dropped from the trial for reasons of futility when this probability drops sufficiently low for all signatures. Drugs will be graduated at an interim point, should this probability reach a sufficient level for one or more signatures. Drugs that have high Bayesian predictive probability of being more effective than standard therapy will graduate along with their corresponding biomarker signatures, allowing these agent–biomarker(s) combinations to be tested in smaller phase III trials. When the drug graduates, its predictive probability will be provided to the company for all the signatures tested. Depending on the patient accrual rate, new drugs can be added at any time during the trial as other drugs are either dropped or graduated.

Barker et al (2009)

- Most of the literature on adaptive designs focus on the prototypical problem of testing a normal mean when the variance is known.
- When variance is unknown , we need "internal pilot" to estimate the variance.
- Problem: $X_1, X_2, \ldots \sim N(\mu_X, \sigma^2)$ and $Y_1, Y_2, \ldots \sim N(\mu_Y, \sigma^2)$. Test $H_0 : \mu_X = \mu_Y$ vs. $H_A : \mu_X \neq \mu_Y$

Stein's two-stage procedure: use first stage (internal pilot) to estimate the variance

- First stage: samples n_0 from each of the two normal populations and computes the usual estimate s_0^2 of σ^2
- Second stage:
  - Sample up to

$$n_1 = n_0 \vee \left[\left(t_{2n_0-2,\alpha/2} + t_{2n_0-2,\beta} \right)^2 \frac{2s_0^2}{\delta^2} \right]$$

    where at $|\mu_X - \mu_Y| = \delta$, $1-\beta$ is the desired power level
  - Reject H_0 if

$$\frac{|\bar{X}_{n_1} - Y_{n_1}|}{\sqrt{2s_0^2/n_1}} > t_{2n_0-2,\alpha/2}$$

- Many modifications of Stein's initial idea: different way to re-estimate the total sample size based on s_0^2

Mid-Course Sample Size Re-Estimation

Re-estimate total sample size based on the data accumulate so far at some interim

- Suppose $\sigma^2 = 1/2$, and $\theta = \mu_X - \mu_Y$
- n=original sample size
- After rn observations, $S_1 = \sum_{i=1}^{rn}(X_i - Y_i)$,

$$n^{-1/2} S_1 \sim N(r\theta\sqrt{n}, r)$$

- If change the second stage sample size to $\gamma(1-r)n$, and $S_2 = \sum_{i=rn+1}^{n^*}(X_i - Y_i)$, then given the first stage data,

$$(n\gamma)^{-1/2} S_2 \sim N((1-r)\theta\sqrt{\gamma n}, 1-r)$$

Mid-Course Sample Size Re-Estimation

Under $H_0 : \theta = 0$, Fisher's (1998) test statistic

$$n^{-1/2}\left(S_1 + \gamma^{-1/2} S_2\right) \sim N(0,1) \tag{1}$$

- Variance spending test: to ensure the variance $r + (1 - r)$
- Jennison and Turnbull (2003): Fisher's test perform poorly with lower efficiency and power compared to group sequential tests.
- The inefficiency is due to the non-sufficient "weighted" statistic (1)

Mid-course modification of the maximum sample size

- Raised by Cui, Hung, and Wang (1999)
- Motivation example: observe at the interim that the drug achieved a reduction that was only half of the target reduction assumed in calculating maximum sample size M
- Increased sample size to $\tilde{M}$
- Allow the future group sizes to be increased of decreased at the interim

Optimal adaptive group sequential designs via dynamic programming

Jennison and Turnbull (2006):

- choose the jth group size and stopping boundary based on the cumulative sample size n_{j-1} and sample sum $S_{n_{j-1}}$
- Solve the problem numerically by backward induction algorithms
- Optimality: minimize a weighted average of the expected sample size subject to prescribed error probabilities
  - Ex: $(E_0(T) + E_{\theta_1}(T) + E_{2\theta_1}(T))/3$
- Efficiency: non-adaptive group sequential tests with optimally chosen first stage ~ optimal adaptive design (but more complicated!)

- **Trade-offs:**
  - **Flexibility vs. efficiency**:
    - ★ Tsiatis & Mehta (2003) showed that standard group sequential tests based on the likelihood ratio statistic are uniformly more powerful than certain adaptive designs, e.g., Cui et al (1999).
    - ★ Jennison & Turnbull (2003) gave a general weighted form of these adaptive designs and demonstrated that they performed much worse than group sequential tests.
    - ★ Jennison & Turnbull (2006a) introduced adaptive group sequential tests that are optimal in the sense of minimizing a weighted average of expected sample sizes over a collection of parameter values.
    - ★ Jennison & Turnbull (2006b) showed standard (non-adaptive) group sequential tests with the first stage chosen optimally are nearly as efficient.
  - **Complexity** in study implementation and analysis

2. Theory of Sequential Testing

- *Sequential Analysis* was born in response to demands for more efficient testing of weapons during World War II
- **Wald's (1943)** *sequential probability ratio test (SPRT)*
  - Suppose $X_1, X_2, \ldots \overset{iid}{\sim} f$
  - Test $H_0 : \ f = f_0$ vs. $H_1 : \ f = f_1$
  - Likelihood ratio $R_n = \prod_{i=1}^{n} \{f_1(X_i)/f_0(X_i)\}$
  - SPRT stops sampling at sample size

$$T = \inf\{n \geq 1 : \ R_n \geq B \text{ or } R_n \leq A\}$$

    Accepts H_0 (or H_1) if $R_T \leq A$ (or $R_T \geq B$).
  - Conjectured SPRT minimizes the expected sample size at H_0 and H_1 among all tests satisfying type I and II error rate constraints
  - Wald's approximations: $A \approx \log(\frac{\tilde{\alpha}}{1-\alpha}), \ B \approx \log(\frac{1-\tilde{\alpha}}{\alpha})$

- **Wald & Wolfowitz (1948):** Optimality of SPRT
  - Minimizes both $E_0(T)$ and $E_1(T)$ under error probability constraints at H_0 and H_1

- **Issue:**
  - $X_1, X_2, \ldots \overset{\text{iid}}{\sim} f_\theta$, a one-parameter exponential family with natural parameter θ.
  - $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta \geq \theta_1 (> \theta_0)$
  - The maximum expected sample size over θ of SPRT can be considerably larger than that of the optimal FSS test.

- **Kiefer-Weiss (1957) problem:**
  - Minimize $E_{\theta^*}(T)$ at a given θ^*, subject to error probability constraints at θ_0 and θ_1.
- **Hoeffding (1960):**
  - Gives a lower bound for $E_{\theta^*}(T)$ subject to error probability constraints at θ_0 and θ_1.
- **Lorden (1976):**
  - An asymptotic solution to the Kiefer-Weiss problem is a 2-SPRT:

$$\widetilde{N} = \inf\left\{n \geq 1 : \prod_{i=1}^{n} \frac{f_{\theta^*}(X_i)}{f_{\theta_0}(X_i)} \geq A_0 \text{ or } \prod_{i=1}^{n} \frac{f_{\theta^*}(X_i)}{f_{\theta_1}(X_i)} \geq A_1\right\}$$

  - In the case of normal mean, it reduces to the triangular test of Anderson (1960), which is close to the optimal boundary in Lai (1973).

- Ideally θ^* should be chosen to be true θ
- **Sequential generalized likelihood ratio (GLR) test:**
  - Replace θ^* with $\hat{\theta}_n$ at stage n
  - The test of $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta \geq \theta_1$ stops at

$$\widetilde{N} = \inf \left\{ n \geq 1 : \prod_{i=1}^{n} \frac{f_{\hat{\theta}_n}(X_i)}{f_{\theta_0}(X_i)} \geq A_0^{(n)} \text{ or } \prod_{i=1}^{n} \frac{f_{\hat{\theta}_n}(X_i)}{f_{\theta_1}(X_i)} \geq A_1^{(n)} \right\}$$

  - With $A_0^{(n)} = A_1^{(n)} = 1/c$, it is an asymptotic solution to the Bayes problem of testing H_0 versus H_1 with 0-1 loss and cost c, as $c \to 0$ (Schwartz, 1962).

- Chernoff (1961, 1965) derived an approximation to the Bayes test of $H_0' : \theta < \theta_0$ versus $H_1' : \theta > \theta_0$.
- **Lai (1988)**: One-parameter exponential family

$$\widehat{N} = \inf \left\{ n \geq 1 : \max \left[\prod_{i=1}^{n} \frac{f_{\hat{\theta}_n}(X_i)}{f_{\theta_0}(X_i)}, \prod_{i=1}^{n} \frac{f_{\hat{\theta}_n}(X_i)}{f_{\theta_1}(X_i)} \right] \geq e^{g(cn)} \right\},$$

where $g(t) \sim \log t^{-1}$ as $t \to 0$.

- In 1950's, it was recognized that sequential hypothesis testing might be useful in clinical trials (Armitage 1960).
- Armitage, McPherson and Rowe (1969) introduced *repeated significance test (RST)*:
  - Rationale: the strength of evidence is indicated by the results of a conventional significance test
  - For testing a normal mean μ with known variance σ^2, the RST of $H_0 : \mu = 0$ has the form

$$T = \inf\{n \le M : |S_n| \ge b\sigma\sqrt{n}\},$$

    rejecting H_0 if $T < M$ or if $T = M$ and $|S_M| \ge b\sigma\sqrt{M}$, where $S_n = X_1 + \cdots X_n$.
  - Developed a recursive numerical integration to compute overall significance level.
- Haybittle (1971) proposed a modification to increase power:
  - Reject H_0 if $T < M$ or if $T = M$ and $|S_M| \ge c\sigma\sqrt{M}$, where $b \ge c$.

- Pocock (1977)
  - In clinical trials, it is typically not feasible to arrange for continuous examination of data
  - Introduced a "group sequential" version of RST:

$$T = \inf\{n \leq M : |S_n| \geq b\sigma\sqrt{n}\},$$

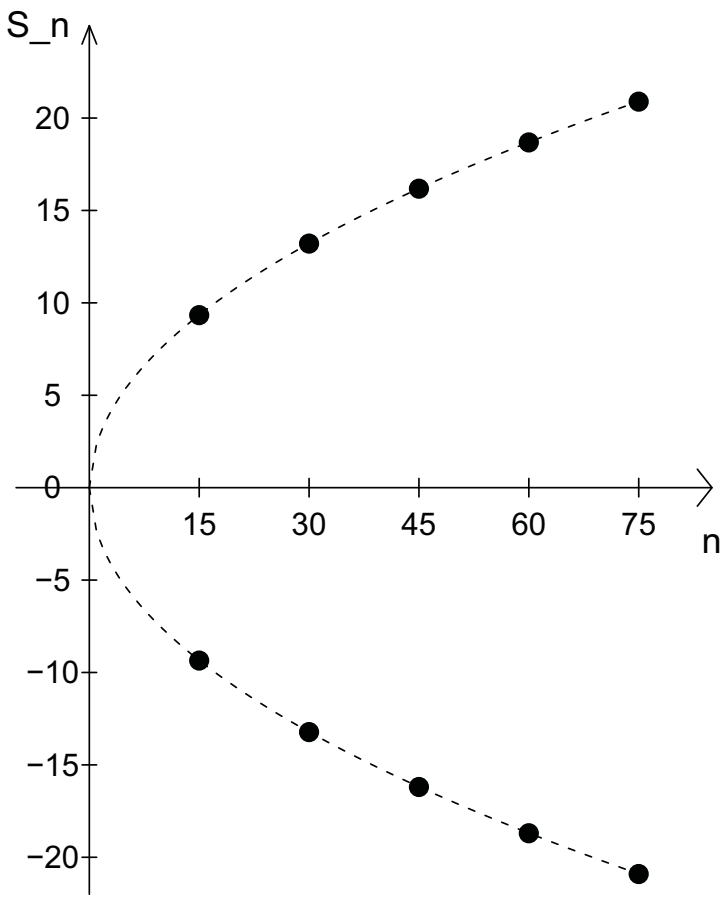
    where X_n is an approximately normally distributed statistic of data of the nth group, and M is the maximum number of looks.

- O'Brien and Fleming (1979)
  - Proposed a constant stopping boundary

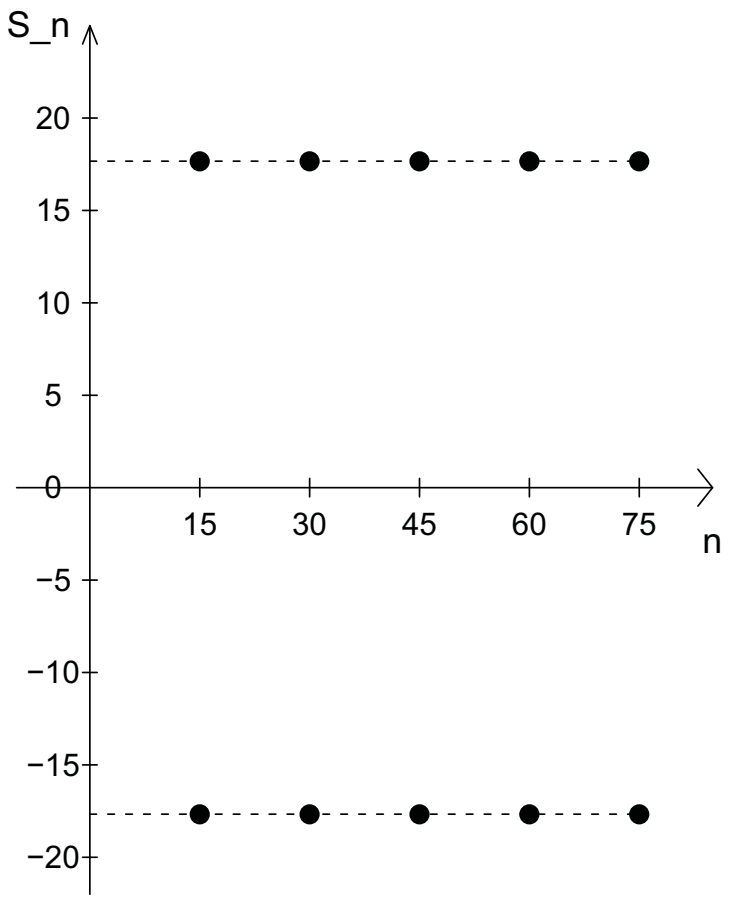
$$T = \inf\{n \leq M : |S_n| \geq b\}.$$

  - Corresponds to the group sequential version of an SPRT

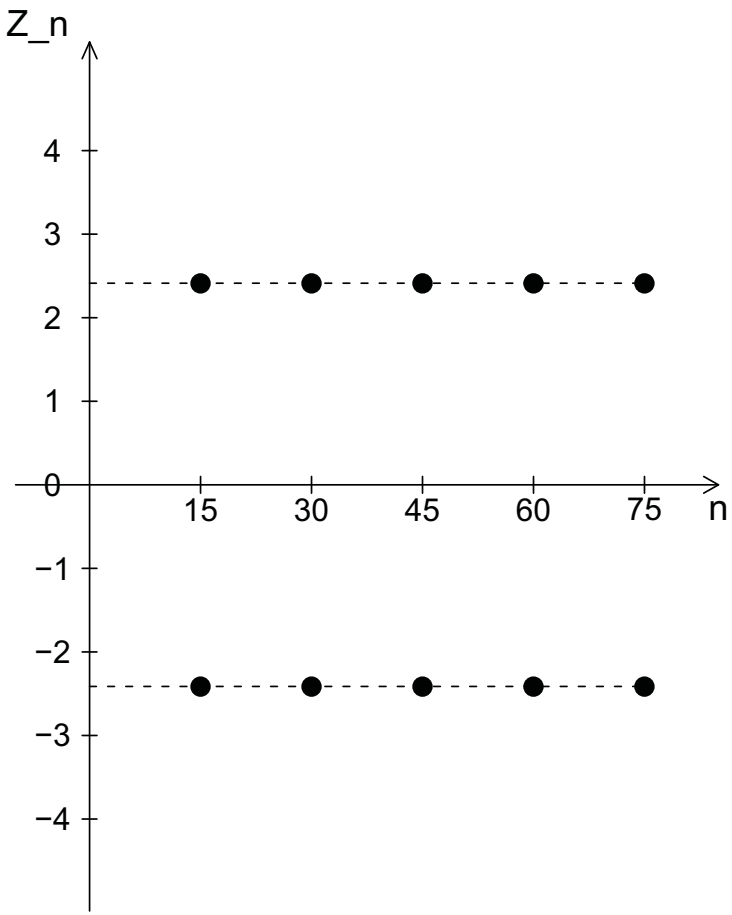
Pocock (1977)



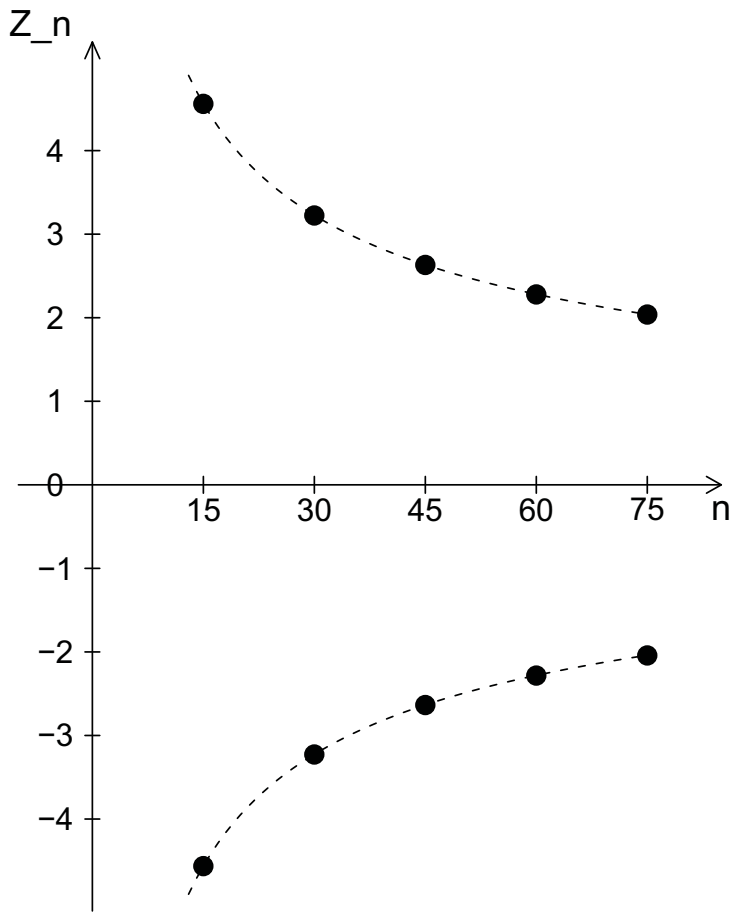
O'Brien−Fleming (1979)



Pocock (1977)



O'Brien−Fleming (1979)



For a group sequential design:

- $X_1, X_2, \ldots, X_M$ indep. $N(\mu, \sigma^2)$
- Want to test $H_0 : \mu = 0$ vs. $H_1 : \mu \neq 0$
- Let $S_n = X_1 + \cdots + X_n$, $\bar{X}_n = S_n/n$
- $(S_n - n\mu)/\sqrt{n\sigma^2} \sim N(0, 1)$
- Suppose there are k looks, with equal group sizes m
- Let $n_i = im$, $M = km$.

- Pocock (1977): Stop and reject H_0 if

$$|S_{n_i}| \geq b\sigma\sqrt{n_i}$$

- O'Brien and Fleming (1979): Stop and reject H_0 if

$$|S_{n_i}| \geq b$$

- Wang and Tsiatis (1987): Stop and reject H_0 if

$$\left|\frac{S_{n_i}}{\sqrt{n_i}}\right| \geq \sigma b \left(\frac{i}{k}\right)^{\delta-\frac{1}{2}},$$

where $0 \leq \delta \leq 0.7$

  - $\delta = 1/2$: Pocock; $\delta = 0$: O'Brien-Fleming

For one-sided hypothesis $H_0' : \mu \leq \mu_0$

- Want to stop not only when S_{n_i} exceeds an *upper* boundary (leading to rejection of H_0'), but also when S_{n_i} falls below a *lower* boundary (suggesting "futility")
- Futility boundary can be determined by considering an alternative $\mu_1 > \mu_0$
- Without loss of generality, assume $\mu_0 = -\mu_1$
- Power family and triangular tests

- **Power family**
  - Emerson and Fleming (1989), Pampallona and Tsiatis (1994)
  - Stop sampling at look $i \leq k-1$ if

$$S_{n_i} + \mu_1 n_i \geq b_i \sigma, \quad \text{rejecting } H_0,$$

$$\text{or } \; S_{n_i} - \mu_1 n_i \leq a_i \sigma, \quad \text{accepting } H_0.$$

  - If stopping does not occur before look k,

$$\text{reject } H_0 \text{ if } S_{n_k} + \mu_1 n_k \geq b_k \sigma.$$

  - The boundaries have the form

$$b_i = c_1(\delta) i^\delta m^{1/2}, \quad a_i = \left\{ 2i\theta_1/\sigma - c_2(\delta) i^\delta \right\} m^{1/2},$$

where $0 \leq \delta \leq 1/2$.

($\delta = 0$: O'Brien-Fleming; $\delta = 1/2$: Pocock)

- **Triangular tests**
  - Whitehead and Stratton (1983)
  - Stop at look $i \leq k-1$ if $|S_{n_i}| \geq b_i\sigma$, where

$$b_i = \left(\frac{\sigma}{\mu_1}\right)\log\left(\frac{1}{2\alpha}\right) - 0.583m^{1/2} - \frac{im\mu_1}{2\sigma}.$$

  - If stopping does not occur before look k,

$$\text{reject } H_0 \text{ if } S_{n_k} > 0.$$

  - This is a special case of Lorden's (1976) 2-SPRT.

The Lan-DeMets (1983) error spending approach

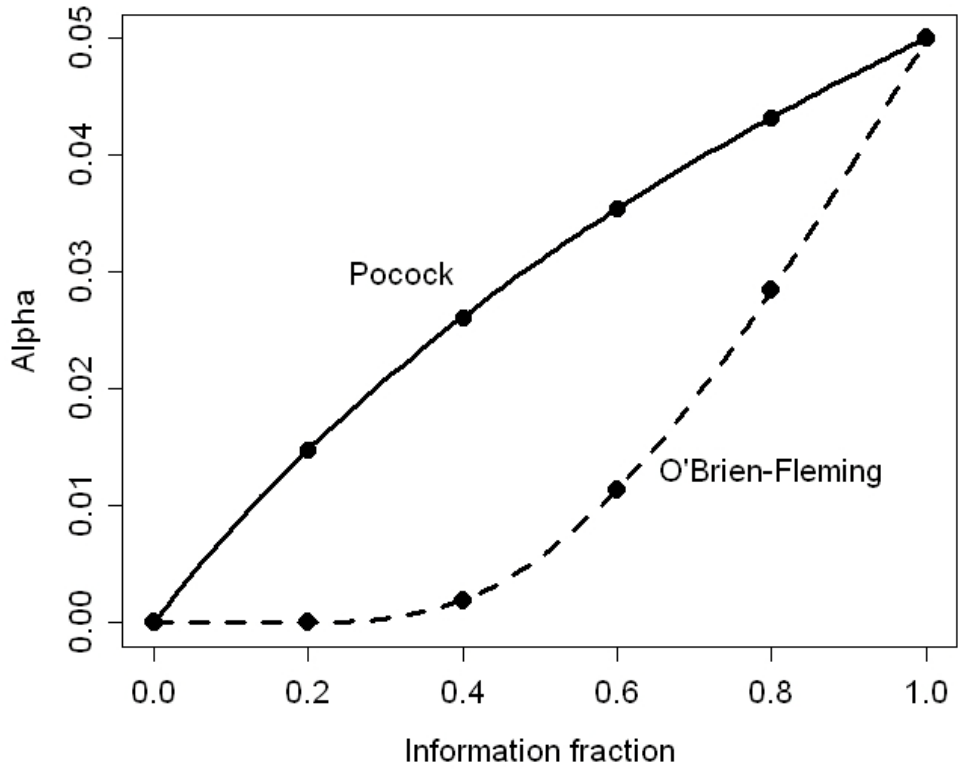
- In practice group sizes are usually unknown in advance and uneven
- **Key observation:** $(S_n/\sqrt{\sigma^2 M},\ 1 \le n \le M)$ has the same distribution as $(B_t,\ t \in \{1/M, \ldots, 1\})$.
- Given any stopping rule τ associated with a sequential test of the drift of a continuous Brownian motion, one can obtain a corresponding stopping rule for mean of X_i.
- Let $\pi(t) = P_0(\tau \le t)$ for $t < 1$.
- Given an error spending function $\pi(t)$, one can transform it to stopping boundaries for S_{n_i} via

$$P_0\left\{|S_{n_i}| \ge a_{n_i},\ \left|S_{n_j}\right| < a_{n_j} \text{ for } 1 \le j < i\right\} = \pi(n_i/M) - \pi(n_{i-1}/M)$$

for $1 \le i \le k-1$.

- Some examples:

$$\begin{aligned}
\pi(t) &= \min\{2 - 2\Phi(z_{\alpha/2}/\sqrt{t}), \alpha\} \quad \text{O'Brien-Fleming} \\
\pi(t) &= \min\{\alpha \log[1 + (e-1)t], \alpha\} \quad \text{Pocock} \\
\pi(t) &= \alpha \min\{t^{\rho}, 1\}, \rho > 0
\end{aligned}$$



Group sequential GLR tests with modified Haybittle-Peto boundaries

- First consider a one-parameter exponential family

$$f_\theta(x) = \exp\left(\theta x - \psi(\theta)\right)$$

- Test $H_0 : \theta \leq \theta_0$ at significance level α
- No more than M observations
- Consider group sequential tests with k analyses and group sizes $n_1, n_2 - n_1, \ldots, n_k - n_{k-1}$ (where $n_k = M$)
- Let $S_n = X_1 + \cdots + X_n$, $\overline{X}_n = S_n/n$
- The Kullback-Leibler information number is

$$I(\gamma, \theta) = E_\gamma\left[\log\left\{f_\gamma(X_i)/f_\theta(X_i)\right\}\right] = (\gamma - \theta)\psi'(\theta) - \left\{\psi(\gamma) - \psi(\theta)\right\}.$$

- Fixed sample size (FSS) test that rejects H_0 if $S_M \geq c_\alpha$ has maximal power at any alternative $\theta > \theta_0$.

- Ideally, want group sequential tests to
  - allow early stopping
  - attain nearly minimal expected sample size
  - have small loss in power compared to FSS test

- Let $\theta(M)$= "implied" alternative by M at which the FSS test with M observations has power $1 - \widetilde{\alpha}$

Group sequential GLR test of $H_0 : \theta \leq \theta_0$, with modified Haybittle-Peto boundary, proceeds as follows:

- At the ith interim analysis with $1 \leq i \leq k-1$,
  - $\widehat{\theta}_{n_i} = (\psi')^{-1}(\overline{X}_{n_i})$ = MLE of θ based on $X_1, \ldots, X_{n_i}$
  - Stop the trial at ith analysis if

$$\hat{\theta}_{n_i} > \theta_0 \text{ and } n_i I(\hat{\theta}_{n_i}, \theta_0) \geq b \quad \text{(rejecting } H_0\text{)},$$

$$\text{or } \hat{\theta}_{n_i} < \theta(M) \text{ and } n_i I(\hat{\theta}_{n_i}, \theta(M)) \geq \tilde{b} \quad \text{(accepting } H_0\text{)}.$$

- If stopping does not occur before kth analysis,

$$\text{reject } H_0 \text{ if } S_{n_k} \geq c.$$

- The thresholds $b, \tilde{b}, c$ are chosen such that
  - $P_{\theta_0}(\text{test rejects } H_0) = \alpha$
  - $P_{\theta(M)}(\text{test rejects } H_0)$ does not differ much from the power $1 - \beta$ of the FSS test at $\theta(M)$.

- Choose $0 < \epsilon \leq \frac{1}{2}$ and define $\tilde{b}$ by

$$P_{\theta(M)}\left\{\hat{\theta}_{n_i} < \theta(M) \text{ and } n_i I\left(\hat{\theta}_{n_i}, \theta(M)\right) \geq \tilde{b} \text{ for some } 1 \leq i \leq k-1\right\} = \epsilon\beta.$$

- After determining $\tilde{b}$, define b and then c by

$$\sum_{j=1}^{k-1} P_{\theta_0}\left\{\hat{\theta}_{n_j} > \theta_0 \text{ and } n_j I\left(\hat{\theta}_{n_j}, \theta_0\right) \geq b,\ n_i I\left(\hat{\theta}_{n_i}, \theta_0\right) 1_{\{\hat{\theta}_{n_i} > \theta_0\}} < b \text{ and } n_i I\left(\hat{\theta}_{n_i}, \theta(M)\right) 1_{\{\hat{\theta}_{n_i} < \theta(M)\}} < \tilde{b} \text{ for } i < j\right\} = \epsilon\alpha,$$

$$P_{\theta_0}\left\{S_{n_k} \geq c,\ n_i I\left(\hat{\theta}_{n_i}, \theta_0\right) 1_{\{\hat{\theta}_{n_i} > \theta_0\}} < b \text{ and } n_i I\left(\hat{\theta}_{n_i}, \theta(M)\right) 1_{\{\hat{\theta}_{n_i} < \theta(M)\}} < \tilde{b} \text{ for } i < k\right\} = (1-\epsilon)\alpha.$$

- For $X_i \overset{\text{iid}}{\sim} N(\theta, 1)$,
  - $I(\theta, \lambda) = (\theta - \lambda)^2/2$
  - $n_i I(\widehat{\theta}_{n_i}, 0) = n_i \bar{X}_{n_i}^2/2 = S_{n_i}^2/(2n_i)$

- To test $H_0 : \theta = 0$, Haybittle (1971) and Peto et al (1976) proposed
  - for $1 \le i \le k-1$, stop & reject H_0 if $|S_{n_i}|/\sqrt{n_i} \ge 3$
  - for $i = k$, reject H_0 if $|S_{n_k}|/\sqrt{n_k} \ge c$

- The above group sequential GLR test is in spirit similar
  - called “modified Haybittle-Peto” test

Group sequential GLR test of $H_0 : \theta \leq \theta_0$, with modified Haybittle-Peto boundary, proceeds as follows:

- At the ith interim analysis with $1 \leq i \leq k-1$,
  - $\widehat{\theta}_{n_i} = (\psi')^{-1}(\overline{X}_{n_i})$ = MLE of θ based on $X_1, \ldots, X_{n_i}$
  - Stop the trial at ith analysis if

$$\hat{\theta}_{n_i} > \theta_0 \text{ and } n_i I(\hat{\theta}_{n_i}, \theta_0) \geq b \quad \text{(rejecting } H_0\text{)},$$

$$\text{or } \hat{\theta}_{n_i} < \theta(M) \text{ and } n_i I(\hat{\theta}_{n_i}, \theta(M)) \geq \tilde{b} \quad \text{(accepting } H_0\text{)}.$$

- If stopping does not occur before kth analysis,

$$\text{reject } H_0 \text{ if } S_{n_k} \geq c.$$

Two-sided tests without futility boundaries

- At the ith interim analysis with $1 \leq i \leq k-1$, stop the trial if

$$n_i I(\hat{\theta}_{n_i}, \theta_0) \geq b \quad \text{(rejecting } H_0\text{)}.$$

- If stopping does not occur before kth analysis,

$$\text{reject } H_0 \text{ if } n_k I(\hat{\theta}_{n_k}, \theta_0) \geq c.$$

Two-sided tests with futility boundaries

- At the ith interim analysis with $1 \le i \le k-1$, stop the trial if

$$n_i I(\hat{\theta}_{n_i}, \theta_0) \ge b \quad \text{(rejecting } H_0\text{)},$$

or

$$n_i I(\hat{\theta}_{n_i}, \theta_0) < b \text{ and } \left\{ n_i I\big(\hat{\theta}_{n_i}, \theta_-(M)\big) \ge \tilde{b}_- \text{ or } n_i I\big(\hat{\theta}_{n_i}, \theta_+(M)\big) \ge \tilde{b}_+ \right\}$$

(accepting H_0).

- If stopping does not occur before kth analysis,

$$\text{reject } H_0 \text{ if } n_k I(\hat{\theta}_{n_k}, \theta_0) \ge c.$$

The modified Haybittle-Peto test

- Uses more flexible boundary $b, \widetilde{b}, c$
- Generalizes to exponential families
  - $n_i I(\widehat{\theta}_{n_i}, \lambda)$ =GLR statistic for testing $\theta = \lambda$
  - Uses efficient statistics for the null and alternative
  - Applies to multi-armed and multi-parameter problems
    - ▷ for testing $u(\theta) = u_0$, GLR statistic is $\inf_{u(\theta)=u_0} n_i I(\widehat{\theta}_{n_i}, \theta)$
- Related to the Kiefer-Weiss problem for fully sequential tests
  - Attains the asymptotically minimal value of the expected sample size at every fixed θ, and has power at $\theta(M)$ comparable to its upper bound $1 - \beta$.

Theory: Lai & Shih (2004 Biometrika)
Appendix A of Bartroff, Lai & Shih

Theorem A.1. *Suppose the possible values of T are $n_1 < \cdots < n_k$, such that*

$$\liminf (n_i - n_{i-1})/|\log(\alpha+\beta)| > 0 \tag{A.14}$$

as $\alpha+\beta \to 0$, where α and β are the type I and type II error probabilities of the test at θ_0 and θ_1. Let $m_{\alpha,\beta}(\theta) = \min\{|\log\alpha|/I(\theta,\theta_0), |\log\beta|/I(\theta,\theta_1)\}$. Let $\varepsilon_{\alpha,\beta}$ be positive numbers such that $\varepsilon_{\alpha,\beta} \to 0$ as $\alpha+\beta \to 0$, and let ν be the smallest $j(\le k)$ such that $n_j \ge (1-\varepsilon_{\alpha,\beta})m_{\alpha,\beta}(\theta)$, defining ν to be k if no such j exists. Then for fixed θ, θ_0 and $\theta_1 > \theta_0$, as $\alpha+\beta \to 0$,

$$P_\theta(T \ge n_\nu) \to 1;$$

If furthermore $\nu < k$, $|m_{\alpha,\beta}(\theta) - n_\nu|/m_{\alpha,\beta}^{1/2}(\theta) \to 0$ and

$$\limsup \frac{m_{\alpha,\beta}(\theta)}{\max\{|\log\alpha|/I(\theta,\theta_0),\ |\log\beta|/I(\theta,\theta_1)\}} < 1, \tag{A.15}$$

then $P_\theta(T \ge n_{\nu+1}) \ge \frac{1}{2} + o(1)$.

Theorem A.2. *Let $\theta_0 < \theta^* < \theta_1$ be such that $I(\theta^*, \theta_0) = I(\theta^*, \theta_1)$. Let $\alpha + \beta \to 0$ such that $\log \alpha \sim \log \beta$.*

(i) The sample size n^* of the Neyman–Pearson test of θ_0 versus θ_1 with error probabilities α and β satisfies $n^* \sim |\log \alpha| / I(\theta^*, \theta_0)$.

(ii) For $L \geq 1$, let $\mathcal{T}_{\alpha,\beta,L}$ be the class of stopping times associated with group sequential tests with error probabilities not exceeding α and β at θ_0 and θ_1 and with k groups and prespecified group sizes such that (A.14) *holds and $n_k = n^* + L$. Then, for given θ and L, there exists $\tau \in \mathcal{T}_{\alpha,\beta,L}$ that stops sampling when*

$$\begin{aligned} &(\theta - \theta_0) S_{n_i} - n_i \{\psi(\theta) - \psi(\theta_0)\} \geq b \\ \text{or} \quad &(\theta - \theta_1) S_{n_i} - n_i \{\psi(\theta) - \psi(\theta_1)\} \geq \tilde{b} \end{aligned} \tag{A.16}$$

for $1 \leq i \leq k-1$, with $b \sim |\log \alpha| \sim \tilde{b}$, and such that

$$E_\theta(\tau) \sim \inf_{T \in \mathcal{T}_{\alpha,\beta,L}} E_\theta(T) \sim n_\nu + \rho(\theta)(n_{\nu+1} - n_\nu), \tag{A.17}$$

where ν and $m_{\alpha,\beta}(\theta)$ are defined in Theorem A.1 and $0 \leq \rho(\theta) \leq 1$.

Theorem A.3. *Let $\alpha + \beta \to 0$ such that $\log \alpha \sim \log \beta$. Suppose that the k group sizes satisfy* (A.14) *with $n_k = M \sim |\log \alpha| / I(\theta^*, \theta_0)$, where $\theta_0 < \theta^* < \theta(M)$ is defined by $I(\theta^*, \theta_0) = I(\theta^*, \theta(M))$.*

(i) For every fixed θ, $E_\theta(\tilde{\tau}) \sim n_\nu + \rho(\theta)(n_{\nu+1} - n_\nu)$, where ν and $\rho(\theta)$ are the same as in Theorem A.2 with $\theta_1 = \theta(M)$.

(ii) $p_{\theta(M)} = 1 - \beta - (\kappa_\varepsilon + o(1))\beta$, where $\kappa_\varepsilon \sim \{1 + (\theta(M) - \theta^*)/(\theta^* - \theta_0)\}\varepsilon$ as $\varepsilon \to 0$.

- **Numerical example:**
  - Test $H_0 : p_1 = p_2$ in a randomized two-armed trial
  - $k = 5$, $M = 100$
  - The sample size n_{ij} for the two treatments can be different at the jth analysis
  - The group size $n_j = n_{1j} + n_{2j}$ can vary over j

Power (%) and expected sample size (in parentheses) of two-sided group sequential tests of $H_0 : p_2 - p_1 = 0$ without futility boundaries.

	$p_2 - p_1$					
	0	0.15	0.2	0.24	0.27	0.30
(a) Equal group sizes, adaptive treatment allocation						
α_1^*	**6.1** (99.0)	34.4 (93.7)	52.2 (89.6)	69.0 (84.2)	78.1 (80.0)	85.5 (75.6)
α_2^*	**7.1** (97.3)	29.2 (90.1)	45.9 (83.4)	59.4 (77.4)	68.6 (73.3)	78.5 (66.0)
ModHP	5.5 (98.9)	36.1 (93.1)	55.9 (89.0)	69.6 (83.7)	79.4 (78.6)	86.2 (74.1)
(b) Unequal group sizes, even treatment allocation						
α_1^*	5.0 (99.4)	36.8 (94.0)	59.0 (89.0)	72.6 (84.8)	81.3 (80.7)	87.2 (76.9)
α_2^*	**7.2** (97.3)	33.3 (88.9)	51.0 (81.3)	64.7 (76.1)	76.7 (70.6)	82.8 (65.7)
ModHP	5.3 (98.7)	38.2 (93.0)	58.1 (88.0)	72.9 (82.5)	80.9 (77.7)	88.4 (72.5)
(c) Unequal group sizes, adaptive treatment allocation						
α_1^*	5.6 (99.3)	35.3 (94.2)	55.0 (89.7)	70.6 (85.0)	79.1 (81.5)	86.2 (77.5)
α_2^*	**6.9** (97.1)	27.9 (90.6)	45.1 (84.4)	59.8 (77.6)	69.5 (73.7)	78.9 (67.9)
ModHP	5.6 (98.6)	35.2 (93.0)	55.3 (87.9)	68.5 (84.0)	77.8 (78.8)	86.7 (74.1)

- The thresholds $b, \tilde{b}, c$ can be calculated via **recursive numerical integration**.

- Consider the prototype model $X_i \sim N(\theta, 1)$:
  - $\tau = \min\{i \le k : S_{n_i} \notin (a_i, b_i)\} \wedge k$
  - Let $f_i(x) = (d/dx)P_\theta\{\tau > i,\ S_{n_i} \le x\}$
  - Then $f_1(x) = \phi((x-\theta)/\sqrt{n_1})$ for $a_1 < x < b_1$
  - For $i > 1$ and $a_i < x < b_i$,

$$f_i(x) = \int_{a_{i-1}}^{b_{i-1}} f_{i-1}(y)\phi\left(\frac{x - y - \theta(n_i - n_{i-1})}{\sqrt{n_i - n_{i-1}}}\right) dy.$$

  - Moreover,

$$P(\tau = i) = \int_{a_{i-1}}^{b_{i-1}} f_{i-1}(y)\left\{\Phi\left(\frac{a_i - y - \theta(n_i - n_{i-1})}{\sqrt{n_i - n_{i-1}}}\right) + 1 - \Phi\left(\frac{b_i - y - \theta(n_i - n_{i-1})}{\sqrt{n_i - n_{i-1}}}\right)\right\} dy.$$

- A major reason why a normal random walk is used as a prototypical case is that the multivariate distribution of many group sequential test statistics has a **limiting normal distribution with independent increments**.
  - Jennison & Turnbull (1999), Scharfstein & Tsiatis (1997): all sequentially computed Wald statistics based on efficient estimates of the parameter of interest have the above asymptotic distribution.
  - The signed root likelihood ratio statistic

$$W_i = \text{sign}\left(u(\hat{\theta}_{n_i}) - u_0\right)\sqrt{2n_i\Lambda_i},$$

in which the GLR statistic Λ_i is

$$\begin{aligned}\Lambda_i &= n_i\left\{\hat{\theta}_{n_i}^T\bar{X}_{n_i} - \psi\left(\hat{\theta}_{n_i}\right)\right\} - \sup_{u(\theta)=u_0} n_i\left\{\theta^T\bar{X}_{n_i} - \psi(\theta)\right\} \\ &= \inf_{u(\theta)=u_0} n_i I\left(\hat{\theta}_{n_i}, \theta\right),\end{aligned}$$

is approximately normal with mean 0 and variance n_i under $H_0 : u(\theta) = u_0$, and that the increments $W_i - W_{i-1}$ are approximately independent under H_0.

3. An efficient approach to adaptive designs

Efficient adaptive designs and GLR tests

- **Bartroff & Lai (2008a):**
  - Efficient tests with at most 3 stages
  - Consider a one-parameter exponential family

$$f_\theta(x) = \exp\left(\theta x - \psi(\theta)\right)$$

  - Want to test $H_0 : \theta \le \theta_0$, with no more than M observations
  - Group sizes:
    - Stage 1: $n_1 = m$
    - Stage 2: $n_2 = m \vee \left\{ M \wedge \left\lceil (1+\rho_m) n\left(\hat{\theta}_m\right) \right\rceil \right\}$ with

$$n(\theta) = \min\left\{ |\log \alpha| / I(\theta, \theta_0), |\log \tilde{\alpha}| / I(\theta, \theta_1) \right\}$$

    - Stage 3 (if $n_2 < M$): $n_3 = M$

Rejection and futility boundaries are similar to Lai & Shih (2004).

- Stop at stage $i \leq 2$ and reject H_0 if

$$n_i < M, \quad \widehat{\theta}_{n_i} > \theta_0, \quad \text{and} \quad n_i I(\widehat{\theta}_{n_i}, \theta_0) \geq b.$$

- Stop at stage $i \leq 2$ and accept H_0 if

$$n_i < M, \quad \widehat{\theta}_{n_i} < \theta_1, \quad \text{and} \quad n_i I(\widehat{\theta}_{n_i}, \theta_1) \geq \widetilde{b}.$$

- Reject H_0 at stage $i = 2$ or 3 if

$$n_i = M, \quad \widehat{\theta}_M > \theta_0, \quad \text{and} \quad M I(\widehat{\theta}_M, \theta_0) \geq c,$$

accepting H_0 otherwise.

- The original idea to use

$$n_2 = m \vee \left\{ M \wedge \left\lceil (1+\rho_m) n\left(\hat{\theta}_m\right) \right\rceil \right\}$$

as the second-stage sample size and to allow the possibility of a third stage to account for uncertainty in the estimate $\hat{\theta}_m$ (and hence n_2) is due to Lorden (1983).

- It can be shown that the three-stage test is asymptotically optimal:

If N is the sample size of the three-stage test above, then

$$E_\theta(N) \sim m \vee \left\{ M \wedge \frac{|\log \alpha|}{I(\theta,\theta_0) \vee I(\theta,\theta_1)} \right\}$$

as $\alpha + \widetilde{\alpha} \to 0$, $\log \alpha \sim \log \widetilde{\alpha}$, $\rho_m \to 0$ and $\rho_m \sqrt{m/\log m} \to \infty$; and if T is the sample size of any test of $H_0 : \theta \le \theta_0$ whose error probabilities at θ_0 and θ_1 do not exceed α and $\widetilde{\alpha}$, respectively, then

$$E_\theta(T) \ge (1 + o(1)) E_\theta(N)$$

simultaneously for all θ.

- **Bartroff & Lai (2008b):**
  - Allow the possibility of increasing the maximum sample size from M to $\widetilde{M}$
  - Efficient tests with at most 4 stages
  - Group sizes:
    - Stage 1: $n_1 = m$
    - Stage 2: $n_2 = m \vee \left\{ M \wedge \left\lceil (1+\rho_m) n\left(\hat{\theta}_m\right) \right\rceil \right\}$
    - Stage 3: $n_3 = n_2 \vee \left\{ M' \wedge \left\lceil (1+\rho_m) \widetilde{n}\left(\hat{\theta}_{n_2}\right) \right\rceil \right\}$ with

$$\widetilde{n}(\theta) = \min\left\{ |\log\alpha| / I(\theta,\theta_0), |\log\tilde{\alpha}| / I(\theta,\theta_2) \right\}$$

    - Stage 4 (if $n_3 < \widetilde{M}$): $n_4 = \widetilde{M}$

4. Comparative Studies

Example: Randomized phase II cancer trial

- Thall & Simon (1994):
  - Phase II trial for treatment of AML
  - Control (standard): fludarabine + ara-C
    Experimental: fludarabine + ara-C + G-CSF
  - From prior data, control response rate $p_0 \approx 0.5$
  - Interested in improvement of $p_1 - p_0 = 0.2$
- $\alpha = 0.05$, $\widetilde{\alpha} = 0.2$
- $m = 25$, $M = 78$

- **Thall et al (1988)**
  - $H_0 : p_1 \leq p_0$ vs. $H_1 : p_1 > p_0$
  - Z_i = approx. normally distributed test statistic at the end of stage i ($i = 1, 2$)
  - At stage 1, stop for futility if $Z_1 \leq y_1$; otherwise continue
  - At stage 2, reject H_0 if $Z_2 > y_2$
  - Choose n_1, n_2, y_1, y_2 to minimize

$$AvSS = \frac{1}{2}\Big[E(N \mid p_1 = p_0) + E(N \mid p_1 = p_0 + \delta)\Big]$$

subject to type I and type II error probability constraints.

Expected sample size, power (in parentheses), expected number of stages (in brackets) and average expected sample size (AvSS).

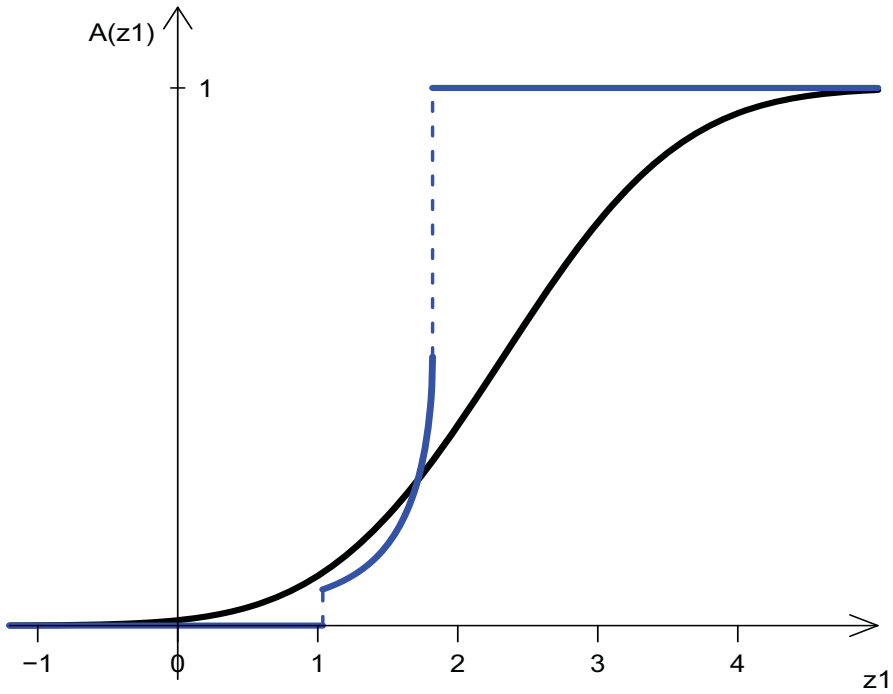
q	p	ADAPT			Opt2		
.4	.3	33.3	(0.4%)	[1.1]	37.8	(0.2%)	[1.1]
	.4	46.1	(5.3%)	[1.5]	48.9	(5.3%)	[1.4]
	.5	57.5	(32.3%)	[1.8]	63.3	(35.6%)	[1.7]
	.6	56.4	(76.0%)	[1.8]	73.5	(78.9%)	[1.9]
	.7	43.8	(97.0%)	[1.5]	77.3	(97.7%)	[2.0]
AvSS		51.3			61.2		
.5	.4	34.7	(0.4%)	[1.2]	38.2	(0.2%)	[1.1]
	.5	47.3	(5.0%)	[1.5]	49.0	(5.6%)	[1.4]
	.6	57.5	(32.2%)	[1.8]	63.3	(35.5%)	[1.7]
	.7	55.1	(77.8%)	[1.8]	73.7	(80.4%)	[1.9]
	.8	41.0	(97.6%)	[1.4]	77.5	(98.2%)	[2.0]
AvSS		51.2			61.4		
.6	.5	34.7	(0.4%)	[1.2]	38.2	(0.2%)	[1.1]
	.6	46.0	(5.2%)	[1.5]	48.9	(5.3%)	[1.4]
	.7	55.8	(33.2%)	[1.7]	63.3	(35.6%)	[1.7]
	.8	52.3	(81.1%)	[1.7]	74.4	(84.2%)	[1.9]
	.9	35.9	(98.5%)	[1.3]	77.8	(99.4%)	[2.0]
AvSS		49.2			61.7		

- **Proschan & Hunsberger (1995):**
  - For testing two normal means
  - **Two-stage design:** uses information about the treatment difference from the first stage to determine the number of additional observations needed and the critical value to use at the end of the study.
  - Conditional power/error:

$$CP_\theta = P_\theta(\text{reject } H_0 \,|\, \text{test statistic at first stage})$$

- Choose a conditional error function $A(\cdot) \in [0, 1]$, such that

$$\int_{-\infty}^{\infty} A(z_1)\, \phi(z_1)\, dz_1 = \alpha.$$



- Choose a conditional error function $A(\cdot) \in [0, 1]$, such that

$$\int_{-\infty}^{\infty} A(z_1)\, \phi(z_1)\, dz_1 = \alpha.$$

  - For a chosen n_2, set $CP_0(n_2, c \mid z_1) = A(z_1)$ to find $c(n_2, z_1)$. This guarantees α-level procedure:

$$\text{Type I error } = \int_{-\infty}^{\infty} CP_0(n_2, c \mid z_1)\, \phi(z_1)\, dz_1.$$

    (Muller & Schafer, 2004)

  - Set $CP_\theta(n_2, c(n_2, z_1) \mid z_1) = 1 - \beta_1$ to find $n_2(z_1)$ to guarantee conditional power of $1 - \beta_1$ to detect θ. May use observed treatment difference for θ.

- **Li et al (2002):**
  - Let $A(z_1)$ has the form

$$A(z_1) = \begin{cases} 0 & z_1 < h \\ CP_0(n_2, c|z_1) & h \le z_1 < k \\ 1 & z_1 \ge k \end{cases}$$

  - The overall type I error probability is

$$\begin{aligned} \alpha &= \alpha_1 + \int_h^k A(z_1)\phi(z_1)dz_1 \\ &= \alpha_1 + \int_h^k \left[1 - \Phi\left(\frac{c\sqrt{n_1+n_2} - z_1\sqrt{n_1}}{\sqrt{n_2}}\right)\right]\phi(z_1)dz_1 \end{aligned}$$

  - For given c, choose $n_2 = n_2(z_1, c)$ to have conditional power $CP_\theta(n_2, c|z_1) = 1 - \beta_1$
  - For given α_1, h, k, choose c such that the above equation holds

Power, expected sample size, and efficiency ratio (in parentheses and at $p_2 > p_1$) of the tests of $H_0 : p_2 \leq p_1$.

p_1	p_2	L	PH	ADAPT
0.20	0.15	0.7%	0.7%	0.3%
		63.4	63.0	98.6
	0.20	5.2%	5.2%	5.0%
		75.8	74.5	158.2
	0.30	53.0%	51.8%	81.8%
		102.0 (89.7)	97.2 (90.8)	206.1 (100)
	0.35	77.1%	76.2%	97.4%
		95.3 (73.3)	90.7 (75.1)	160.5 (100)
0.25	0.20	0.8%	1.0%	0.4%
		64.7	64.5	111.2
	0.25	5.2%	5.1%	5.0%
		77.3	75.8	171.2
	0.35	**48.3%**	**47.0%**	**79.2%**
		97.7 (90.5)	**93.3 (91.9)**	**213.1 (100)**
	0.40	72.7%	71.7%	96.7%
		94.1 (74.1)	89.7 (76.3)	170.3 (100)
0.30	0.25	0.9%	0.9%	0.4%
		65.5	64.7	122.2
	0.30	5.1%	5.0%	5.0%
		75.1	73.7	177.0
	0.40	45.3%	44.3%	76.6%
		96.4 (92.7)	92.0 (95.1)	218.3 (100)
	0.45	70.9%	69.9%	96.2%
		96.1 (75.2)	91.4 (77.6)	176.9 (100)

Conclusions

- GLR statistics are efficient statistics for adaptation
  - Comparable to the benchmark optimal adaptive test of Jennison and Turnbull (2006a,b)
  - The benchmark test needs to assume a specified alternative.
  - Fulfills the seemingly disparate requirements of flexibility and efficiency on a design.
  - Rather than achieving exact optimality at a specified collection of alternatives through dynamic programming, they achieve asymptotic optimality over the entire range of alternatives, resulting in near-optimality in practice.
- Versatility of GLR tests
  - Phase I-II and phase II-III trials
  - Development and validation of biomarker-guided therapies

Conclusions

- Major drawback of conditional power approach to two-stage adaptive designs is that the estimated alternative at the end of the first stage can be quite different from the actual alternative; it may even fall in H_0 and mislead one to stop for futility, resulting in substantial lose of power. The three-stage test makes use of M to come up with an implied alternative and adjust for the uncertainty in the parameter estimates. Moreover, we estimate the second-stage sample size by using an approximation to Hoeffding's lower bound rather than the conditional power.

Conclusions

- This new approach to adaptive design is built on the foundation of sequential testing theory. it can serve to bridge the gap between the "efficiency camp" in the adaptive design estimation with the "flexibility camp" that focuses on addressing the difficulty of comping up with realistic alternative at the design stage. An important innovation is that it uses the Markov property to compute error probabilities when the fixed sample size is replaced by a data-dependent sample size that is based on an estimated alternative at the end of the first stage, like the "flexibility camp".