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 Conference on
Statistics in Health Sciences
Nantes, June 2004

http://www.bath.ac.uk/∼mascj

Plan of talk

1. Motivation for adaptive sample size designs that increase power.

2. Methods for adaptive re-design.

3. *Examples:* Group sequential tests adapting to

   (1) external factors,

   (2) internal information.

   Overall efficiency of these procedures.

4. Theoretical results.

5. Pre-planned 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. 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

3. Conditional type I error probability

If, in our 2-stage example, the second stage sample size is modified and a test defined that preserves the conditional type I error probability

$$P_{\theta=0}\{S_1 + S_2 > z_\alpha \sqrt{n} \mid S_1 = s_1\},$$

then the overall type I error rate α is maintained.

- Cui et al's design and Shen & Fisher's (1999, Biometrics) “variance spending” method do this. Indeed, Jennison & Turnbull (2003, SiM) show any unplanned design modification *must* have this property.
- Müller & Schäfer (2001, Bmcs) and Denne (2001, SiM) use this construction to create adaptive group sequential designs.

Comments on flexible adaptive designs

- 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.

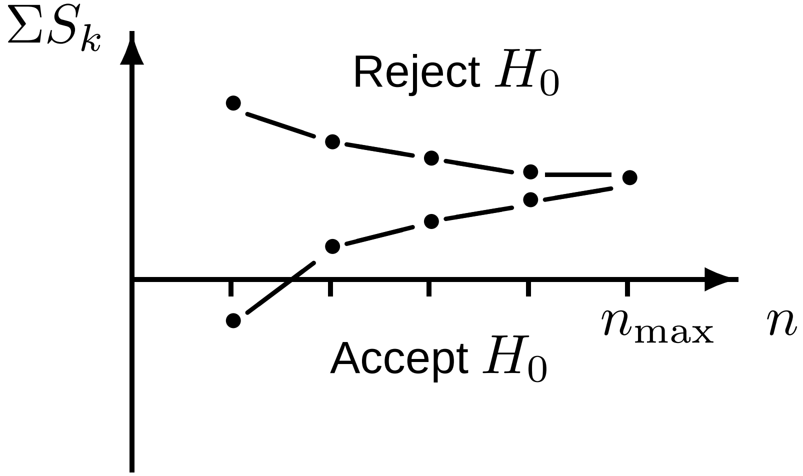
§3 Example 1: Re-design in response to external information

Original error-spending design:

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

Five group error-spending test, ρ-family with $\rho = 3$ (JT, Ch. 7.3),

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.

Cui et al. design change at analysis 2:

Group 3

Original plan: $S_3 =$ sum of $n_{\max}/5$ terms $(X_{Ai} - X_{Bi})$

Revised plan: $S_3' =$ sum of $\gamma\,(n_{\max}/5)$ terms $(X_{Ai} - X_{Bi})$

Use $\gamma^{-1/2}\,S_3'$ in place of S_3, preserving the null distribution.

Groups 4 and 5 — similarly.

Design modification (external information)

We are now aiming for power 0.9 at $\theta = \delta/2$.

So, choose the modification factor γ such that conditional power given observed data S_2 and $\theta = \delta/2$ is 0.9.

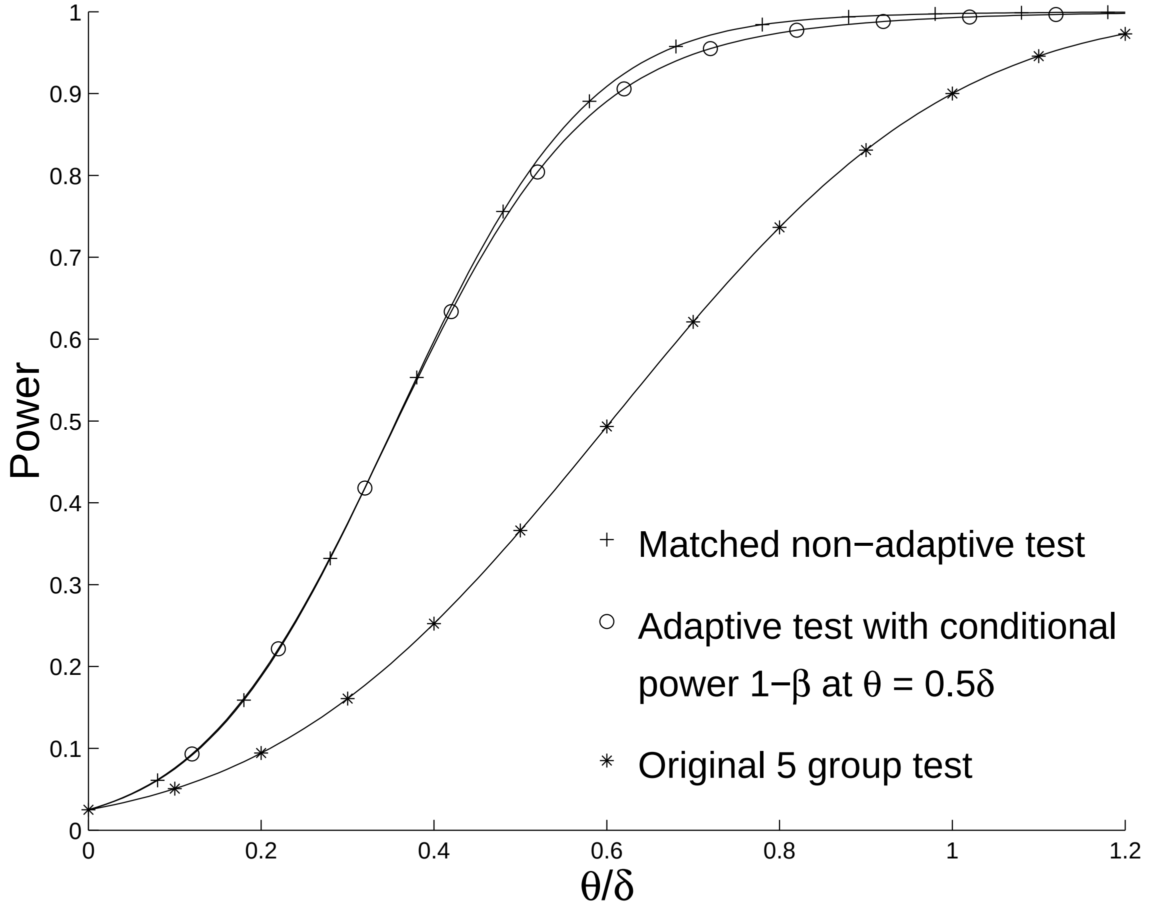
Truncate γ to

≥ 1 i.e., no decrease in sample size for high values of S_2,

≤ 6 so total sample size increases by at most a factor of 4.

Note: the likelihood of stopping by analysis 2 under the original group sequential rule is quite small for θ in the range $\delta/2$ to δ.

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



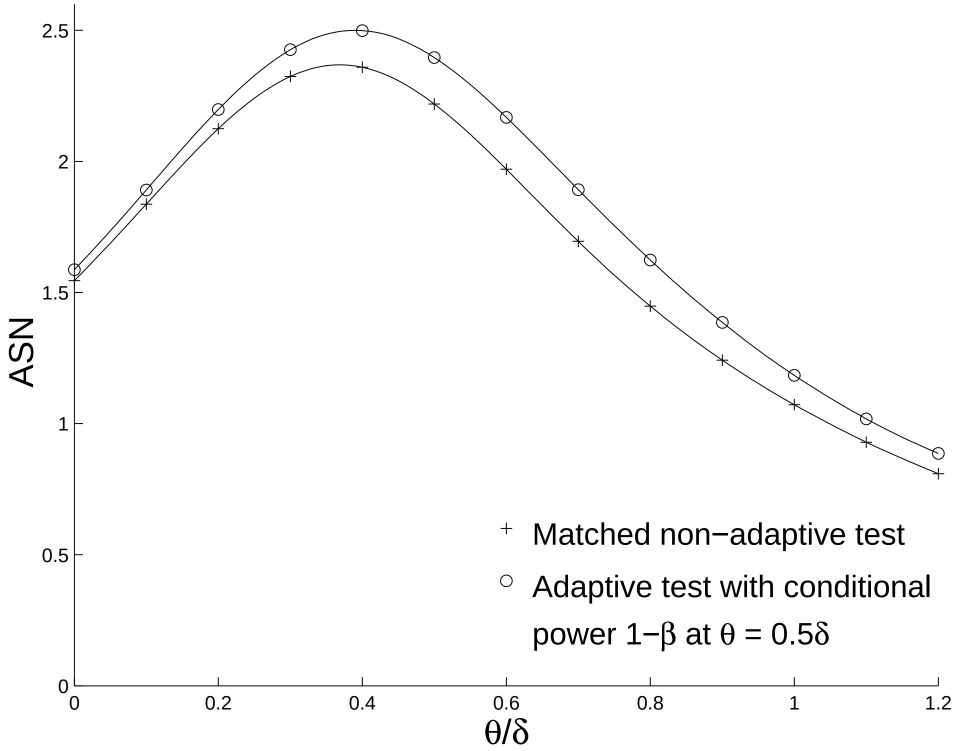
A “matched” non-adaptive test

Suppose the need for power at $\theta = \delta/2$ had been known initially: how much more efficiently could we have attained the power of the adaptive test?

The power curve of the Cui et al. test is matched by a non-adaptive test with power 0.9 at $\theta = 0.59\,\delta$.

Choosing a 5-group, ρ-family error-spending test with $\rho = 0.75$ and analyses at 0.1, 0.2, 0.45 and 0.7 of the maximum sample size gives an expected sample size curve similar in shape to that of the adaptive test.

Figure 2. Average Sample Number (ASN) curves of Cui et al. adaptive test and matched non-adaptive 5 group test with power 0.9 at $\theta = 0.59\,\delta$.



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

Comparing both power and ASN

In "Adaptive and non-adaptive group sequential tests" (2004), Jennison & Turnbull propose an overall measure for comparing power curves and ASN curves of two tests.

If tests A and B with type I error rate α have:

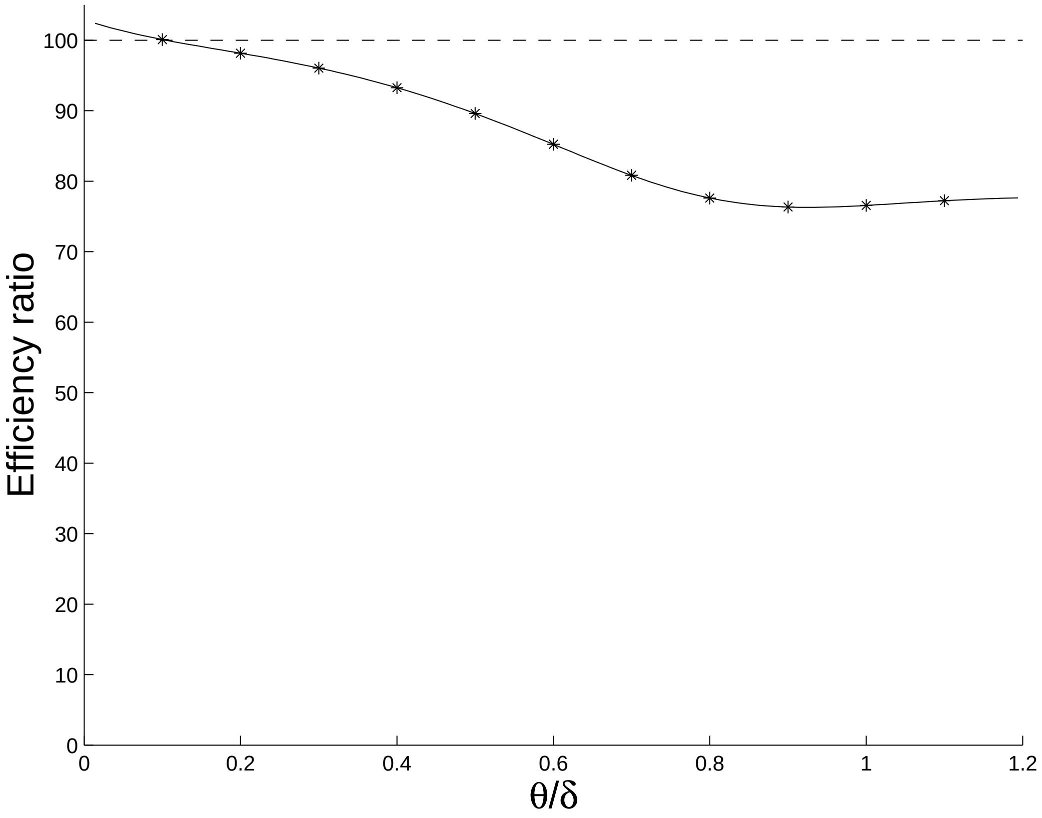
power curves $\quad 1 - b_A(\theta),\ 1 - b_B(\theta),$

and ASN curves $\quad E_{A,\theta}(N), \quad E_{B,\theta}(N),$

their efficiency ratio at θ is defined as

$$ER_{A,B}(\theta) \ = \ \frac{E_{B,\theta}(\mathcal{I})}{E_{A,\theta}(\mathcal{I})}\ \frac{\{z_\alpha + z_{b_A(\theta)}\}^2}{\{z_\alpha + z_{b_B(\theta)}\}^2} \times 100.$$

Figure 3. Efficiency ratio between Cui et al. adaptive test and matched non-adaptive test.



The efficiency ratio compares ASN with adjustment for differences in power.

Example 2: Re-design in response to internal information

Original error-spending design: As before.

Intervention: At analysis 2, set target of power 0.9 at $\theta = \widehat{\theta}_2$.

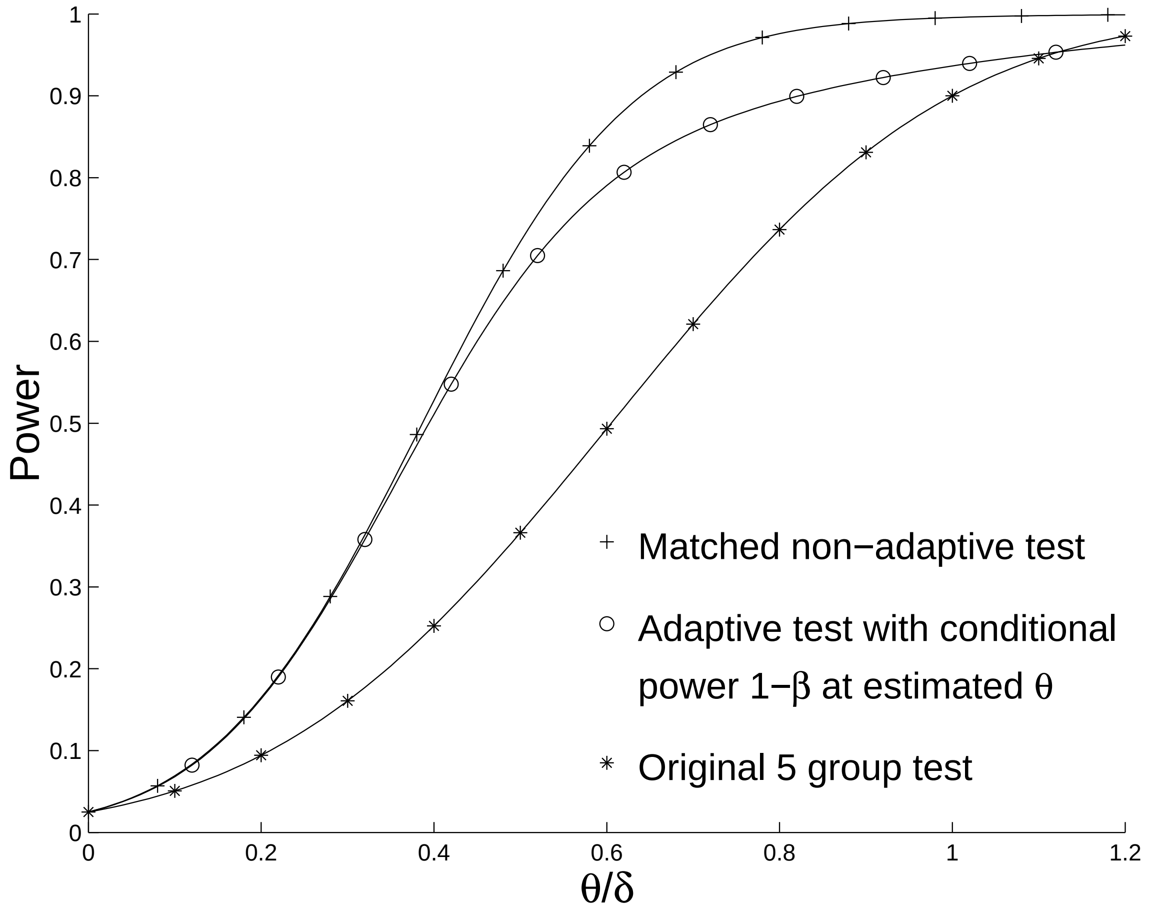
Aim to achieve this by choosing γ such that conditional power given observed data S_2 under $\theta = \widehat{\theta}_2$ is 0.9.

Allow $\gamma < 1,$ i.e., a decrease in sample size for high values of S_2.

Restrict to $\gamma \leq 6,$ so total sample size rises by at most a factor of 4.

Note: $\theta = \widehat{\theta}_2$ is a noisy *estimate* of θ.

Figure 4. Power functions of original $\rho = 3$ error-spending test and Cui et al. adaptive test.



A “matched” non-adaptive test

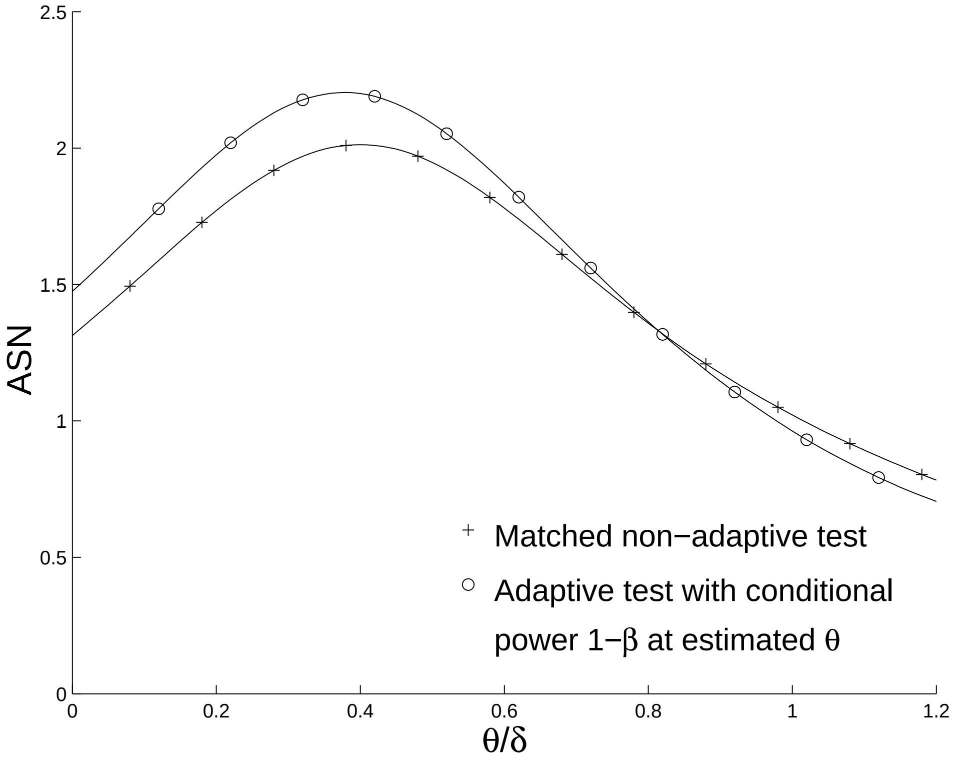
Investigators could have thought ahead about how they would react to disappointing results.

Suppose the power curve of the adaptive test is in keeping with such considerations. Are there efficient non-adaptive designs that could have been chosen at the outset?

The power curve of the Cui et al. test is matched by a non-adaptive test with power 0.9 at $\theta = 0.64\,\delta$.

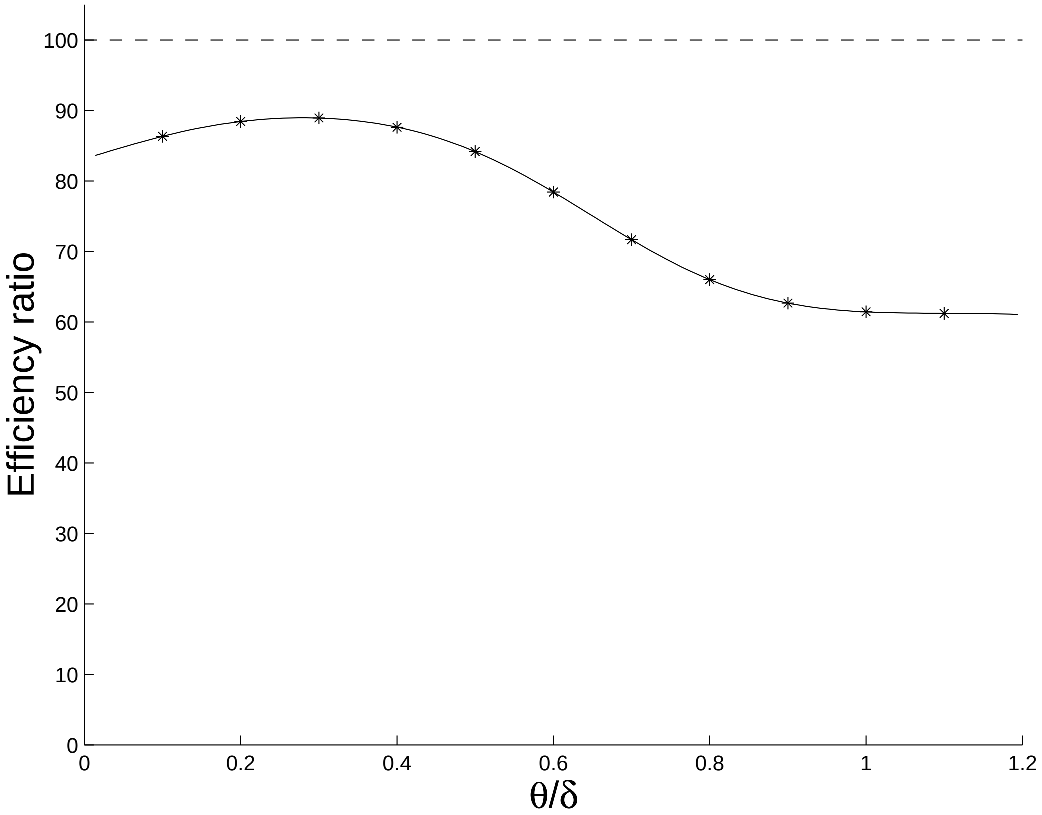
Choosing a 5-group, ρ-family error-spending test with $\rho = 0.75$ and analyses at 0.1, 0.2, 0.45 and 0.7 of the maximum sample size gives an expected sample size curve similar in shape to that of the adaptive test.

Figure 5. Average Sample Number (ASN) curves of Cui et al. adaptive test and matched non-adaptive 5 group test with power 0.9 at $\theta = 0.64\,\delta$.



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

Figure 6. Efficiency ratio between Cui et al. adaptive test and matched non-adaptive test.



The efficiency ratio compares ASN with adjustment for differences in power.

Discussion of examples

We noted the motivation for adaptive designs — to respond to external changes or to rescue an under-powered study.

Proposals for adaptive methods go beyond this, suggesting

- an appealing, flexible approach for running clinical trials,
- an alternative methodology to standard group sequential tests.

In Examples 1 and 2, adaptivity leads to inefficiency.

- Must this always be the case?
- Can adaptivity actually be beneficial for efficiency?

§4 Theory

Consider tests of H_0: $\theta = 0$ against $\theta > 0$ with

- a maximum of K analyses,
- cumulative sample sizes chosen from $\{n_1, \ldots, n_M\}$ in a data-dependent manner.

A good procedure has

- low $P_\theta\{\text{Reject } H_0\}$ for $\theta \leq 0$,
- high $P_\theta\{\text{Reject } H_0\}$ for $\theta > 0$,
- low $E_\theta(N)$ for all θ.

Theory: Admissible tests

A test is INADMISSIBLE if another test is at least as good in all respects and superior in some.

A test which is not INADMISSIBLE is ADMISSIBLE.

Jennison & Turnbull (2004) prove that

- any test which is not a function of sufficient statistics is INADMISSIBLE,
- each ADMISSIBLE test solves a BAYES decision problem for some choice of prior, cost function for wrong decisions, and sampling cost function.

Theory: Implications

(a) Adaptivity can be beneficial if used well, i.e., the best K-group adaptive test is superior to the best K-group non-adaptive test.

(b) With many analyses $(K = M)$ this advantage is lost. Hence,

non-adaptive tests with extra analyses do just as well,

non-adaptive tests with large K out-perform adaptive tests based on non-sufficient statistics (cf Tsiatis & Mehta, 2003).

Questions

- How great are the benefits in (a) for small values of K ?
- Does this provide a margin of error for flexible designs using non-sufficient statistics?

§5 Optimal planned adaptive tests (Schmitz, 1993)

In *Optimal Sequentially Planned Decision Procedures*, Schmitz proposes procedures which run as follows.

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

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

then 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.

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 — note the link with our theoretical results that admissible frequentist tests are solutions of Bayes decision problems.

Jennison & Turnbull (2004) extend this approach, with rather more computation, to yield optimal adaptive group sequential tests.

Optimal tests: Example

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

with type I error rate $\alpha = 0.025$

and power $1 - \beta = 0.9$ at $\theta = \delta$.

Aim for low values of

$$\int E_\theta(N) f(\theta)\, d\theta,$$

where $f(\theta)$ is the density of a $N(\delta,\, \delta^2/4)$ distribution.

Constraints:

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

Maximum number of analyses $= K$.

Optimal average ASN

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

Number of analyses, K	*Optimal adaptive, group sequential design (Schmitz)*	*Non-adaptive, optimised group sizes*	*Non-adaptive, equal group sizes*
2	72.5	73.2	74.8
3	64.8	65.6	66.1
4	61.2	62.4	62.7
6	58.0	59.4	59.8
8	56.6	58.0	58.3
10	55.9	57.2	57.5

Conclusions

- Effective non-adaptive group sequential tests are readily available.
- Pre-planned adaptive designs can do a little better, but perhaps not enough to compensate for their complexity.
- There are dangers in uninformed use of flexible adaptive designs.
  We recommend against using such methods to put off the decision on a study’s power requirement until one sees some interim data.
- A key role for adaptive methods is in adapting to a change in objectives due to *external* factors — with protection of the type I error rate.
- They can also rescue a study found to lack power at an interim stage.