Flexible Clinical Trial Design

Christopher Jennison

Department of Mathematical Sciences,

University of Bath, UK

http://people.bath.ac.uk/mascj

Föreningen för Medicinsk Statistik

Göteberg,

September 24, 2007

1. Protecting the type I error probability

The importance of avoiding inflation of the type I error rate in an adaptive or flexible design cannot be over-emphasized.

- ICH E9 (p. 25)

  "The procedures selected should always ensure that the overall probability of type I error is controlled."

- PhRMA White paper (*J. Biopharmaceutical Statistics*, 2006)

  "The key issue in most contexts is preservation of the type I error rate."

- Pocock & Hughes (*Controlled Clinical Trials*, 1989)

  "Control of type I error is a vital aid to prevent a flood of false positives into the medical literature."

2. Combining P-values: Bauer & Köhne's method

Bauer & Köhne (*Biometrics*, 1994).

Initial design

Stipulate that Bauer & Köhne's combination test will be used.

Define the null hypothesis H_0.

Design Stage 1, fixing sample size and test statistic for this stage.

Stage 1

Observe the P-value, P_1.

Under H_0, $P_1 \sim U(0,\, 1)$.

Design Stage 2 in the light of Stage 1 data.

Stage 2

Observe the P-value, P_2.

Under H_0, $P_2 \sim U(0,\, 1)$ and P_2 is independent of P_1.

Bauer & Köhne's two-stage method

Overall test

If $P \sim U(0,\, 1)$, then

$$-\ln(P) \;\sim\; \text{Exp}\,(1) = \frac{1}{2}\,\chi^2_2.$$

Thus, under H_0,

$$-\ln(P_1\,P_2) \;\sim\; \frac{1}{2}\,\chi^2_4$$

and we combine the two P-values in an overall test, rejecting H_0 if

$$-\ln(P_1\,P_2) \;>\; \frac{1}{2}\,\chi^2_{4,\,1-\alpha}.$$

This χ^2 test was originally proposed for combining results of several studies by R. A. Fisher (1932) *Statistical Methods for Research Workers*.

3. Sample size re-estimation on estimating response variance

Consider a two-treatment comparison with normal responses of unknown variance, $X_{Ai} \sim N(\mu_A,\, \sigma^2)$ and $X_{Bi} \sim N(\mu_B,\, \sigma^2)$ on treatments A and B, respectively.

Initial design

Investigators decide to test H_0: $\theta = \mu_A - \mu_B \leq 0$ against $\theta > 0$ with type I error rate α, using a Bauer & Köhne two-stage design.

Sample size, n_0 per treatment, is determined that will give power $1 - \beta$ at $\theta = \delta$ if the variance is equal to their initial estimate σ_0^2.

Stage 1 is planned with a sample size of $n_1 = n_0/2$ per treatment.

Stage 1

Yields estimates $\hat{\theta}_1$ and $\hat{\sigma}_1^2$, plus the t-statistic t_1 for testing H_0 vs $\theta > 0$.

Convert t_1 to a P-value, $P_1 = Pr_{\theta=0}\{T_{2\,n_1-2} > t_1\}$.

Sample size re-estimation

Stage 1 . . .

Now use the variance estimate $\hat{\sigma}_1^2$ to re-calculate sample size.

One may simply use this value in the original calculation, in place of σ_0^2.

Or, perhaps, choose the Stage 2 sample size to give conditional power $1 - \beta$, given P_1, assuming $\theta = \hat{\theta}_1$ and $\sigma^2 = \hat{\sigma}_1^2$.

This defines an additional sample size of n_2 per treatment arm in Stage 2.

Stage 2

Calculate the t-statistic t_2 for testing H_0 vs $\theta > 0$ based on Stage 2 data alone.

Convert t_2 to a P-value, $P_2 = Pr_{\theta=0}\{T_{2\,n_2-2} > t_2\}$.

The overall test — which has type I error rate exactly α — rejects H_0 if

$$-\ln(P_1\,P_2) \;>\; \frac{1}{2}\,\chi^2_{4,\,1-\alpha}.$$

Other approaches to sample size re-estimation

Adaptive designs add to the variety of methods for dealing with an unknown parameter, ϕ, affecting the sample size needed to attain a specific power.

Internal pilot:

Wittes & Brittain (*Statistics in Medicine*, 1990) propose a simple "plug in" of the current estimate $\hat{\phi}$ to update the remaining sample size.

A small inflation of the type I error rate can result from bias in the final $\hat{\phi}$.

Information monitoring:

Mehta & Tsiatis (*Drug Information J.*, 2001) use a similar "plug in" of estimated information within an error spending group sequential design.

For normal response with unknown σ^2, the small error rate inflation can be reduced by use of better approximations (Denne & Jennison, *Biometrika*, 2000).

4. Sample size modification in response to new information

Consider a study to compare failure rates p_c and p_t on control and experimental treatments, respectively. Historical data indicate $p_c \approx 0.25$.

Writing $\theta = p_c - p_t$, it is desired to test H_0: $\theta \leq 0$ against $\theta > 0$ with type I error rate $\alpha = 0.025$ and power $1 - \beta = 0.9$ if $\theta = 0.05$.

Initial design

A Bauer & Köhne two-stage design is specified.

With type I error rate and power as specified above, assuming $p_c = 0.25$, a fixed sample test needs 1461 subjects per treatment arm.

Stage 1 is planned with 730 subjects per treatment, with a view to re-assessing requirements for the remainder of the study in the light of their responses.

Sample size modification

Stage 1

Yields $\hat{p}_c = 0.253$ and $\hat{p}_t = 0.219$.

Hence $\hat{\theta} = 0.034$ with standard error 0.0222.

A test of H_0: $\theta \leq 0$ has Z-statistic $0.034/0.0222 = 1.53$ and P-value

$$P_1 \; = \; 1 - \Phi(1.53) = 0.0629.$$

The overall test will reject H_0 if $-\ln(P_1\, P_2) > 0.5\, \chi^2_{4,\,0.975} = 5.57$.
Since $-\ln(0.0629) = 2.77$, results thus far are promising. However, a positive outcome is by no means certain.

It is learnt that trials of competing treatments have been unsuccessful.

Based on perceived costs and benefits, it is decided to increase the second phase sample size.

Sample size is chosen to give higher probability of a positive outcome under the original alternative, $\theta = 0.05$ — and under smaller effect sizes.

Sample size modification

Planning Stage 2 sample size

p_c	p_t	θ	Stage 2 sample size	Conditional power
0.25	0.22	0.03	750	0.43
			1000	0.51
			1250	0.59
0.25	0.21	0.04	750	0.62
			1000	0.72
			1250	0.80
0.25	0.20	0.05	750	0.78
			1000	0.87
			1250	0.93

Sample size modification

Stage 2

Chosen sample size: 1000 patients per treatment arm in Stage 2.

Yields $\hat{p}_c = 0.251$ and $\hat{p}_t = 0.221$ from Stage 2 data alone.

Hence $\hat{\theta} = 0.030$ with standard error 0.0190.

A test of H_0: $\theta \leq 0$ has Z-statistic $0.030/0.0190 = 1.58$ and, thus, P-value

$$P_2 \;=\; 1 - \Phi(1.58) = 0.0570.$$

Applying the overall test,

$$-\ln(P_1\, P_2) \;=\; -\ln(0.0629) - \ln(0.0570) \;=\; 2.77 + 2.87 \;=\; 5.64.$$

This is greater than $0.5\, \chi^2_{4,\,0.975} = 5.57$, so H_0 is rejected and it is concluded that the new treatment has a lower failure rate.

5. Increasing sample size for disappointing results

Example: *(JT, Biometrika, 2006, Ex. 2)*

Consider a study to investigate a new treatment *vs* control, where responses are $N(\mu_t, \sigma^2)$ on treatment, $N(\mu_c, \sigma^2)$ on control, and $\theta = \mu_t - \mu_c$.

The study will test H_0: $\theta \leq 0$ against $\theta > 0$ with

one-sided type I error probability $\alpha = 0.025$,

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

A fixed sample size test needs n_f observations per treatment where

$$n_f = (z_\alpha + z_\beta)^2 \, 2\sigma^2/\delta^2.$$

Suppose a group sequential design is specified with 5 analyses.

A group sequential design

An "error spending" group sequential design

Boundaries are set such that cumulative type I and type II error rates follow a given function of information (equivalently, sample size).

We assume a ρ-family error spending design is chosen with $\rho = 3$, then error is spent in proportion to $\mathcal{I}^3$.

This gives a slow rate of spending error early on, so boundaries are wide initially.

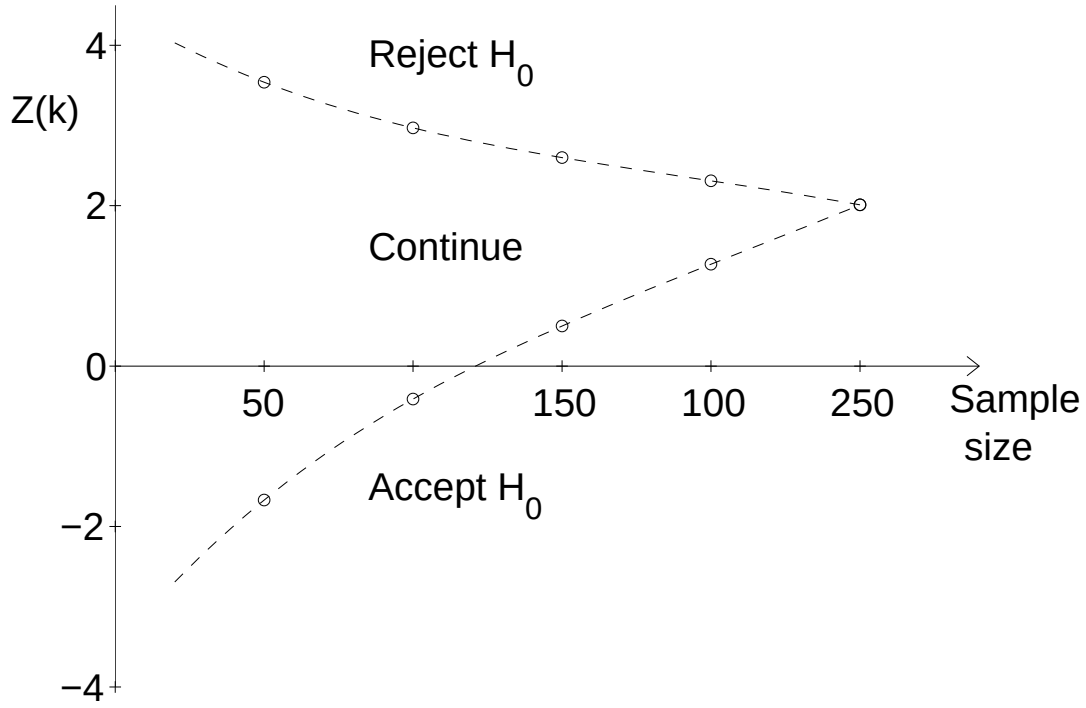
Overall sample size

The sample size "inflation factor" for this design is 1.049, that is, the design has a maximum sample size per treatment of $1.049\, n_f$.

We suppose δ and σ^2 are such that this maximum sample size is 250.

An error spending group sequential test

The 5-group error spending test has the following stopping boundary:



However, in response to the interim estimate of effect size, $\hat{\theta}_2$, at the second analysis, investigators consider changing sample size in the remaining groups.

Can later group sizes be increased?

Motivation

A low interim estimate $\hat{\theta}_2$ prompts investigators to consider the trial's power at effect sizes below the point where power 0.9 was originally set since:

lower effect sizes now plausible,

conditional power under these effect sizes, using the current design, is low.

(We assume smaller effect sizes are still of clinical significance.)

Questions

Can group sizes be increased without affecting the type I error rate?

How advisable is this "wait and see" approach?

Protecting the type I error rate

The method of Cui, Hung and Wang, Biometrics, 1999

As planned, each group 3 to 5 would provide a summary statistic

$$Y \;=\; \sum_{i=1}^{50} X_{ti} - \sum_{i=1}^{50} X_{ci} \;\sim\; N(50\,\theta,\, 100\,\sigma^2).$$

Suppose groups 3 to 5 are multiplied by a factor γ. We can define:

$$Y' \;=\; \gamma^{-1/2} \left(\sum_{i=1}^{50\gamma} X_{ti} - \sum_{i=1}^{50\gamma} X_{ci} \right) \;\sim\; N(50\,\gamma^{1/2}\theta,\, 100\,\sigma^2).$$

Using the down-weighted sum Y' in place of the original Y will:

retain the original distribution of Y under $\theta = 0$,

increase the mean of Y for $\theta > 0$, and so enhance power.

Implementing mid-study re-design

The Cui et al. (1999) approach can be applied to increase power when $\hat{\theta}_2$ is low.

We shall assume this is done for values of $Z(2)$ in the continuation region $(-0.42, 2.97)$.

Sample size rule

Suppose the value of γ is chosen to provide conditional power of 0.9 given current data, if θ is equal to $\hat{\theta}_2$.

We allow a decrease in group size ($\gamma < 1$).

We impose an upper limit $\gamma = 6$, restricting the total sample size to 1000 per treatment, i.e., 4 times the original maximum sample size.

Conditional properties at the re-design point

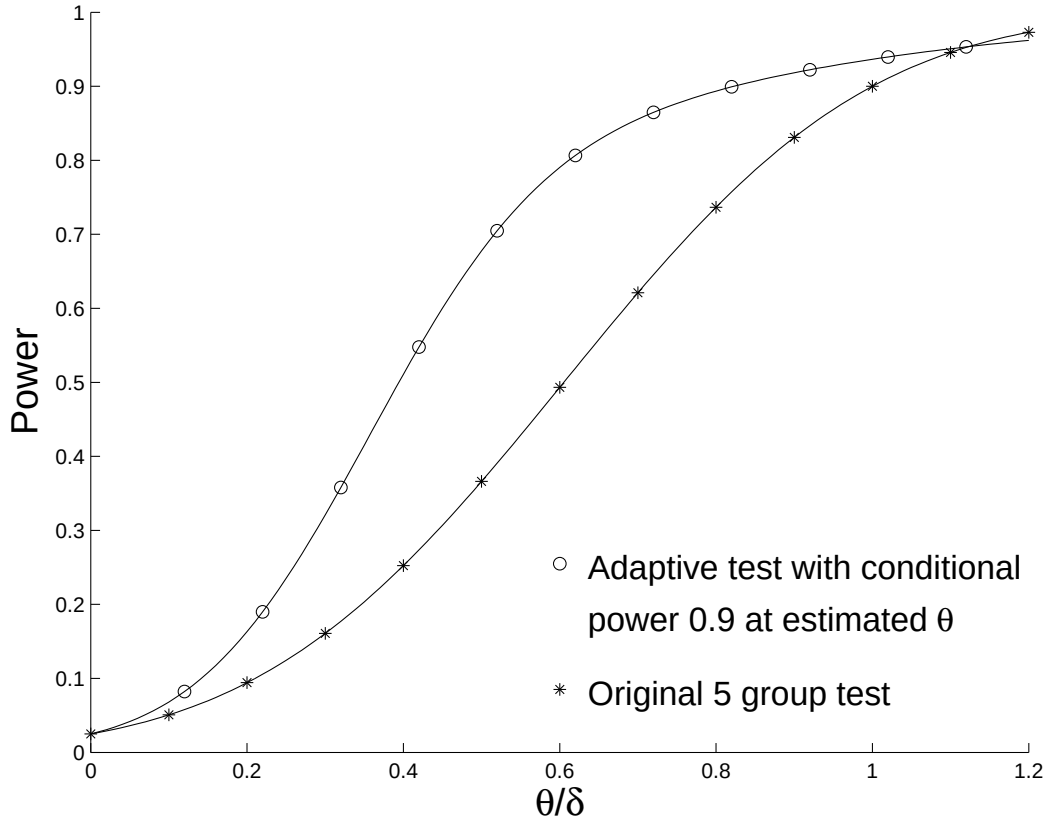
The re-designed test has the following features:

$\hat{\theta}/\delta$	z	*Conditional type I error probability*	*Conditional power at $\theta = \hat{\theta}$ before re-design*	γ	*Conditional power at $\theta = \hat{\theta}$ after re-design*
1.40	2.94	0.5707	0.9998	0.12	0.9000
1.20	2.52	0.3856	0.9959	0.30	0.9000
1.00	2.10	0.2329	0.9597	0.66	0.9000
0.80	1.68	0.1272	0.8051	1.46	0.9000
0.60	1.26	0.0630	0.4908	3.48	0.9000
0.40	0.84	0.0279	0.1825	6.00	0.7085
0.20	0.42	0.0109	0.0477	6.00	0.2236
0.00	0.00	0.0036	0.0195	6.00	0.1218
-0.20	-0.42	0.0010	0.0066	6.00	0.0554

NB, investigators will have focused on conditional properties given $Z(2) = z$.

Results of re-design

Re-design has raised the power curve.



Overall power at $\theta = \delta/2$ has increased from 0.37 to 0.68.

Is there an efficiency cost in following this adaptive approach?

Reasons for re-design arose from observing $\hat{\theta}_2$. A group sequential test responds to such interim estimates — in the decision to stop the trial or to continue.

Investigators could have considered at the design stage how they would react to low interim estimates of effect size.

If they had thought this through and chosen the above adaptive procedure, they could also have examined its overall power curve. Assuming this power curve is acceptable, how else might it have been achieved?

An alternative group sequential design

A design matching key features of the adaptive test is

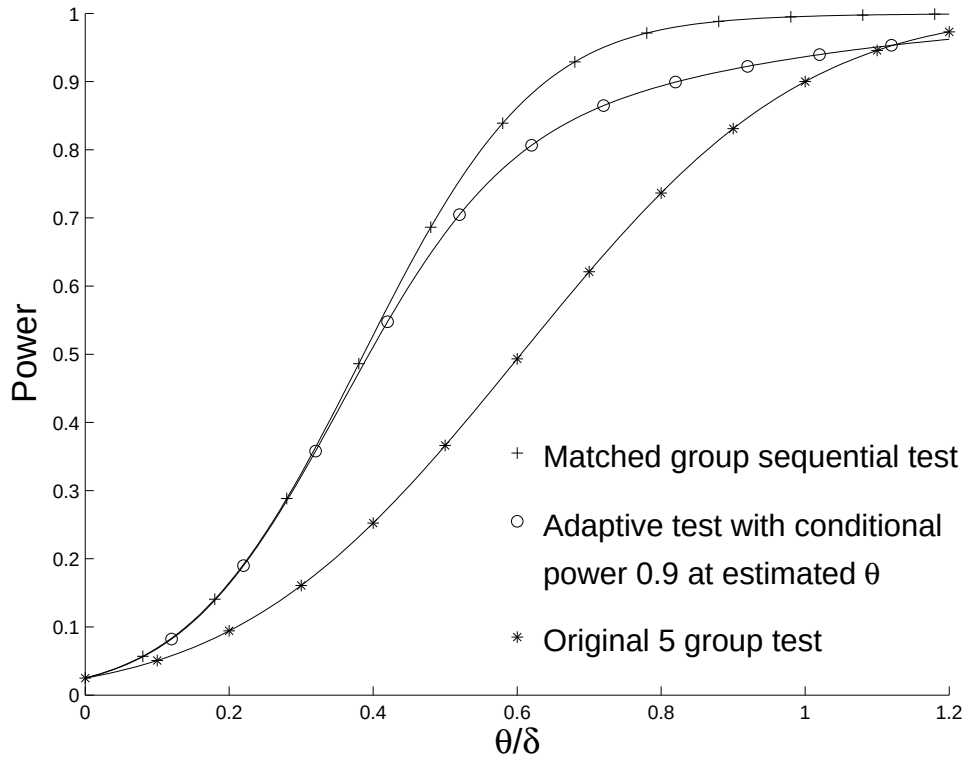
A ρ-family error spending with $\rho = 0.75$, power 0.9 at $\theta = 0.64\,\delta$,

5 analyses, the first four at $0.1, 0.2, 0.45$ and 0.7 times $n_{\max}$,

$n_{\max} = 3.21\,n_f$ (compared to $4.2\,n_f$ for the adaptive design).

A matched group sequential design

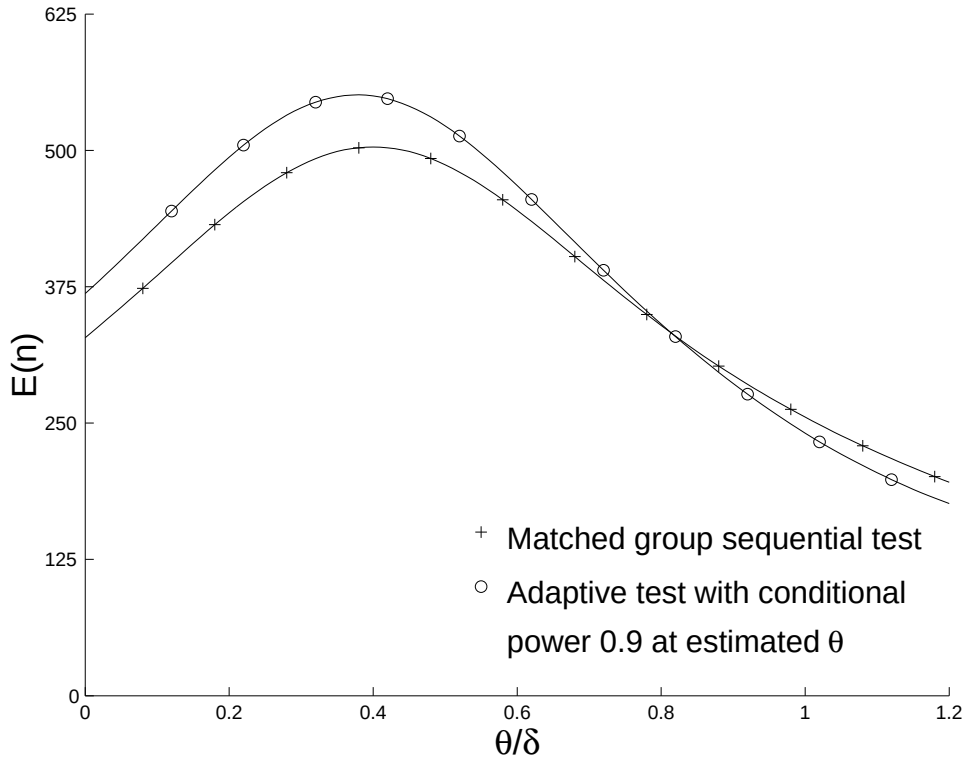
Power of the “matched” group sequential design is as high as that of the adaptive design at all effect sizes — and substantially higher at the largest θ values.



A matched group sequential design

The group sequential design has significantly lower expected sample size than the adaptive design over a range of effect sizes.

The group sequential design has slightly higher expected sample size for $\theta > 0.8\,\delta$ where its power advantage is greatest.

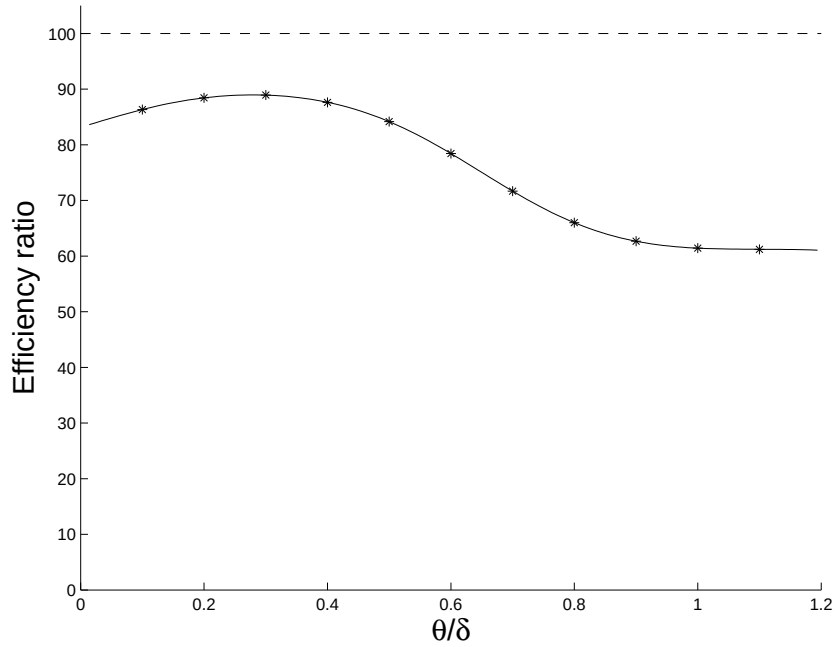


Efficiency ratio

Attained power and expected sample size can be combined in an “efficiency ratio” (JT, *Biometrika*, 2006).

The adaptive design is up to 39% less efficient than the non-adaptive, group sequential alternative.

Efficiency ratio of adaptive design vs group sequential test



Sample size adaptation in response to a low $\hat{\theta}$

We have studied a variety of proposed adaptive designs and found similar inefficiencies to the above example. These include methods of:

Proschan and Hunsberger (*Biometrics*, 1995),

Shen and Fisher (*Biometrics*, 1999) — see Jennison and Turnbull (*Bmcs*, 2006),

Li, Shih, Xie and Lu (*Biostatistics*, 2002).

Smaller adaptations increase power less but efficiency loss is still present.

Conclusion

Adaptive methods can be used to rescue an under-powered study.

We do ***not*** recommend their use for planned sample size modifications.

When there is uncertainty about the likely treatment effect, group sequential tests achieve power at the appropriate effect size — with early stopping if the true effect size is very high. (JT, *Statistics in Medicine*, 2006).

6. Credibility of a study after adaptive re-design

Study results may lack credibility if data are treated in unusual ways.

Unequal weighting of similar observations

Down-weighting some groups of observations looks strange when all subjects provide the same amount of statistical information.

> Burman and Sonesson (*Biometrics*, 2006) present extreme examples of adaptive designs which lead to highly counter-intuitive conclusions.
>
> In one example H_0: $\theta \leq 0$ is rejected, even though the estimate of θ based on the sample mean is negative!

Use of non-sufficient statistics

Unequal weighting, and many other forms of combination rule, will usually lead to conclusions based on non-sufficient statistics.

Making the most of the adaptive toolbox

Adaptive opportunities

Adaptation methods offer new approaches to

- treatment selection,
- restriction to a sub-population — "enrichment",
- change of endpoint

during the course of a study.

Need for special methods

Special methods are required to account for the testing of multiple hypotheses.

Although some complications are inevitable, the more procedures for adaptation are pre-specified, the more "reasonable" they can appear.