Adaptive re-design of clinical trials

Chris Jennison,

Department of Mathematical Sciences, University of Bath, UK

and

Bruce Turnbull,

Department of Statistical Science, Cornell University, Ithaca, NY

International Biometrical Society

German Region

Magdeburg

16–17 October 2003

http://www.bath.ac.uk/∼mascj

A variety of adaptive and flexible procedures

- Adaptive randomisation rules designed to allocate fewer subjects to the inferior treatment.
- Adapting the sample size to estimates of nuisance parameters.
- Re-assessing the power requirement in response to interim data or external information.
- Flexibility to change treatment, outcome or response during a study.

Plan of talk

1. Motivation for adaptive sample size designs.

2. Methods for adaptive re-design.

3. *Examples:* Group sequential tests adapting to

   (1) internal information,

   (2) external factors.

   Overall efficiency of these procedures.

4. Pre-designed group sequential tests with adaptive group sizes.

§1 Motivation: Prototype example

Balanced parallel design

$$X_{Ai} \sim N(\mu_A, \sigma^2), \quad X_{Bi} \sim N(\mu_B, \sigma^2)$$
$$Y_i = X_{Ai} - X_{Bi} \sim N(\theta, 2\sigma^2)$$
$$\theta = \mu_A - \mu_B$$

The MLE of θ is $\widehat{\theta} = \overline{X}_A - \overline{X}_B$.

Without loss of generality, suppose $2\sigma^2 = 1$.

Aim: to Test H_0: $\theta = 0$ versus H_1: $\theta > 0$

with type I error rate α, e.g. $\alpha = 0.025$.

Fixed sample design

Initially aim for power $1 - \beta$ at target effect size $\theta = \delta$.

Hence set sample size

$$n \;=\; (z_\alpha + z_\beta)^2 \, \frac{2\sigma^2}{\delta^2} \;=\; \left(\frac{z_\alpha + z_\beta}{\delta} \right)^2$$

per treatment arm, where $z_\alpha = \Phi^{-1}(1 - \alpha)$, etc.

(Recall $2\sigma^2 = 1$.)

Data at an intermediate stage

After a fraction r of the sample size (information) is collected,

$$\begin{aligned} \widehat{\theta}_1 &\sim N(\theta, \tfrac{1}{rn}), \\ S_1 &\sim N(\theta rn, rn). \end{aligned}$$

Intermediate results may be examined, even though a formal interim analysis was not planned.

Disappointing results

- Suppose $\widehat{\theta}_1$ is positive but smaller than the hoped for effect size δ.
- It is unlikely that H_0 will be rejected (low conditional power).
- However, the magnitude of $\widehat{\theta}_1$ is clinically meaningful.
- It appears the original target effect size δ was over-optimistic.

Can this trial be “rescued” ?

External changes

- Suppose external information about a competing treatment or changes in the manufacturer's circumstances imply it would be worthwhile to find a smaller treatment effect than δ.
- Alternately, the same change in objective may be motivated by, say, safety information internal to the current study.
- Interim data have been seen, so the investigators do know the current estimate $\widehat{\theta}_1$.

Can the trial be enlarged without loss of credibility?

Revising the sample size

- At an interim stage, we wish we had designed the test with power $1-\beta$ at $\theta = \delta/\xi \quad (\xi > 1)$ rather than at $\theta = \delta$.

  E.g., $\delta/\xi = \widehat{\theta}_1$ where this is > 0 and $< \delta$.

- This would have required the larger sample size $\xi^2 n$ instead of n.

- One might collect extra observations in the remainder of the study to make a total sample size of $\xi^2 n$.

Naive test leads to inflated type I error

Suppose we behave as if the sample size $\xi^2 n$ was pre-planned and compute

$$Z = \left(\overline{X}_A - \overline{X}_B\right) \sqrt{\xi^2 n}.$$

If ξ is a function of the first stage data, Z is *not* $N(0,1)$.

The test that rejects when $Z > z_\alpha$ does not have type I error α.

Type I error rate is inflated

- typically by 30% to 40% (Cui, Hung & Wang, *Biometrics,* 1999)
- can more than double (Proschan, Follmann & Waclawiw, *Bmcs,* 1992).

§2 Methods for adaptive re-design

1. Bauer & Köhne (Biometrics, 1994)

Design the study in two stages.

Calculate two separate P-values for H_0 from the two stages, p_1 and p_2.

Use R. A. Fisher's test based on

$$-\ln(p_1 p_2) \sim 0.5\,\chi^2_4.$$

Note the second stage can be re-designed in light of first stage results as long as, conditionally, $p_2 \sim U(0,\,1)$ under H_0.

But: this way of combining the two stages has to be pre-specified.

Adaptive re-design

2. L. Fisher: Variance spending (Stats in Medicine, 1998)

A fixed sample of n observations can be divided into

stage 1: $$S_1 = \sum_{i=1}^{rn}(X_{Ai} - X_{Bi}) \sim N(rn\theta,\, rn),$$

stage 2: $$S_2 = \sum_{i=rn+1}^{n}(X_{Ai} - X_{Bi}) \sim N(\{1-r\}n\theta,\, \{1-r\}n).$$

Under H_0: $\theta = 0$,

$$Z = \frac{S_1 + S_2}{\sqrt{n}} = \left(\frac{S_1}{\sqrt{n}}\right) + \left(\frac{S_2}{\sqrt{n}}\right) \sim N(0,1).$$

Variance spending — continued

At stage 1, we observe $S_1/\sqrt{n}$ from its $N(rn\theta,\, rn)$ distribution.

If we then modify the stage 2 sample size to $\gamma(1-r)n,$ conditionally

$$S_2' \sim N(\gamma\{1-r\}n\theta,\, \gamma\{1-r\}n).$$

Under H_0: $\theta = 0$,

$$\gamma^{-1/2}\, S_2'/\sqrt{n} \sim N(0,\, \{1-r\}),$$

just like the originally planned $S_2/\sqrt{n}$. Hence,

$$Z = \frac{S_1 + \gamma^{-1/2}\, S_2'}{\sqrt{n}} \sim N(0,1).$$

Note this method can be used for *unplanned* adaptation.

Adaptive re-design

3. Cui, Hung & Wang, *Biometrics*, 1999

Consider a group sequential design planned for the sequence of information levels $\{\mathcal{I}_1, \ldots, \mathcal{I}_K\}$.

Score statistic increments are independent with

$$S_1 \sim N(\theta\mathcal{I}_1, \mathcal{I}_1),$$

$$S_k - S_{k-1} \sim N(\theta(\mathcal{I}_k - \mathcal{I}_{k-1}), \mathcal{I}_k - \mathcal{I}_{k-1}).$$

Suppose re-design takes place at analysis j and future increments in information are increased by a factor γ.

Denote new score statistics by $S'_{j+1}, S'_{j+2}, \ldots, S'_K$.

Cui et al. — continued

Then

$$S'_k - S'_{k-1} \sim N(\,\theta\,\gamma\,(\mathcal{I}_k - \mathcal{I}_{k-1}),\ \gamma\,(\mathcal{I}_k - \mathcal{I}_{k-1}))$$

independently of other increments (taking $S'_j = S_j$).

Defining

$$S_k = S_j + \sum_{i=j+1}^{k} \gamma^{-1/2}(S'_i - S'_{i-1}), \quad k = j+1, \ldots, K,$$

recovers the original joint distribution, under H_0, of $S_1, \ldots, S_K$.

Applying the original boundary to these statistics maintains the type I error probability.

Adaptive re-design

4. Conditional type I error probability

In our 2-stage example, conditional type I error probability after stage 1 is

$$P_{\theta=0}\{S_1 + S_2 > z_\alpha \sqrt{n} \mid S_1 = s_1\}. \qquad (1)$$

If stage 2 sample size is modified and a test defined that preserves the conditional error probability (1), overall type I error rate α is maintained.

- The methods of L. Fisher and Cui et al. do this.
- Jennison & Turnbull (2003, SiM) show that any unplanned design modification *must* have this property.
- Müller & Schäfer (2001, Bmcs) and Denne (2001, SiM) use this construction in adaptive group sequential designs.

Variance spending — notes

- If on re-design future sample sizes are multiplied by $\gamma > 1$, later observations are down-weighted. The final statistic Z is not sufficient for θ — so the efficiency of this approach is suspect.
- The distribution of Z under $\theta \neq 0$ is not simple. The inter-relation of stages 1 and 2 needs to be properly treated in calculating overall properties of adaptive procedures.

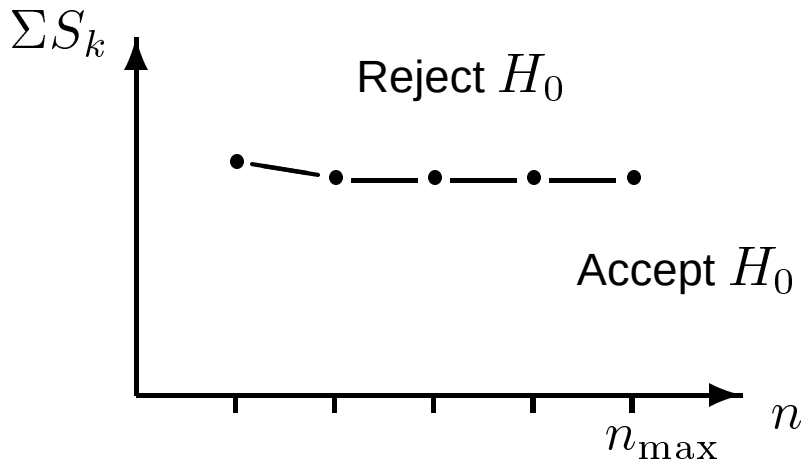
We shall report results on power and average sample size for examples with specific rules for sample size adaptation.

§3 Example 1: A Cui, Hung & Wang (1999) design

Original group sequential design:

To test H_0: $\theta = 0$ with type I error rate 0.025 and power 0.9 at $\theta = \delta$.

Observations taken in 5 groups; early stopping allowed to *reject* H_0.



$n_{\max} = 10.8/\delta^2$, cf fixed sample size, $n_f = 10.5/\delta^2$.

Design modification

Cui et al. suggest adjusting the design at just one interim analysis.

Changing design at stage 3:

Group 4

Original plan: $S_4 =$ sum of $n_{\max}/5$ terms $(X_{Ai} - X_{Bi})$

Revised plan: $S_4' =$ sum of $\gamma\,(n_{\max}/5)$ terms $(X_{Ai} - X_{Bi})$

Use $\gamma^{-1/2}\,S_4'$ in place of S_4, preserving the null distribution.

Group 5 — similarly.

Re-design in response to $\widehat{\theta}_3$ (internal information)

Aim for the sample size needed in the original test to attain power 0.9 at $\theta = \widehat{\theta}_3$ with a minimum value of $\theta = \delta/2$.

So, set $\xi(\widehat{\theta}_3) = \min(\delta/\widehat{\theta}_3,\, 2)$.

To achieve total sample size $\xi(\widehat{\theta}_3)^2 n_{\max},$ with a correction for weighting by $\gamma^{-1/2},$ take

$$\gamma(\widehat{\theta}_3) \;=\; \frac{\{\xi(\widehat{\theta}_3) - 0.6\}^2}{(1-0.6)^2}.$$

Hence $\gamma \in (0, 12.25)$ and total sample size $\in (0.6n_f,\, 5.6n_f)$.

Figure 1. Power functions of original group sequential test and Cui et al. adaptive test.

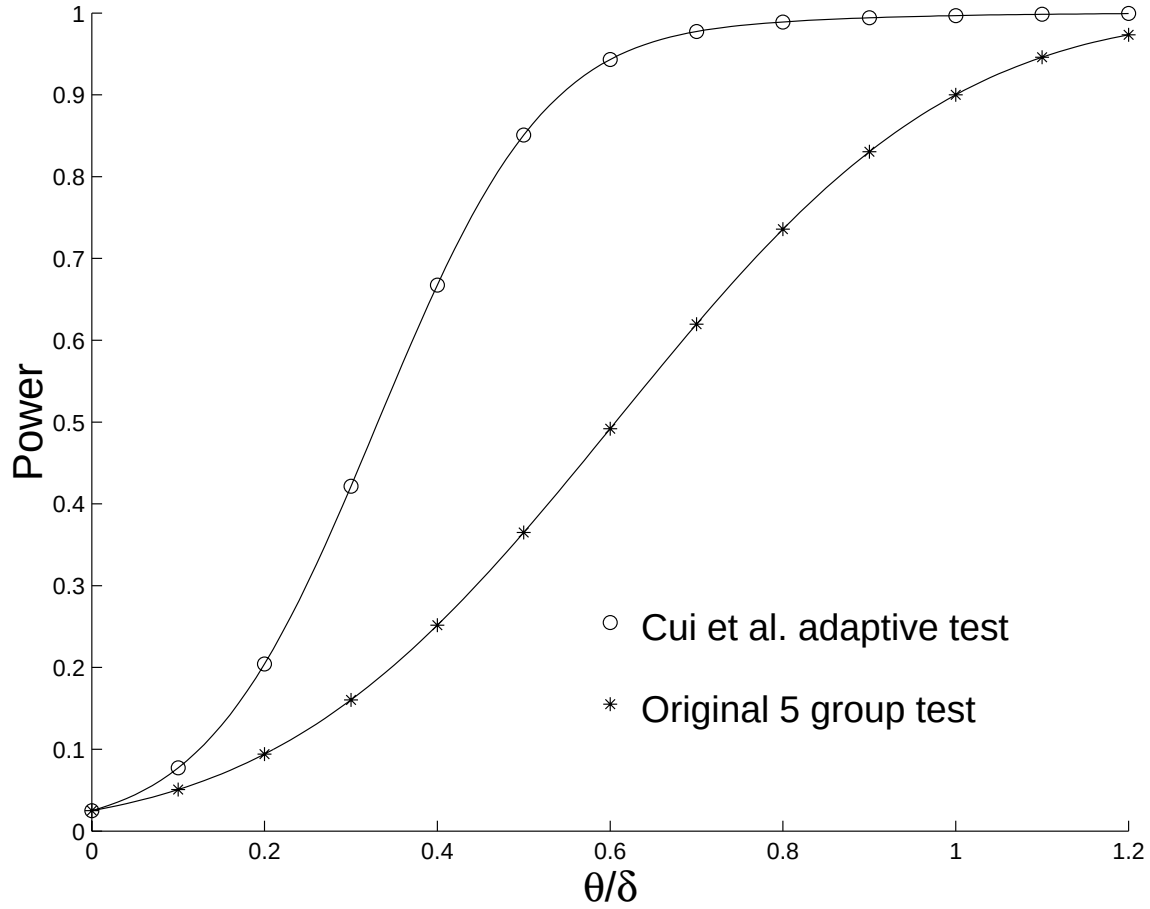
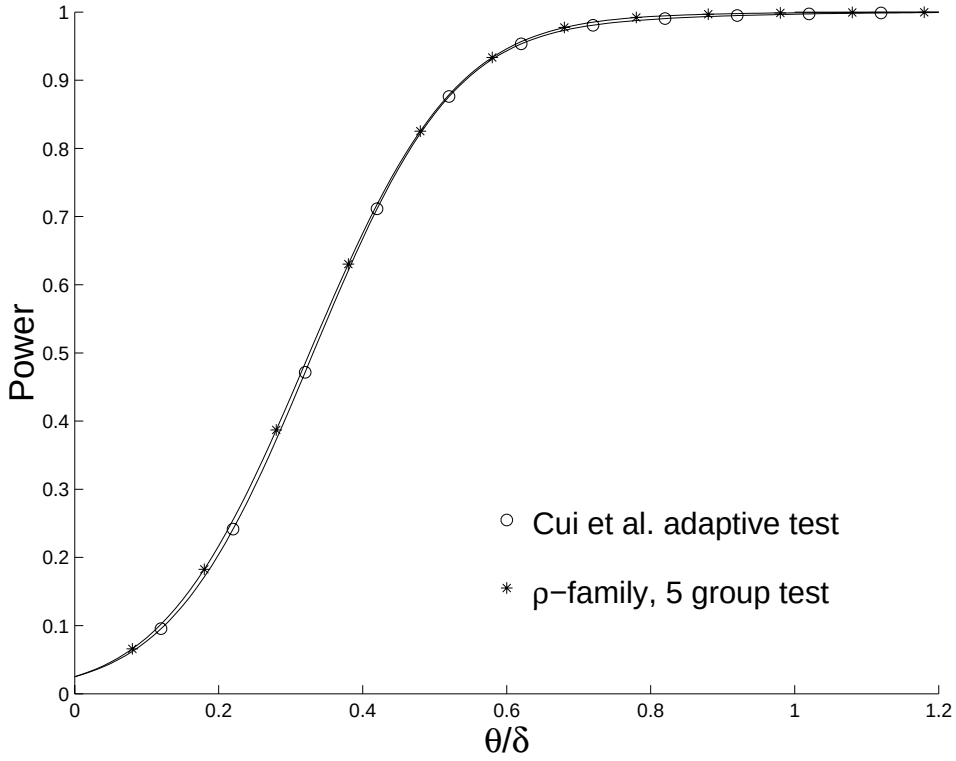
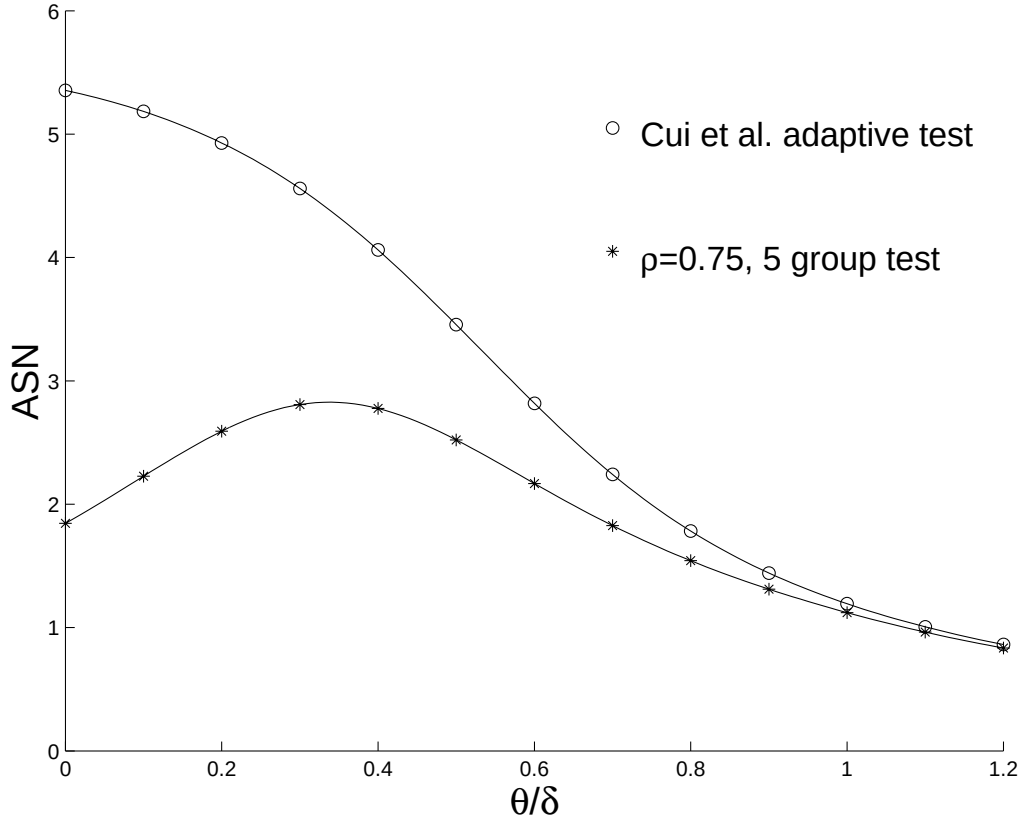


Figure 2. Power functions of Cui et al. adaptive test and a non-adaptive 5 group test with power 0.9 at $\theta = 0.54\,\delta$.



The non-adaptive test is a ρ-family error spending test with $\rho = 0.75$ and interim analyses at 0.1, 0.2, 0.45 and 0.7 of the maximum sample size.

Figure 3. Average Sample Number (ASN) curves of Cui et al. adaptive test and matched non-adaptive test.

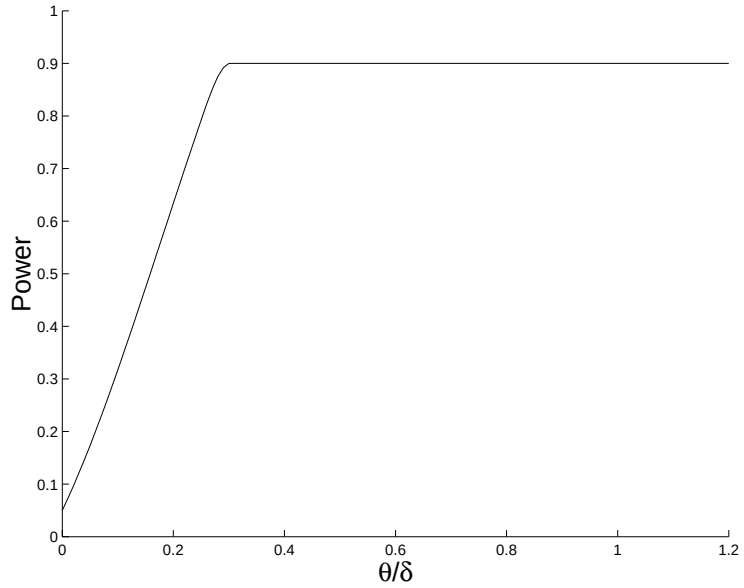


ASN scale is in multiples of the original fixed sample size, n_f.

Setting power: a strange philosophy

Shen and Fisher (1999, Biometrics) refer to setting power $1-\beta$ at effect size δ where δ is an *estimate* of θ. In Example 1 we tried to attain power $1-\beta$ at $\theta = \widehat{\theta}_3$.

This suggests we are aiming for a power function of the following form (!)



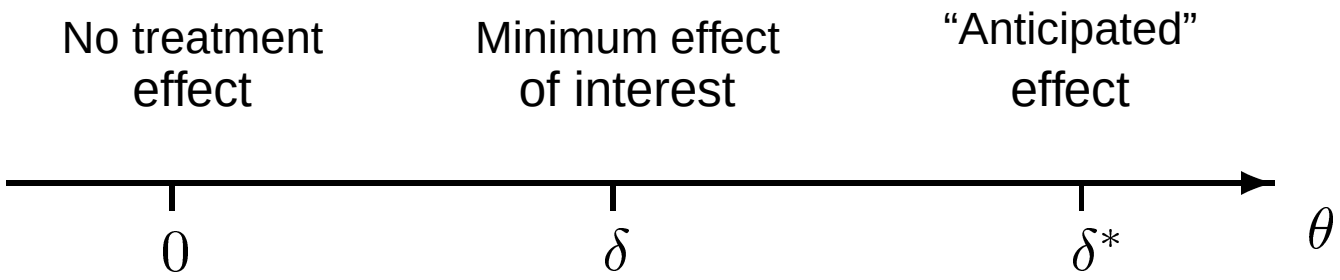
A preferable formulation

Test H_0: $\theta = 0$ with:

type I error rate α,

power $1 - \beta$ at $\theta = \delta$,

low ASN at $\theta = \delta^* \gg \delta$.



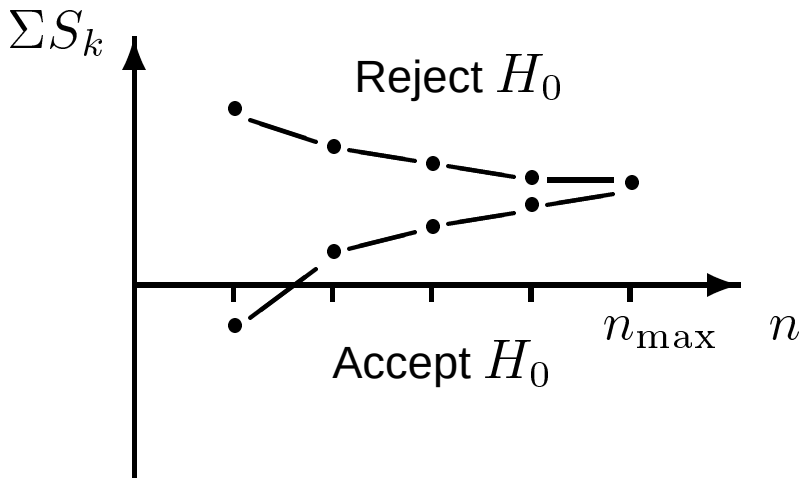
It should not be necessary to see $\widehat{\theta} = \delta$ before realising a treatment effect of this size is (just) worth pursuing.

Example 2: Müller-Schäfer adaptation to external information

Original error spending design:

To test H_0: $\theta = 0$ with type I error rate 0.025 and power 0.9 at $\theta = \delta$.

5 group error spending test, $\rho = 3$, early stopping to accept or reject H_0.



$n_{\max} = 11.0/\delta^2$, cf fixed sample size, $n_f = 10.5/\delta^2$.

Design modification (external information)

At analysis 2, suppose external factors prompt interest in lower θ values and we now aim for power 0.9 at $\delta/2$ rather than δ.

On observing $S_2 = s_2$ in the continuation region:

Calculate conditional type I error rate

$$\tilde{\alpha}(s_2) = P_{\theta=0}\left\{\text{Reject } H_0 \,|\, S_2 = s_2\right\}.$$

Set up a new design based on future observations with

3 further analyses, type I error $\tilde{\alpha}(s_2)$, power 0.9 at $\delta/2$.

Design modification

New design, conditional on $S_2 = s_2$:

Use an error spending test with $\rho = 3$.

Required future group sizes depend on s_2 through $\tilde{\alpha}(s_2)$.

For values of s_2 in the continuation region, the total sample size (including groups 1 and 2) varies up to a maximum of $7.5n_f$.

Figure 4. Power functions of original $\rho = 3$ error spending test and Müller-Schäfer adaptive test.

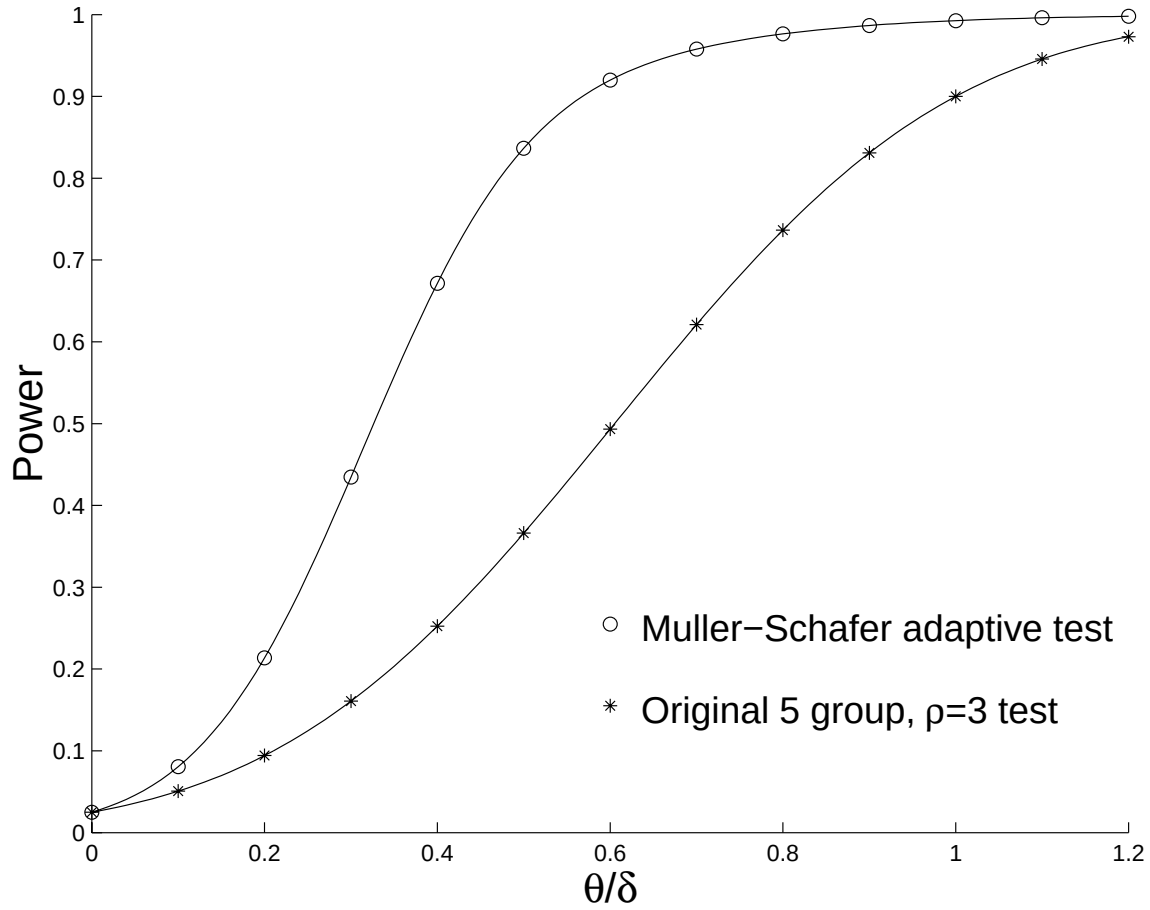
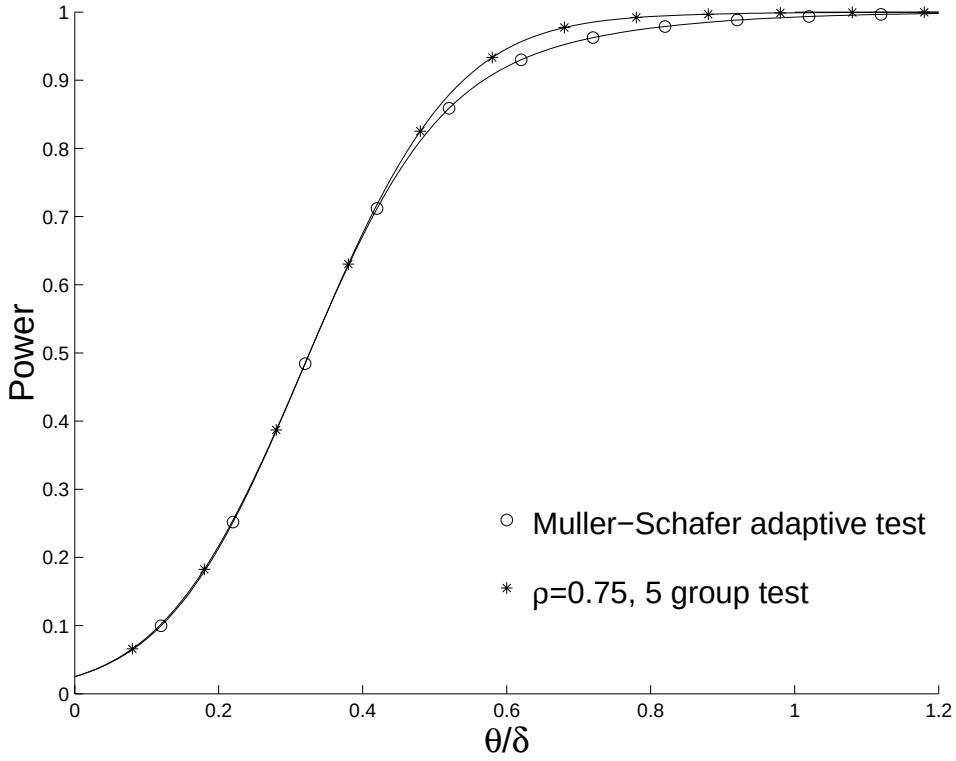
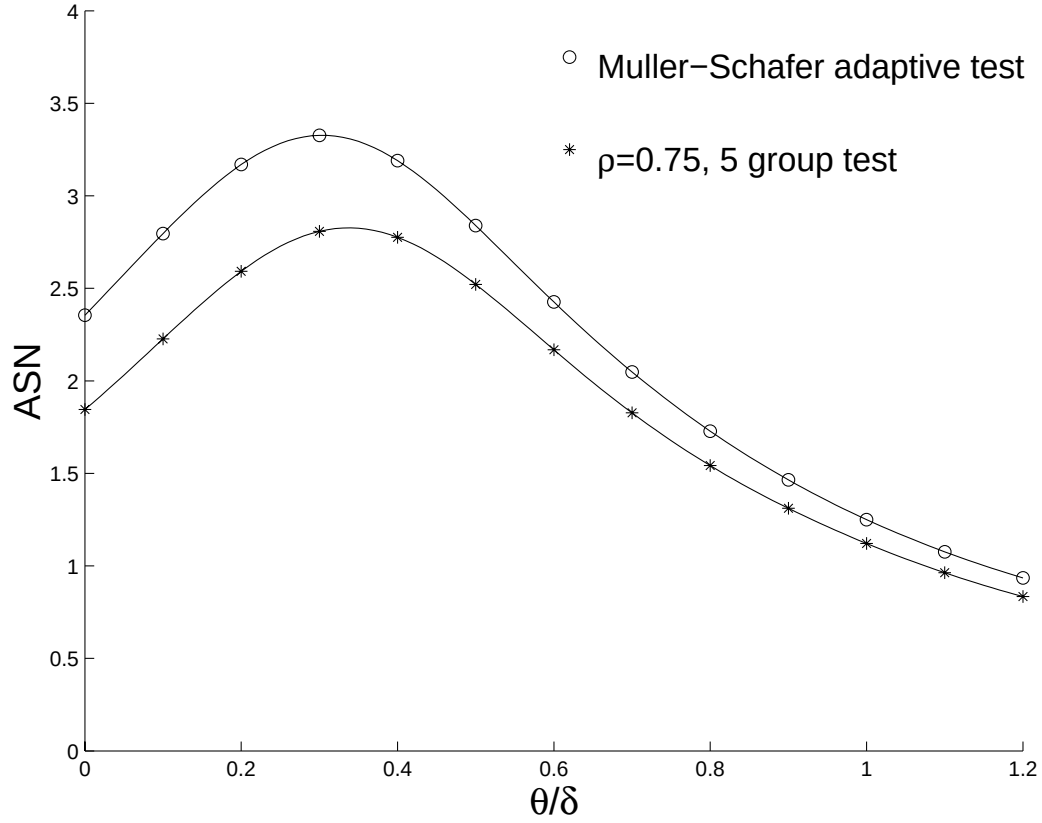


Figure 5. Power functions of Müller-Schäfer adaptive test and a non-adaptive 5 group test with power 0.9 at $\theta = 0.54\,\delta$.



The non-adaptive test is a ρ-family error spending test with $\rho = 0.75$ and interim analyses at 0.1, 0.2, 0.45 and 0.7 of the maximum sample size.

Figure 6. Average Sample Number (ASN) curves of Müller-Schäfer adaptive test and matched non-adaptive test.



ASN scale is in multiples of the original fixed sample size, n_f.

§4 Planned adaptive tests

Use of adaptive methods is not confined to “rescuing” studies.

Adaptive designs can be considered in their own right.

In *Optimal Sequentially Planned Decision Procedures* (Springer-Verlag, 1993), Schmitz proposes tests where the size of group k is allowed to depend on data seen at analysis $k-1$.

But what are the advantages over standard group sequential tests where group sizes are pre-specified (or vary in a way that does not depend on $\widehat{\theta}$) ?

Adaptive designs, Schmitz (1993)

Sequentially planned sequential tests

Initially, fix $\mathcal{I}_1$,

observe $S_1 \sim N(\theta\mathcal{I}_1, \mathcal{I}_1)$,

choose $\mathcal{I}_2$ as a function of S_1, observe S_2 where

$$S_2 - S_1 \sim N(\theta(\mathcal{I}_2 - \mathcal{I}_1), (\mathcal{I}_2 - \mathcal{I}_1)),$$

and so forth.

Specify sampling rule and stopping rule to achieve desired *overall* type I error and power.

Theoretical considerations

If a study has many, frequent analyses there is very little to be gained from adaptive sampling.

With a fixed, small number of analyses, adaptive choice of group size does extend the range of possible designs — offering potential gains in efficiency.

In any pre-specified design, it is efficient to define sampling and stopping rules in terms of the sufficient statistic for θ.

Computing optimal adaptive and non-adaptive designs

Eales & Jennison (*Biometrika*, 1992) and Barber & Jennison, (*Biometrika*, 2002) derive optimal, non-adaptive group sequential tests.

They use Dynamic Programming to solve Bayes sequential decision problems, the solutions of which are optimal frequentist tests.

This approach extends, with rather more computation, to yield optimal adaptive group sequential tests.

Example

To test H_0: $\theta = 0$ versus H_1: $\theta > 0$

with type I error rate $\alpha = 0.025$

and power $1 - \beta = 0.8$ at $\theta = \delta$.

Aim for low values of:

$$\frac{1}{3}\{E_{\theta=0}(N) + E_{\theta=\delta}(N) + E_{\theta=2\delta}(N)\}.$$

Constraints:

Maximum sample size $= 1.2 \times$ fixed sample size.

Maximum number of analyses $= K$.

Optimal average E(N)

Results are stated as a percentage of the fixed sample size.

Number of analyses, K	*Non-adaptive, equally spaced analyses*
2	70.7
3	59.8
4	55.8
6	52.6
8	51.1
10	50.3

Optimal average E(N)

Results are stated as a percentage of the fixed sample size.

Number of analyses, K	*Non-adaptive, equally spaced analyses*	*Optimal adaptive group sequential design*
2	70.7	66.1
3	59.8	57.8
4	55.8	54.0
6	52.6	50.8
8	51.1	49.4
10	50.3	48.6

Optimal average E(N)

Results are stated as a percentage of the fixed sample size.

Number of analyses, K	*Non-adaptive, equally spaced analyses*	*Non-adaptive, optimised group sizes*	*Optimal adaptive group sequential design*
2	70.7	66.4	66.1
3	59.8	58.5	57.8
4	55.8	55.1	54.0
6	52.6	52.1	50.8
8	51.1	50.7	49.4
10	50.3	49.8	48.6

Conclusions

- One can rescue a study found to lack power at an interim stage.

  But, this has a price and investigators really should consider power requirements properly *before* a study gets under way.

- Adaptive methods can help when objectives change in response to *external* factors. The resulting designs lose some efficiency — but this is inevitable when circumstances change without warning.

- Pre-planned adaptive designs have some benefits, but perhaps not enough to compensate for their complexity.