TO ADAPT OR NOT TO ADAPT

Professor Christopher Jennison,

Dept of Mathematical Sciences,

University of Bath, UK

Boston

25 January 2005

Classical group sequential tests (GSTs)

Pocock (1977), O'Brien & Fleming (1979):

Two-sided tests with early stopping to reject H_0.

Then

One-sided tests, equivalence testing. Early stopping for futility.

Unpredictable group sizes — error spending tests.

General response distributions, survival data.

Flexible monitoring: repeated confidence intervals, stochastic curtailment.

Optimized designs.

Sample size re-estimation (internal pilots).

Multiple endpoints, multi-arm trials.

Adaptive treatment allocation: for balance, to reduce inferior treatment.

Modern adaptive methods

Bauer (1989), Bauer & Köhne (1994), Lehmacher & Wassmer (1999):

Re-defining treatment or major endpoint.

Responding to external events or to internal information (e.g., safety).

Adapting to nuisance parameters (e.g., variance).

Adapting to interim estimates of effect size.

Cui, Hung & Wang (1999):

Rescuing an under-powered study.

Possibility of using this approach to delay the final choice of power until there is interim information on the effect size.

Modern adaptive methods, continued

Proschan & Hunsberger (1995), Shen & Fisher (1999):

Designs with modification of sample size built-in.

Setting sample size for conditional power under estimated effect size.

Ad hoc proposals tend to improve on fixed sample designs but to be less efficient than competing GSTs.

Then

Posch, Bauer & Brannath (2003):

Optimizing within defined classes of designs.

Liu, Anderson & Pledger (2004):

Optimizing designs for “commercial utility”.

Remarks on GSTs and Adaptive tests

Development of GSTs has matured to meet practical needs and to address specialised problems.

Recently proposed adaptive designs have additional breadth, particularly:

Re-defining treatment or major endpoint.

Responding to external events.

Rescuing an under-powered study.

Some of these features could be incorporated in classical GSTs, especially the ability to response to external events by re-design preserving the conditional type I error probability.

Criticisms of adaptive designs

First a positive comment

In improving on a fixed sample design, any sensible step towards a scheme of interim monitoring and possible early stopping should be helpful. Further gains from finding the best sequential design may be relatively small.

Thus, recent publicity for adaptive designs is good as it draws attention to the benefits of interim monitoring and early stopping.

Critical comments

1. Many proposals of adaptive designs in the literature are *demonstrably inferior* to standard GSTs.

2. The notion that one can use interim data to refine a study’s power requirement leads to confused thinking and inefficient study designs.

1. Efficiency of GSTs and adaptive tests

Problem formulation:

$\theta =$ effect size

to test H_0: $\theta = 0$ against $\theta > 0$ with

type I error rate α under $\theta = 0$ and power $1 - \beta$ at $\theta = \delta$,

minimizing the objective function

$$H = \sum_{\theta} w(\theta) E_{\theta}(N) \quad \text{or} \quad H = \int w(\theta) E_{\theta}(N)\, d\theta .$$

Fix maximum number of analyses K and

maximum sample size $= R \times$ (fixed sample size).

Special cases: $K = \infty$ for continuous monitoring,

$R = \infty$ for unconstrained maximum sample size.

Efficiency of GSTs and adaptive tests

Both adaptive and non-adaptive tests have the same basic form:

Sample the first stage with its specified group size,
decide whether to stop or continue,
[†]choose group sizes and stopping boundary values
for second stage in the light of first stage data.

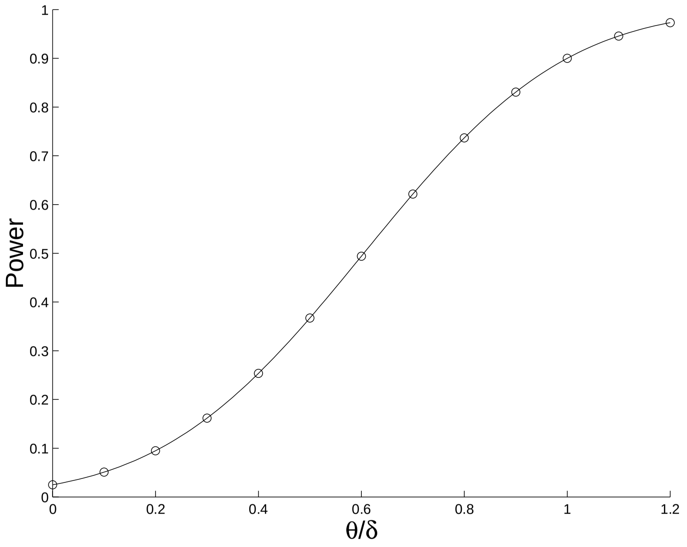
Sample the second stage with its specified group size,
decide whether to stop or continue,
[†]choose group sizes and stopping boundary values
for third stage in the light of second stage data.

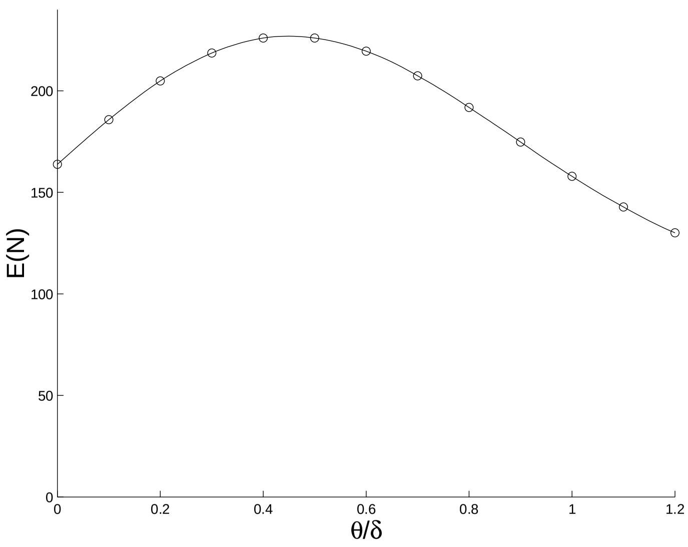
. . .

[†]Only for adaptive designs — group sizes and boundaries have to be pre-specified in the non-adaptive case.

Efficiency of GSTs and adaptive tests

Any design, adaptive or non-adaptive, has an *overall* power function and expected sample size function.





Matching the power curves at α for $\theta = 0$ and $1 - \beta$ for $\theta = \delta$ gives comparability, as does fixing K and R.

Efficiency of GSTs and adaptive tests

Since the class of adaptive tests is larger, it should yield a lower minimized objective function H.

But we still need to ask:

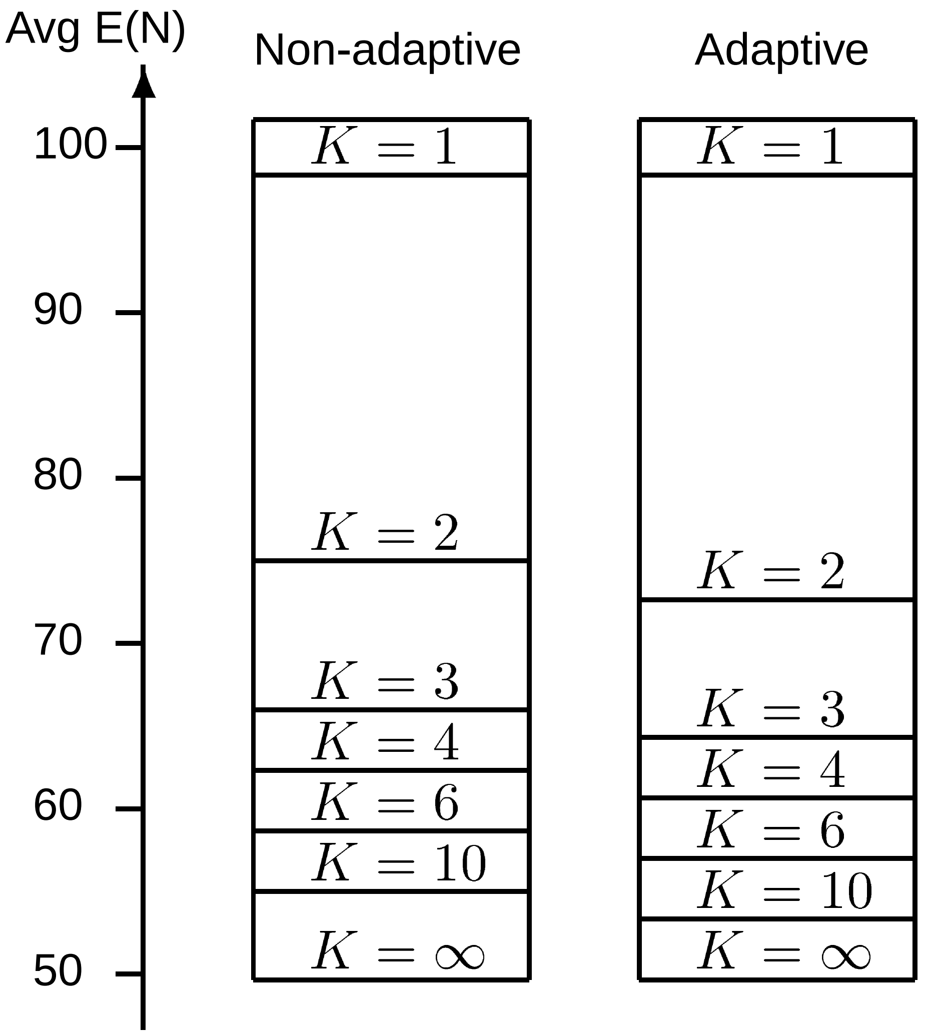
How great can the benefit be of adaptively choosing group sizes?

How efficient are adaptive tests proposed in the literature?

Are they better or worse than matched non-adaptive GSTs?

What are the characteristics of efficient designs, both adaptive and non-adaptive?

Efficiency of GSTs and adaptive tests



All optimal designs are based on sufficient statistics.

Optimal adaptive designs have *very specific* sample size rules.

Various proposals for adaptive designs are not close to optimal.

Jennison & Turnbull (2005) “Adaptive and non-adaptive group sequential tests”.

Commercial utility

1. Jennison & Turnbull derived optimal adaptive and non-adaptive tests in a decision formulation with costs for

$c_1\, P_{\theta=0}(\text{Reject } H_0)$ *penalty for type I error*,

$-\, c_2\, P_{\theta=\delta}(\text{Reject } H_0)$ *reward for positive study outcome*,

$\int_\theta w(\theta) E_\theta(N)\, d\theta$ *sampling cost*.

2. Liu, Anderson & Pledger adopt a “commercial utility” taking the same type I error and sampling costs but with

$-\int_{\theta>0} c(\theta) P_\theta(\text{Reject } H_0)\, d\theta$ as *reward for a positive outcome*

and adding

$\int_{\theta>0} k(\theta) E_\theta\{N.I(\text{Reject } H_0)\}\, d\theta$ *opportunity cost of study length*.

Commercial utility

In both cases, cost of a type I error is adjusted to give type I error rate α.

In (1), the type II error cost is adjusted to give a fixed type II error rate β.

In (2), the costs are obtained from a commercial model and the type II error rate is not constrained.

For either formulation, we can optimise over non-adaptive group sequential designs *or* optimise over adaptive designs which allow data-dependent choice of group sizes.

In both cases, the gain from adaptivity is small when expressed in percentage terms. Since the sums involved in (2) are millions of dollars, this can still be interesting — so sound statistical advice has a high value!

2. Using interim data to refine a power requirement

Thinking: We would like power 0.9 under the *true* effect size, θ, but we don't know what this is. Example: an optimistic view is $\theta = 20$, the minimal clinically significant effect is $\theta = 10$.

Start the study off with power 0.9 under $\theta = 20$.

At the half-way stage, examine the interim estimate $\widehat{\theta}_I$ and re-size the study. Preserve the conditional type I error probability and achieve conditional power 0.9 at

$$\begin{aligned} \theta &= 20 && \text{if } \widehat{\theta}_I \geq 20, \\ \theta &= \widehat{\theta}_I && \text{if } 10 < \widehat{\theta}_I < 20, \\ \theta &= 10 && \text{if } \widehat{\theta}_I \leq 10. \end{aligned}$$

Possibly stop for futility if $\widehat{\theta}_I$ is really low.

Using interim data to refine a power requirement

(i) ***A practical problem*** — $\widehat{\theta}_I$ is a highly variable estimate of θ.

Suppose 100 observations are enough to test H_0: $\theta = 0$ against $\theta > 0$ with one-sided type I error rate 0.025 and power 0.9 at $\theta = 20$.

Then, with 100 observations, the standard error of $\widehat{\theta}$ is

$$\text{s.e.}\,(\widehat{\theta}) \;=\; 6.2.$$

After just 50 observations, the standard error of $\widehat{\theta}_I$ is

$$\text{s.e.}\,(\widehat{\theta}_I) \;=\; 6.2 \times \sqrt{2} \;=\; 8.7.$$

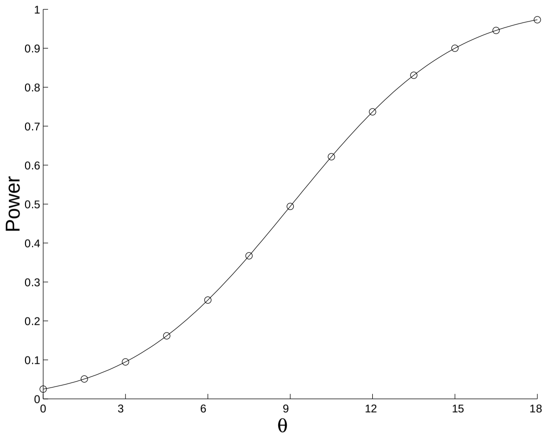
So if $\widehat{\theta}_I = 12,$ a 95% confidence interval for θ is $(-5,\; 29)$.

Treating this estimate as accurate is a source of inefficiency.

Using interim data to refine a power requirement

(ii) ***This is confused thinking anyway.***

Once the rule for re-calculating sample size is stated, a design has been specified. This has a power curve.



So, the power at each possible value of θ is determined at the outset anyway — just as in a more standard approach.

And, we could have achieved this power curve directly.