Adaptive Sample Size Re-estimation: Design and Inference

Sarah Emerson and Gregory Levin

Disclaimer

This presentation reflects the views of the author and should not be construed to represent FDA's views or policies.

Outline

1. Statistical Efficiency of Adaptation
2. Complete Inference after Adaptation
   - Inference for Pre-specified Design
   - Inference after Unplanned Adaptation
3. Evaluating Inferential Methods
4. Additional Issues

1. Statistical Efficiency of Adaptation

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

3. Evaluating Inferential Methods

4. Additional Issues

Competing Issues in Clinical Trials

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

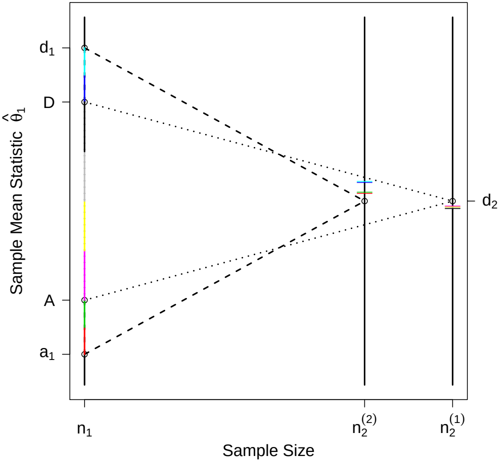
Considering Adaptation

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

Efficiency of Adaptive Testing

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

Violation of Sufficiency Principle



Our Research on Efficiency

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

Setting and Notation

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

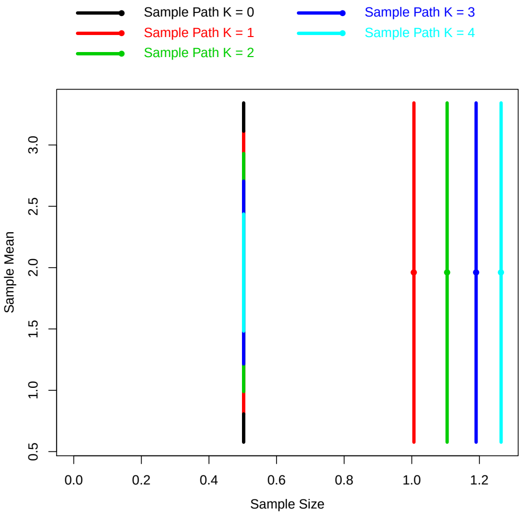
Setting and Notation

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

A Class of Pre-specified Adaptive Designs

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

Example of Adaptive Design



Sampling Density

- $N_j^* = n_j^{k*}$ is fixed conditional on $S_{j-1} = s \in C_{j-1}^k$
- Appealing to the central limit theorem,
  - $S_1^* \sim N(n_1^0\,\theta\,/\,2,\ n_1^0\,\sigma^2)$
  - $S_j^* \mid S_{j-1} \sim N(n_j^{k*}\,\theta\,/\,2,\ n_j^{k*}\,\sigma^2)$

Sampling Density

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

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

where the (sub)density is recursively defined as

$$f_{M,S,K}(1,\ s,\ 0;\ \theta) \quad = \quad \frac{1}{\sqrt{n_1^0}\,\sigma}\,\phi\left(\frac{s - n_1^0\,\theta\,/\,2}{\sqrt{n_1^0}\,\sigma}\right)$$

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

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

Sampling Density

Easy to show the following relation:

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

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

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

Computations

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

Efficiency of Adaptive Testing

Our research on efficiency...

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

Setting 1: Optimality Criteria

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

Setting 1: Optimal GSD

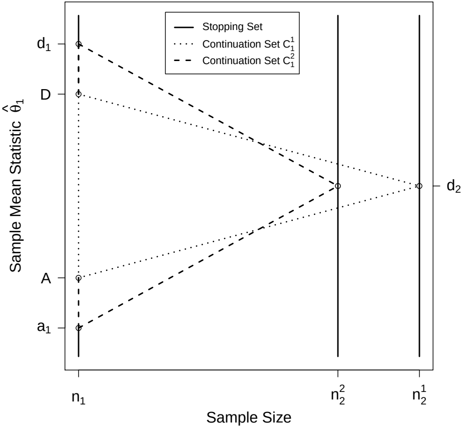
- 2-analysis GSD with Pocock-like stopping boundaries
- Analyses at 50% and 118% of original fixed sample size n
- Stopping boundaries for futility and efficacy at first analysis 0.21Δ and 0.79Δ on sample mean scale
  - $(0.57, 2.21)$ on Z-scale
  - $(4.9\%, 95.1\%)$ on conditional power scale assuming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

Finding the “Optimal” Adaptive Design

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

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

Finding the "Optimal" Adaptive Design



Setting 1: Results

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

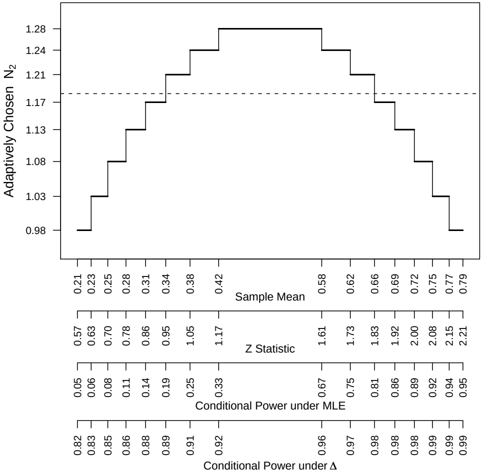
	Number of Continuation Regions								
	0^a	1^b	2	3	4	5	6	7	8
$\text{ASN}_{\theta=0,\Delta}$	1	0.6854	0.6831	0.6828	0.6825	0.6824	0.6824	0.6824	0.6824
% Difference	+45.9%	*Ref*	-0.34%	-0.38%	-0.42%	-0.43%	-0.43%	-0.44%	-0.44%
Maximal N	1	1.18	1.24	1.24	1.26	1.26	1.26	1.26	1.28

a. Fixed Sample Design

b. Group Sequential Design (*Reference* design)

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

Setting 1: The Optimal Adaptive Design



Setting 1: Describing the Design

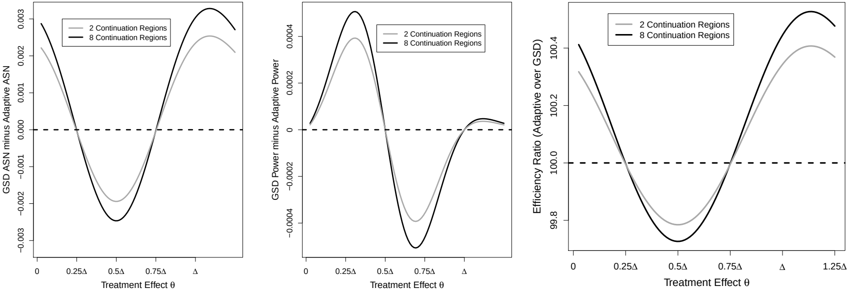
- Increasing # regions only modestly increases maximal N
  - Designs frequently proposed in literature allow $\geq$ 2-fold increase
- Largest maximal sample sizes chosen near center of group sequential continuation region, smallest near boundaries
- $\sim$ Optimal thresholds for increasing N_2 (relative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

Setting 1: Other Efficiency Considerations

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

Setting 1: Other Efficiency Considerations

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



Setting 2: Optimality Criteria

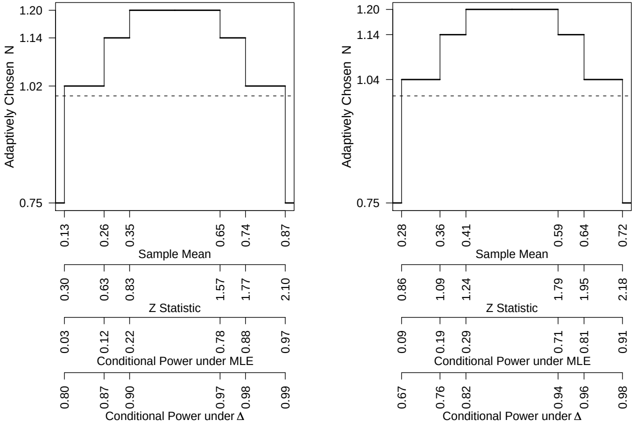
- Only one "stopping analysis"
- Earlier "adaptation" analysis to determine optimal sample size
- $\alpha = 0.025$, $\beta = 0.975$ at $\theta = \Delta$, candidate fixed $n = 4\,\frac{(z_{1-\alpha}+z_\beta)^2}{\Delta^2}$
- Minimum sample size for stopping of $n_{min} < n$ required for adequate safety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

Setting 2: Results

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

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

Setting 2: The Optimal Adaptive Designs



(a) q = 0.5

(b) q = 0.8

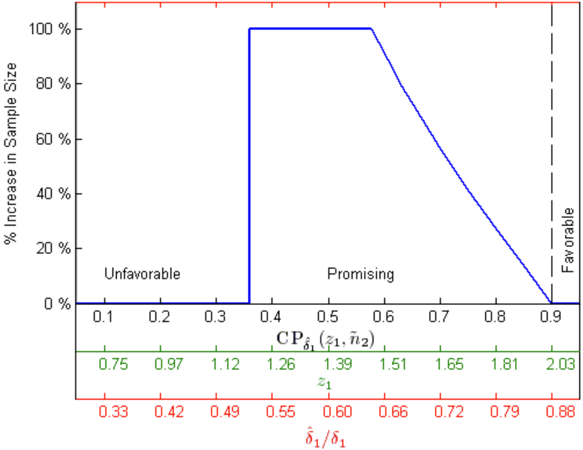
Setting 2: Describing the Designs

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

Results on Efficiency

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

Design in Literature (MP 2010)



Limitations

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

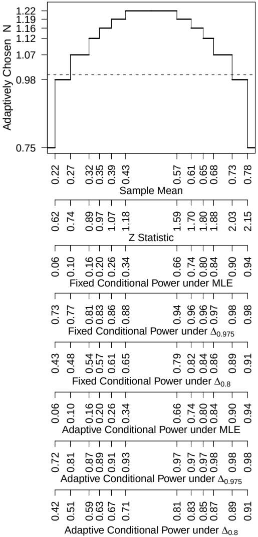
Other Results: Jennison and Turnbull 2006

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

Efficiency in Survival Setting

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

Stochastic Curtailment and Conditional Power



Stochastic Curtailment and Conditional Power

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

1-1 Correspondence Between Scales

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

Collaborate, Evaluate, and Iterate

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

Schizophrenia Example (Mehta and Pocock 2010)

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

Schizophrenia Example

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

Schizophrenia Software Example

Demonstration of calculations of important operating characteristics in R

1. Statistical Efficiency of Adaptation

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

3. Evaluating Inferential Methods

4. Additional Issues

Motivation for Additional Research

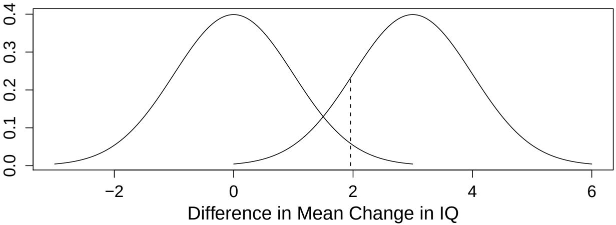
- Despite unclear efficiency gains, adaptive designs implemented in practice, so research needed to propose, evaluate estimation methods
  - Desire for "innovative" designs
  - One sponsor even requires justification if adaptation not included?
- False positive rate and statistical efficiency not only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

Case Study: NECAT

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

Fixed Sample Hypothesis Testing

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



Testing versus Estimation

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

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

Testing versus Estimation

- Four scenarios: What do you conclude?

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

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

The Need for Good Estimates

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

Complete Inference

Four numbers (with good properties):

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

Sequential Analyses: Statistical Challenges

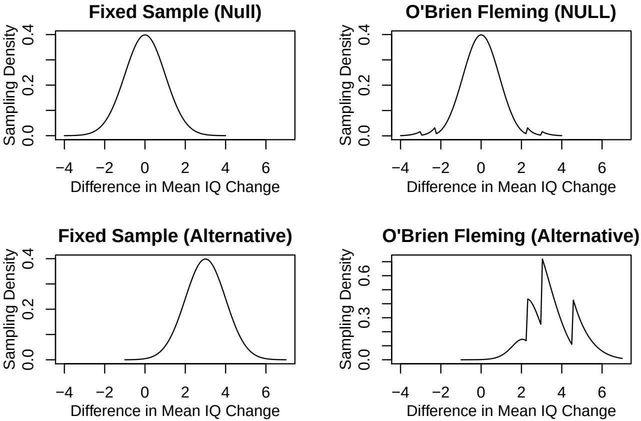
- Sequential testing has implications on estimation of the treatment effect in addition to hypothesis testing
- We stop early only if extreme results are observed
  - Fixed sample estimates such as the sample mean tend to be biased (to the extreme)
  - Confidence intervals do not have correct coverage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

Connection to Other Types of Studies

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

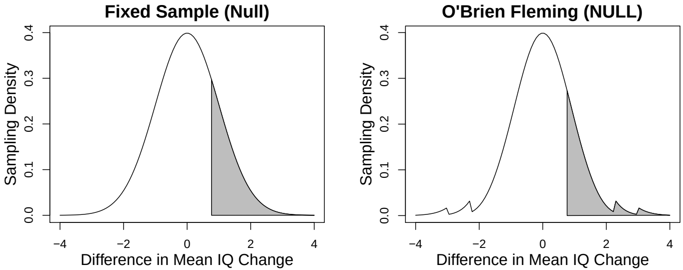
Estimation after Sequential Hypothesis Testing

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



Estimation after Sequential Testing

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



Well-understood Methods

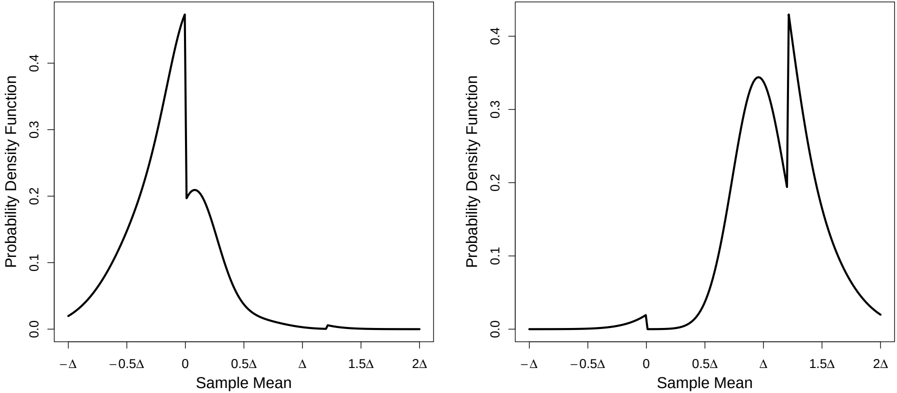
- Extensive literature on estimation after group sequential test
  - Different methods to compute bias-adjusted point estimates, and correct (adjusted) confidence intervals and P-values
  - Extensive evaluation of properties assessing the reliability and precision of estimates, CIs, P-values
  - Variety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

Extension of Group Sequential Approaches

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

Adaptive Sampling Density of Sample Mean

- Under null (left) and alternative (right)



Duality of Testing and Confidence Sets

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

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

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

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

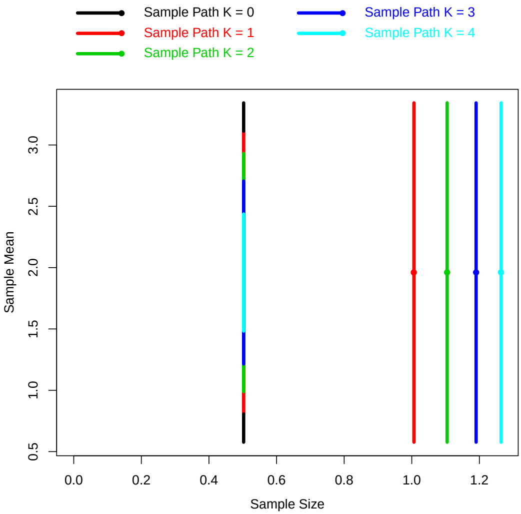
Exact Confidence Sets

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

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

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

Orderings of Outcome Space



Orderings of Outcome Space

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

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

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

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

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

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

Orderings of Outcome Space

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

Stage-wise Orderings

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

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

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

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

- *Statistical information ordering (N)*:

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

Point Estimates and P-values

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

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

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

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

Statistics as Usual

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

Schizophrenia Software Example

Demonstration of calculations in R to compute estimates, confidence intervals, and P-values after fixed, group sequential, and adaptive sampling plans

Interactive Exercise

Interactive exercise to illustrate concepts discussed thus far

1. Statistical Efficiency of Adaptation
2. Complete Inference after Adaptation
   - Inference for Pre-specified Design
   - Inference after Unplanned Adaptation
3. Evaluating Inferential Methods
4. Additional Issues

Unplanned Adaptation: Motivation

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

Testing versus Estimation

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

Brannath, Mehta, and Posch (BMP) 2009

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

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

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

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

Gao, Liu, and Mehta (2012)

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

1. Statistical Efficiency of Adaptation
2. Complete Inference after Adaptation
   - Inference for Pre-specified Design
   - Inference after Unplanned Adaptation
3. Evaluating Inferential Methods
4. Additional Issues

Assessing Reliability and Precision of Inference

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

Assessing Reliability and Precision of Inference

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

Approach to Evaluating Inferential Methods

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

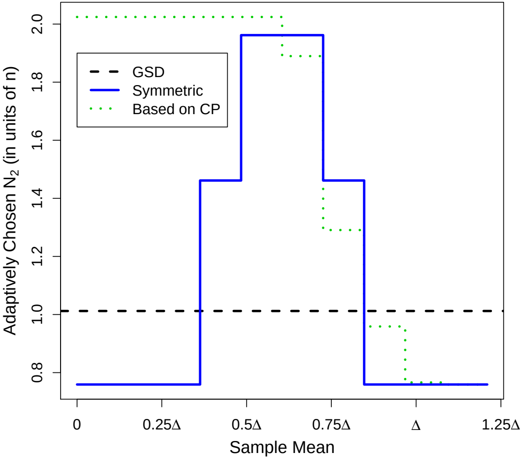
Comparison Framework

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

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

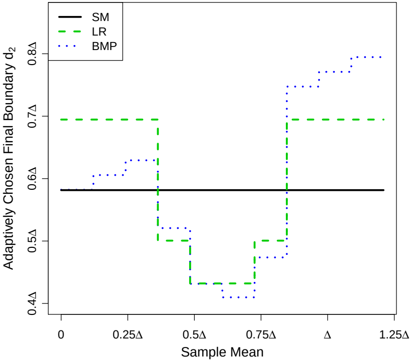
Adaptively Chosen Sample Size

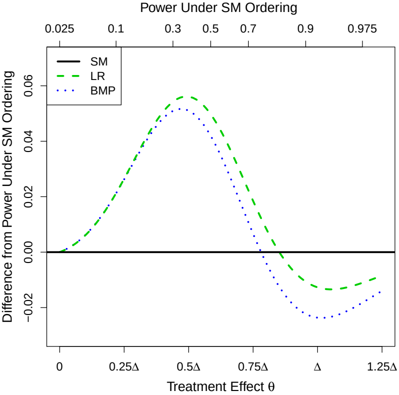
- Example of symmetric and CP-based N_2 functions



Differences in Boundaries and Power

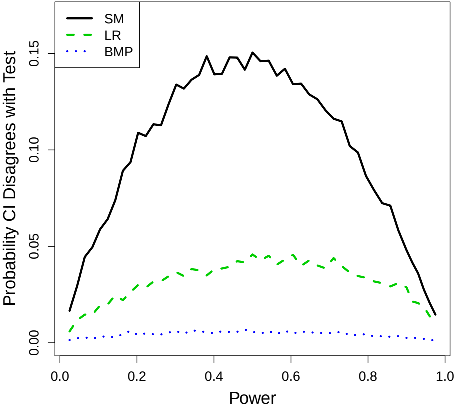
- Comparing testing based on different orderings of outcome space (OF reference, symmetric N_J rule, 100% maximal increase)...





Avoiding Inconsistent Inference

- Should use same ordering for testing as for estimation



Confidence Intervals: Correct Coverage

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

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

Estimates: Median-unbiased

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

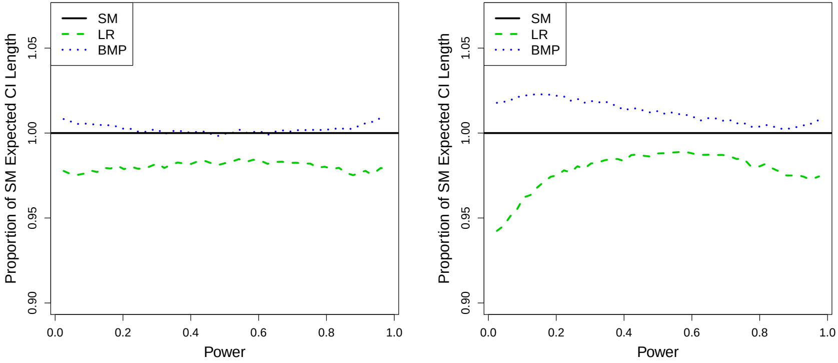
	OF Reference GSD			Pocock Reference GSD		
Power	SM	LR	BMP	SM	LR	BMP
			Symmetric N_J function, up to 100% Increase			
0.0250	0.4956	0.4993	0.4960	0.4983	0.4986	0.4960
0.5000	0.5082	0.5076	0.5081	0.5100	0.5093	0.5095
0.9000	0.5019	0.5006	0.4970	0.5034	0.5028	0.5011
			CP-based N_J function, up to 100%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

Results: Naive Inference

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

Comparing Confidence Intervals: Example

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

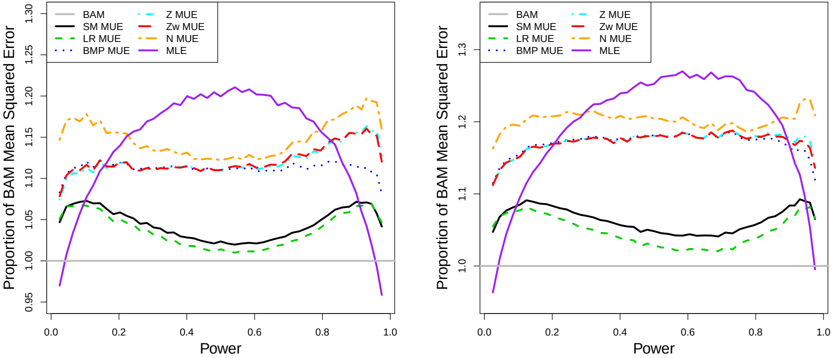


Comparing Confidence Intervals: Trends

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

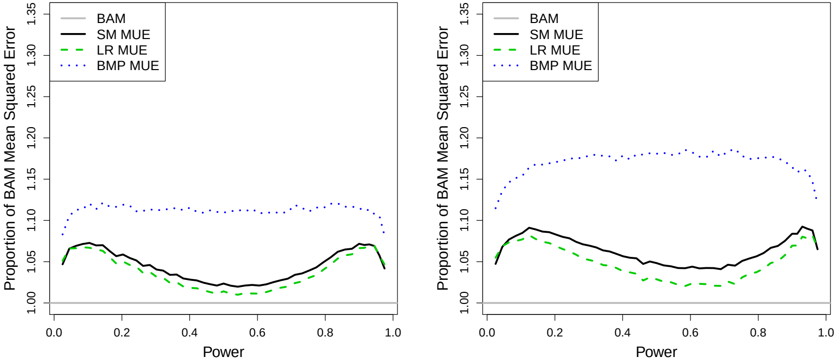
Comparing Point Estimates: Example

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



Comparing Point Estimates: Example

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

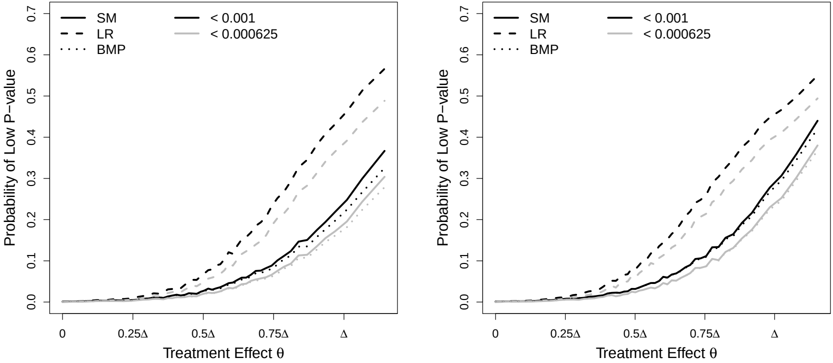


Comparing Point Estimates: Trends

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

Comparing *P*-values: Example

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



Comparing *P*-values: Trends

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

Summary and Conclusions

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

Cost of Planning not to Plan

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

1. Statistical Efficiency of Adaptation

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

3. Evaluating Inferential Methods

4. Additional Issues

Case Study: An Antidepressant in MDD

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

Case Study: An Antidepressant in MDD

Candidate designs:

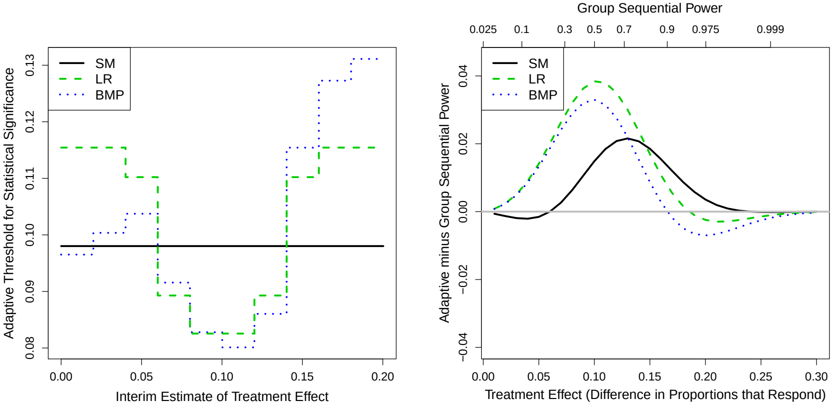
- Fixed sample design with 176 participants per arm
  - $\alpha = 2.5\%$ type I error, $\beta = 90\%$ power at $\theta = 0.165$, threshold for statistical significance of 10%
- Two-analysis O'Brien and Fleming and Pocock group sequential designs with same α, β, significance threshold
- Adaptive designs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

Statistical *versus* Clinical Significance

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

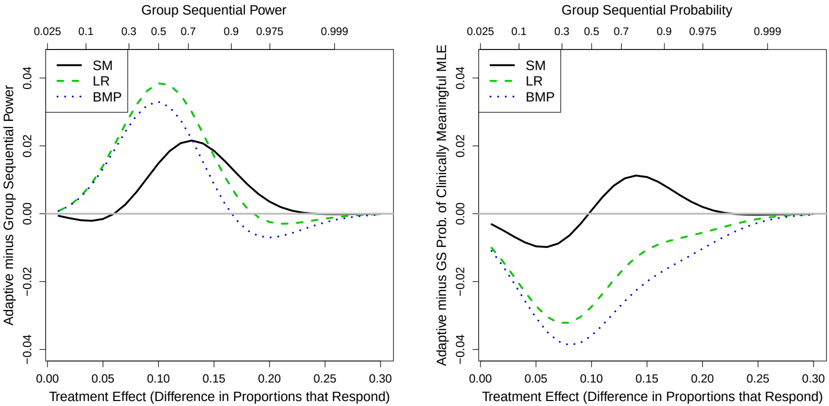
Statistical *versus* Clinical Significance

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



Statistical *versus* Clinical Significance

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



Maintaining Confidentiality

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

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

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

Maintaining Confidentiality

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

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

Logistical and Ethical Issues

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

Additional Challenges: Summary

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

Summary and Conclusions

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

Editorial

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

Thank you