Adaptive Sample Size Re-estimation

Christopher Jennison

Department of Mathematical Sciences,

University of Bath, UK

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

European Leukaemia Network Workshop,

Munich

October 30–31, 2008

Initial choice of sample size

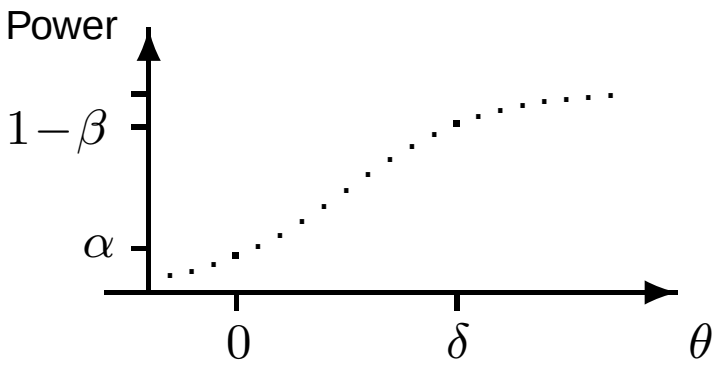
Type I error and power

Suppose θ represents the effect of a new treatment vs control.

A study is to test H_0: $\theta \leq 0$ against $\theta > 0$ with

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

The choice of sample size determines the power curve,



in particular, the effect size δ where power is $1 - \beta = 0.9$, say.

Information and sample size

Information for the effect size θ is defined as

$$\mathcal{I} \;=\; 1/\mathsf{Var}(\hat{\theta}).$$

To achieve power $1-\beta$ at $\theta = \delta$, a fixed sample size test needs information

$$\mathcal{I} \;=\; (z_\alpha + z_\beta)^2/\delta^2.$$

Information is related to sample size:

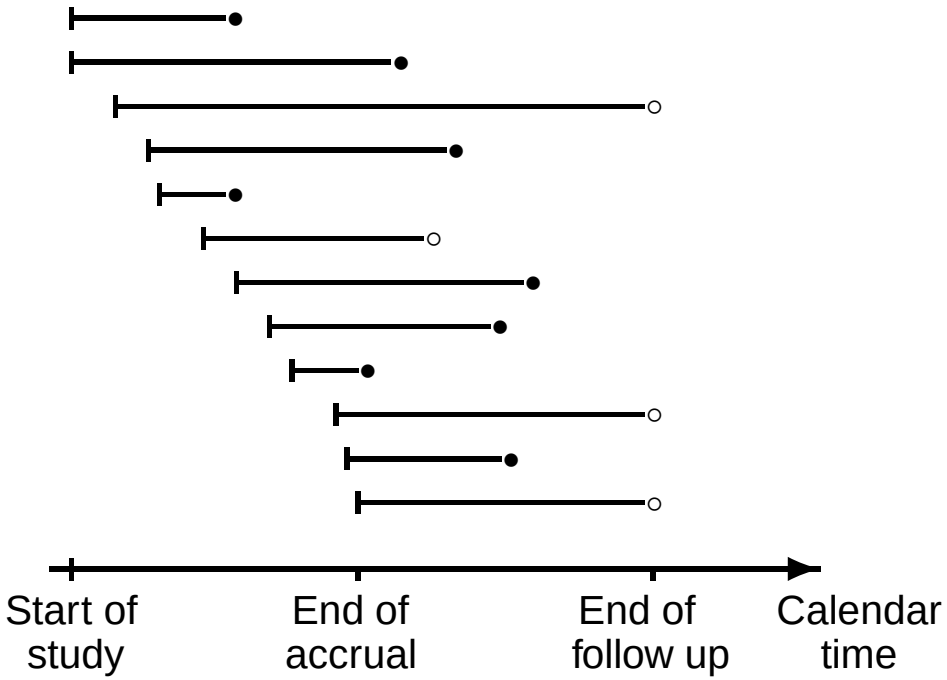
In a two-treatment comparison with ***normal responses*** of variance σ^2, a sample size of n per treatment gives

$$\mathcal{I} \;=\; n/(2\,\sigma^2),$$

In a ***survival study***, information for the log hazard ratio when d failures have been observed is approximately

$$\mathcal{I} \;=\; d/4.$$

Survival data: Accrual and follow up

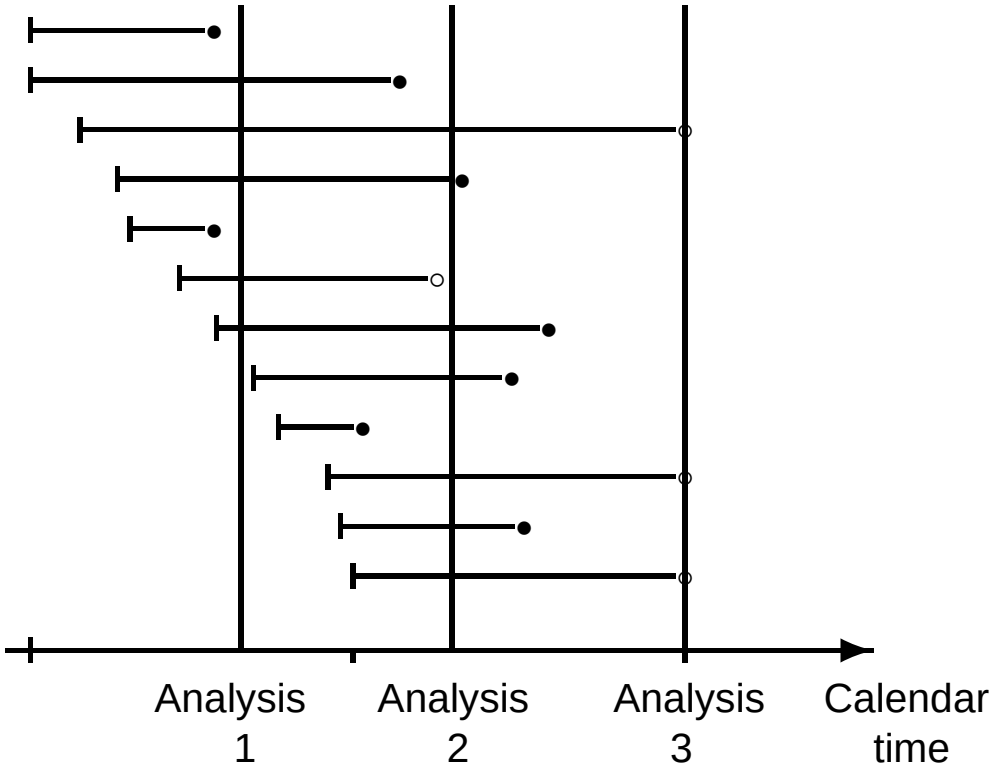


Subjects enter the study and are randomised to a treatment group.

Survival is measured from entry to the study.

Key: ● death time observed,
○ censored observation.

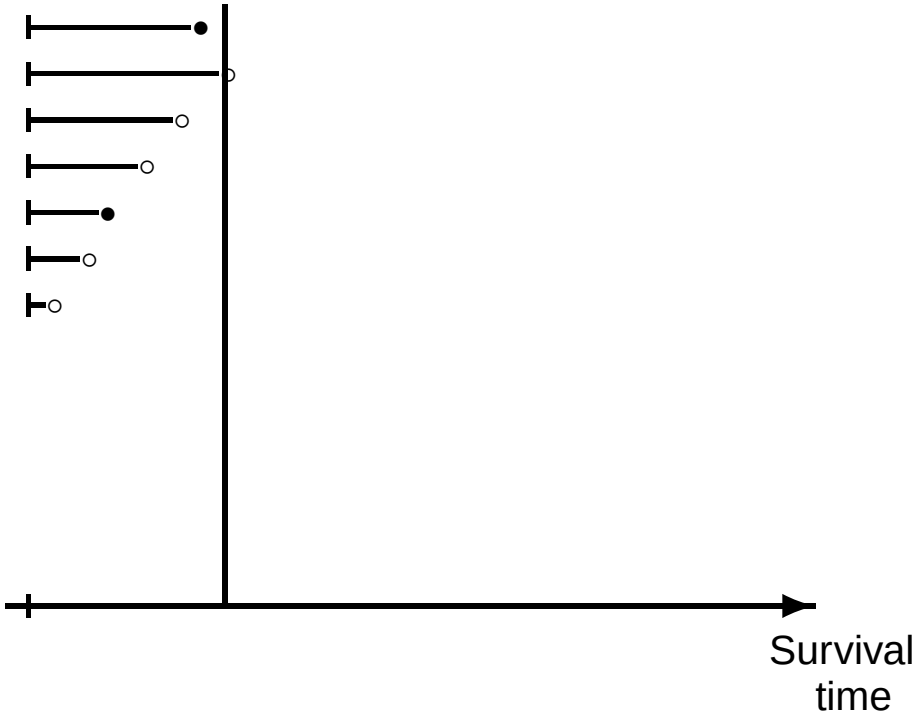
Interim analyses in a survival study



At an interim analysis, subjects are censored if they are still alive at this point.

Information on such patients will continue to accrue at later analyses.

Interim analysis 1

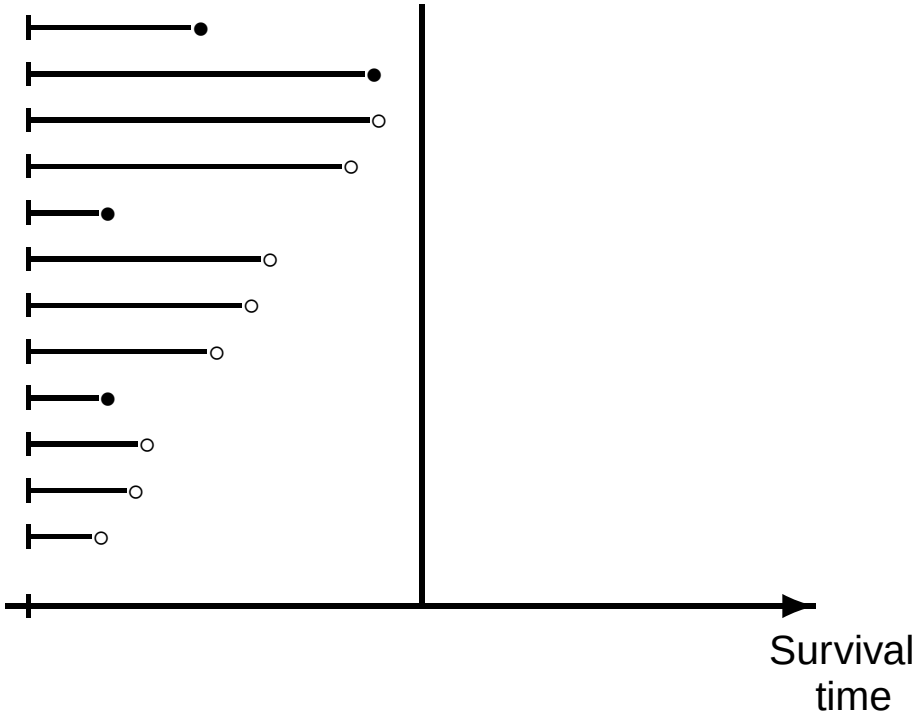


At the first analysis, we analyse available data on survival from randomisation.

These survival times have a common starting point of zero and “analysis time”

censoring occurs for subjects surviving past the first analysis.

Interim analysis 2



At the second analysis, we analyse updated data on survival from randomisation.

These times have a common starting point of zero and “analysis time” censoring occurs for subjects surviving past the second analysis.

And so on, through further analyses . . .

The logrank statistic

At stage k, denote the d_k failure times (measured from entry to study) by

$$\tau_{1,k} < \tau_{2,k} < \ldots < \tau_{d_k,k}.$$

Define for stage k:

$r_{iA,k}$ and $r_{iB,k}$	Numbers at risk on Treatments A and B at $\tau_{i,k}-$
O_k	Observed deaths on Treatment B
$E_k = \sum_{i=1}^{d_k} r_{iB,k}/(r_{iA,k} + r_{iB,k})$	"Expected" deaths on Treatment B
$V_k = \sum_{1}^{d_k} r_{iA,k} r_{iB,k}/(r_{iA,k} + r_{iB,k})^2$	"Variance" of O_k
$Z_k = (O_k - E_k)/\sqrt{V_k}$	Standardised logrank statistic

Proportional hazards model

Assume hazard rates h_A on Treatment A and h_B on Treatment B are related by

$$h_B(t) = \lambda\, h_A(t).$$

The log hazard ratio is $\theta = \ln(\lambda)$.

Then, approximately,

$\{Z_1, \ldots, Z_K\}$ are multivariate normal,

$$Z_k \sim N(\theta\sqrt{\mathcal{I}_k},\, 1), \qquad k = 1, \ldots, K,$$

$$\text{Cov}(Z_{k_1}, Z_{k_2}) = \sqrt{(\mathcal{I}_{k_1}/\mathcal{I}_{k_2})} \quad \text{for } k_1 < k_2,$$

where

$\mathcal{I}_k = V_k$, the variance of the unstandardised logrank statistic $O_k - E_k$.

Proportional hazards model

We can define the estimate of the hazard ratio

$$\hat{\theta}_k = Z_k / \sqrt{\mathcal{I}_k}.$$

Then, approximately,

$\{\hat{\theta}_1, \ldots, \hat{\theta}_K\}$ are multivariate normal

$$\hat{\theta}_k \sim N(\theta, \mathcal{I}_k^{-1}), \qquad k = 1, \ldots, K,$$

$$\text{Cov}(\hat{\theta}_{k_1}, \hat{\theta}_{k_2}) = \mathcal{I}_{k_2}^{-1} \quad \text{for } k_1 < k_2.$$

When $\lambda \approx 1$, we have $\mathcal{I}_k = V_k \approx d_k/4$ and the familiar result

$$\hat{\theta}_k \sim N(\theta, (d_k/4)^{-1}).$$

Sample size re-estimation for nuisance parameters

One can design a study in terms of “required information”.

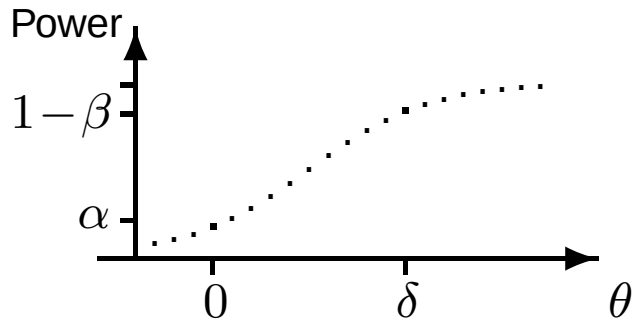
During the trial, sample size or length of follow up for survival data may be modified to reach the target information level.

Internal pilots: Wittes & Brittain (*Statistics in Medicine*, 1990)

Information monitoring: Mehta & Tsiatis (*Drug Information J.*, 2001)

Adaptive designs: Bauer & Köhne (*Biometrics*, 1994)

These modifications are intended to achieve the power curve originally specified.



What if requirements for the power curve change during a trial?

Changing power in response to external events

Example 1. *(JT, Biometrika, 2006, Ex. 1)*

A group sequential study to investigate effect size θ using an "error spending" test.

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

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

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

A fixed sample size test would need information

$$\mathcal{I}_f = (z_\alpha + z_\beta)^2/\delta^2 = 10.51/\delta^2.$$

The trial has 5 planned analyses, and spends type I and II error in proportion to $\mathcal{I}^3$.

The information level must be able to reach

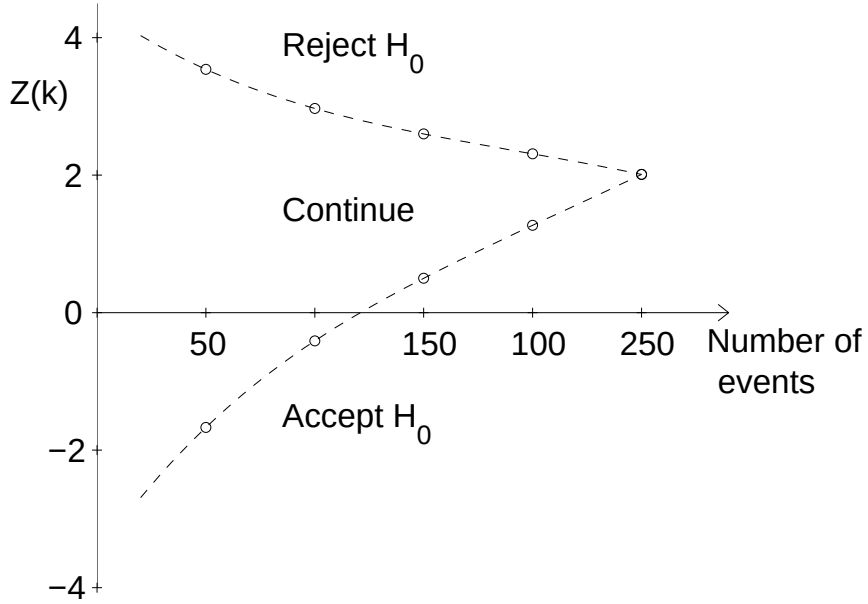
$$\mathcal{I}_{\max} = 1.049\,\mathcal{I}_f = 11.02/\delta^2.$$

Adapting sample size to external events

Suppose θ is a log hazard ratio, so $\hat{\theta}_k \sim N(\theta,\, 4/d_k)$ approximately.

With $\delta = 0.42$, i.e., power 0.9 at a hazard ratio of 1.52, maximum information is $\mathcal{I}_{\max} = 11.02/\delta^2$, equating to 250 failure events.

Then, the test has the following stopping boundary.



Now suppose external information arrives at the time of the second analysis.

Increasing later group sizes

Suppose, at the time of analysis 2:

The external environment has changed,

Investigators want power 0.9 at $\theta = \delta/2$ rather than $\theta = \delta$.

Re-design must protect the type I error rate, remembering that investigators knew the value of $Z(2)$ when deciding to do this.

The method of Cui, Hung and Wang (*Biometrics*, 1999)

Numbers of events in the remaining groups, 3 to 5, are multiplied by a factor γ, which may depend on $Z(2)$ — so increments in information are multiplied by γ.

Down-weighting observations from these groups by $\gamma^{-1/2}$ maintains the type I error rate.

Despite down-weighting, the larger numbers of events give increased power.

The Cui et al. method

In the original design, "group k" provides a score statistic for θ

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

The sum of these for groups 1 to k is the overall score statistic at analysis k.

When group size is increased by a factor γ, we have

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

However,

$$\gamma^{-1/2}\,S'_{(k)} \sim N(\gamma^{1/2}\,\mathcal{I}_{(k)}\,\theta,\, \mathcal{I}_{(k)}),$$

which has the same null distribution under $\theta = 0$ as the original $S_{(k)}$.

The higher mean of $\gamma^{-1/2}\,S'_{(k)}$ for $\theta > 0$ increases power, as desired.

Applying the Cui et al. method

Sample size "re-estimation" occurs at analysis 2.

The objective is to deliver power at the new alternative $\theta = \delta/2$.

Information (i.e., numbers of events) planned from groups 3 to 5 is multiplied by a factor γ, the value of which depends on responses observed thus far.

We restrict γ to the range 1 to 6, implying:

- No reduction of information,
- Total information is at most 4 times the original maximum value.

Within this restriction, we endeavour to achieve a conditional power of 0.9 given current data under $\theta = \delta/2$.

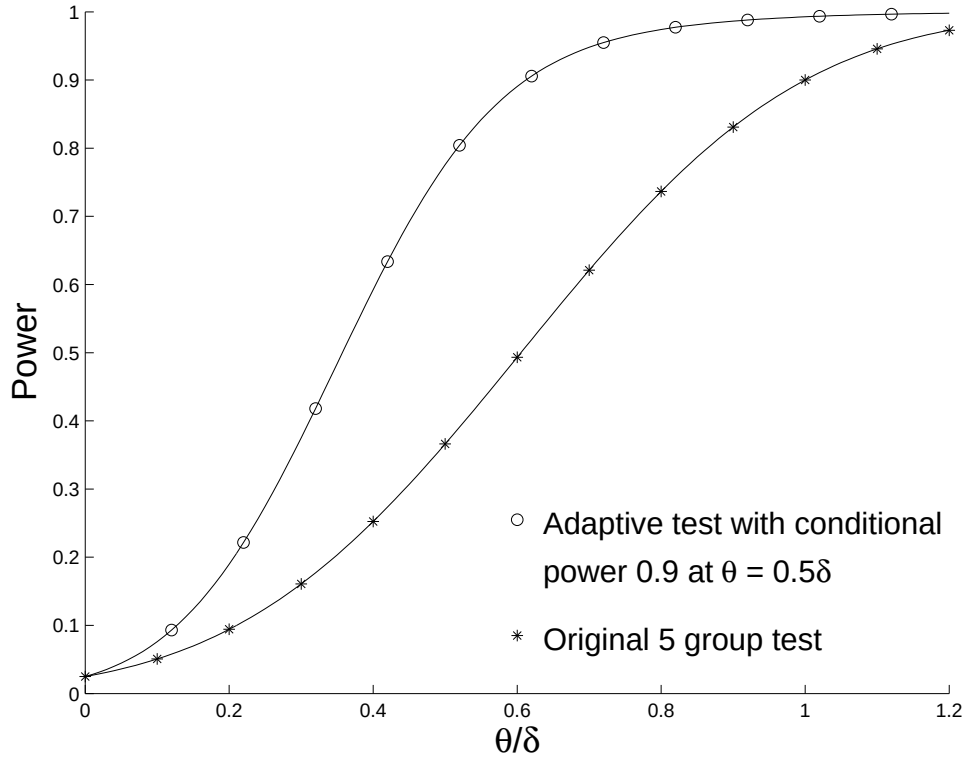
Conditional properties at the re-design point

Over the range of values for $Z(2)$ in the continuation region $(-0.42, 2.97)$, the re-designed test has the following features:

$\hat{\theta}/\delta$	$z(2)$	*Conditional type I error probability*	*Conditional power at $\theta = \delta/2$ before re-design*	γ	*Conditional power at $\theta = \delta/2$ after re-design*
1.40	2.94	0.5707	0.9100	1.00	0.9100
1.20	2.52	0.3856	0.8177	1.70	0.9000
1.00	2.10	0.2329	0.6903	2.65	0.9000
0.80	1.68	0.1272	0.5421	3.74	0.9000
0.60	1.26	0.0630	0.3917	5.00	0.9000
0.40	0.84	0.0279	0.2565	6.00	0.8762
0.20	0.42	0.0109	0.1490	6.00	0.7721
0.00	0.00	0.0036	0.0745	6.00	0.6204
−0.20	−0.42	0.0010	0.0310	6.00	0.4350

Results of re-design

The re-design is successful in raising the power curve at all effect sizes.



Overall power at $\theta = \delta/2$ is 0.78, below the 0.9 aimed for due to previous early stopping and truncation of γ to the range 1 to 6.

Is there a cost for learning power objectives late?

The adaptive approach has allowed investigators to respond to new information, even though the trial was already under way.

Has the delay in learning the power requirement had an impact on efficiency?

If the ultimate objective of power 0.9 at $\theta = \delta/2$ had been known when the study was first planned, a suitable group sequential design could have been chosen.

An alternative group sequential design

We consider an error spending test which matches features of the adaptive test's power curve and expected sample size function. This is

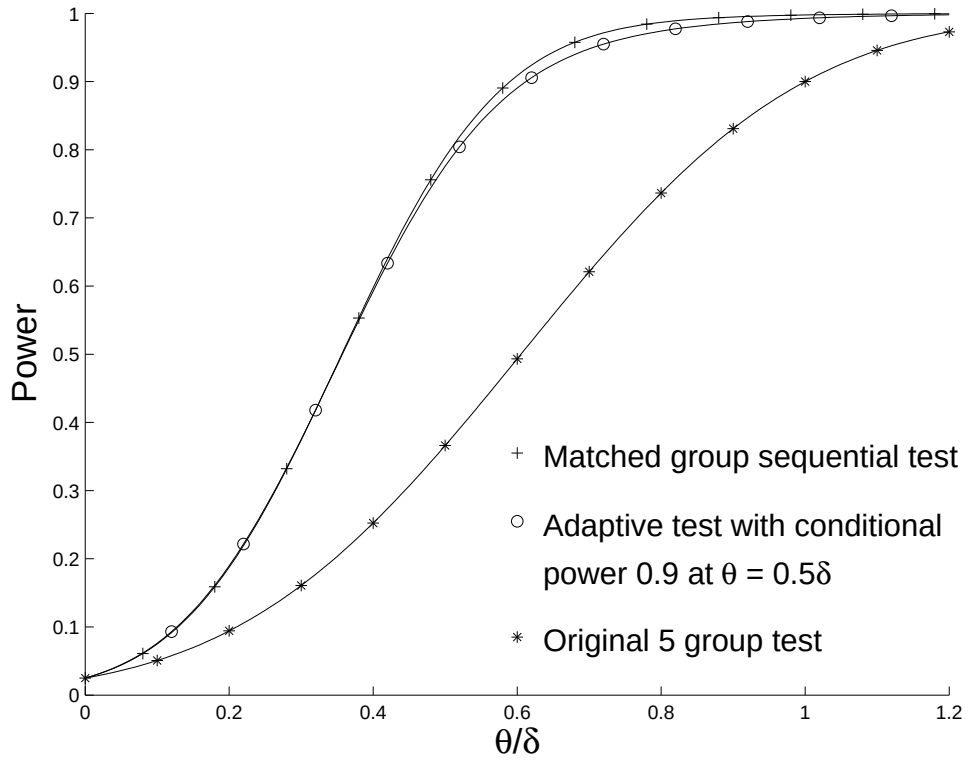
A design which spends error $\propto \mathcal{I}^{0.75}$ with power 0.9 at $\theta = 0.59\,\delta$,

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

$\mathcal{I}_{\max} = 3.78\,\mathcal{I}_f$ (compared to $4.2\,\mathcal{I}_f$ for the adaptive design).

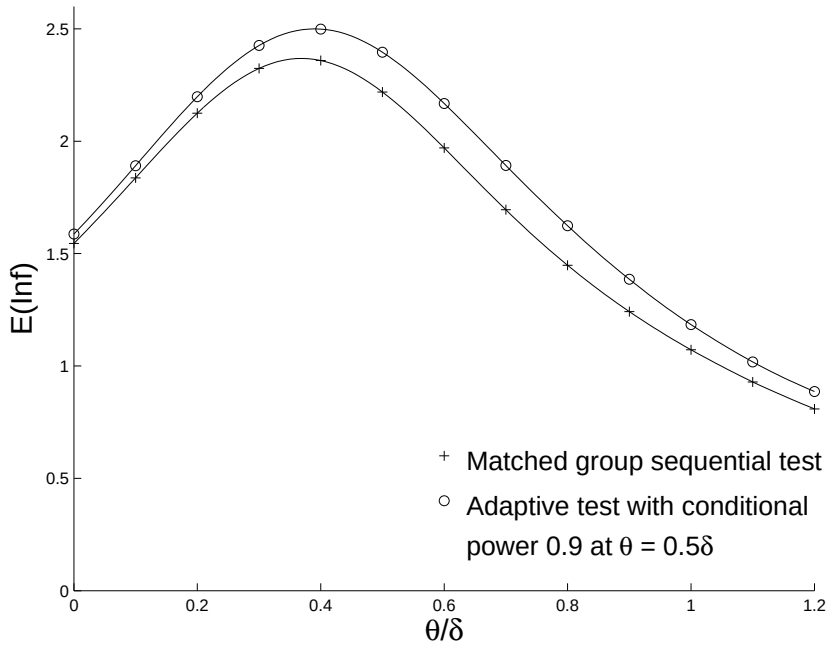
A matched group sequential design

Power of the “matched” group sequential design is as high as that of the adaptive design at all effect sizes.



A matched group sequential design

The group sequential design has lower expected information than the adaptive design at all effect sizes.



For survival data, information translates to “number of observed events”.

The lower $\mathcal{I}_{\max}$ for the group sequential design should imply lower recruitment.

Comparing efficiencies of sequential designs

Suppose test A has

Type I error probability α and

Power $1 - b_A(\theta)$ and expected information $E_{A,\theta}(\mathcal{I})$ at effect size θ.

A level α fixed sample test needs $\mathcal{I} = (z_\alpha + z_{b_A(\theta)})^2/\theta^2$ to achieve this power.

Hence, we define the Efficiency Index of test A at effect size θ to be

$$EI_A(\theta) \;=\; \frac{(z_\alpha + z_{b_A(\theta)})^2}{\theta^2}\,\frac{1}{E_{A,\theta}(\mathcal{I})}.$$

We use this index to define the Efficiency Ratio at θ between tests A and B as

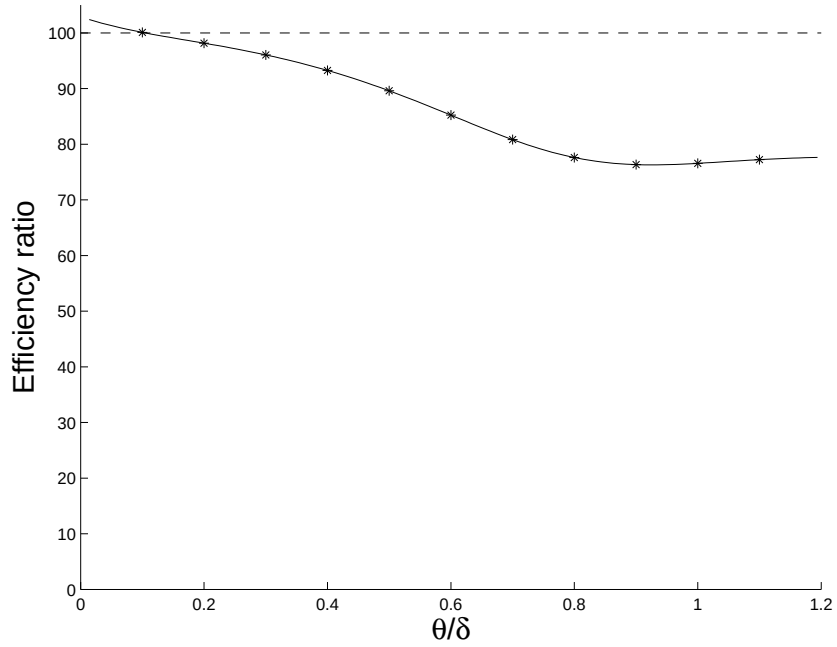
$$ER_{A,B}(\theta) \;=\; \frac{EI_A(\theta)}{EI_B(\theta)} \times 100 \;=\; \frac{E_{B,\theta}(\mathcal{I})\,(z_\alpha + z_{b_A(\theta)})^2}{E_{A,\theta}(\mathcal{I})\,(z_\alpha + z_{b_B(\theta)})^2} \times 100.$$

Comparing efficiencies of sequential designs

The Efficiency Ratio combines information on attained power and expected information at each effect size θ.

The cost of delay in learning the real power requirement is seen to be an efficiency loss of 20% at higher effect sizes.

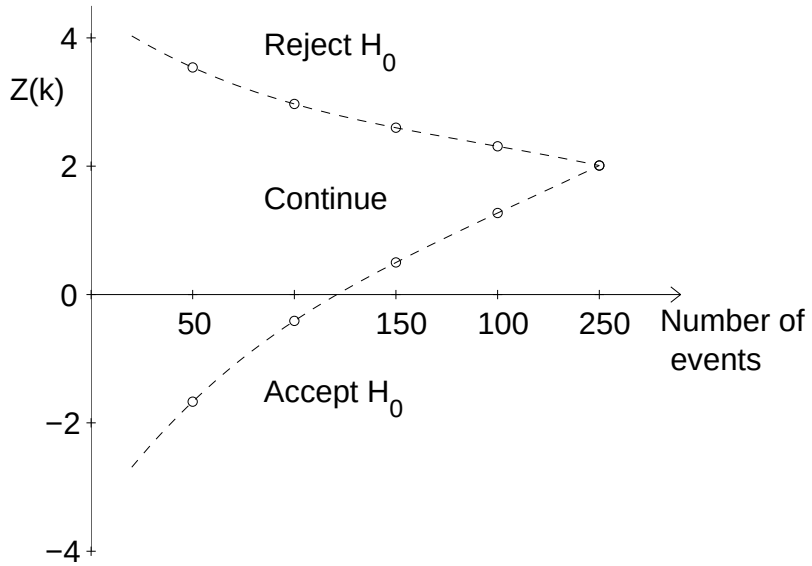
Efficiency ratio of adaptive design vs group sequential test



Changing power in response to internal events

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

We start with the same initial group sequential design as in Example 1.



Suppose investigators wish to re-design the remainder of the trial in response to the interim estimate of effect size, $\hat{\theta}_2$, at analysis 2.

Increasing later group sizes

A lower than anticipated interim estimate $\hat{\theta}_2$ prompts investigators to consider the trial's power at effect sizes below δ, where power 0.9 was originally set:

Lower effect sizes start to appear plausible,

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

Applying the Cui et al. method:

Numbers of events in groups 3 to 5 are multiplied by a factor γ, and increments in the score statistic from these groups are down-weighted by $\gamma^{-1/2}$ to maintain the type I error rate.

The value of the factor γ is chosen so that conditional power is 0.9, given current data, **if $\boldsymbol{\theta}$ is equal to $\hat{\boldsymbol{\theta}}_{\mathbf{2}}$**.

A decrease ($\gamma < 1$) is allowed but an upper limit $\gamma = 6$ is imposed, restricting the number of events to at most 4 times the original maximum.

Conditional properties at the re-design point

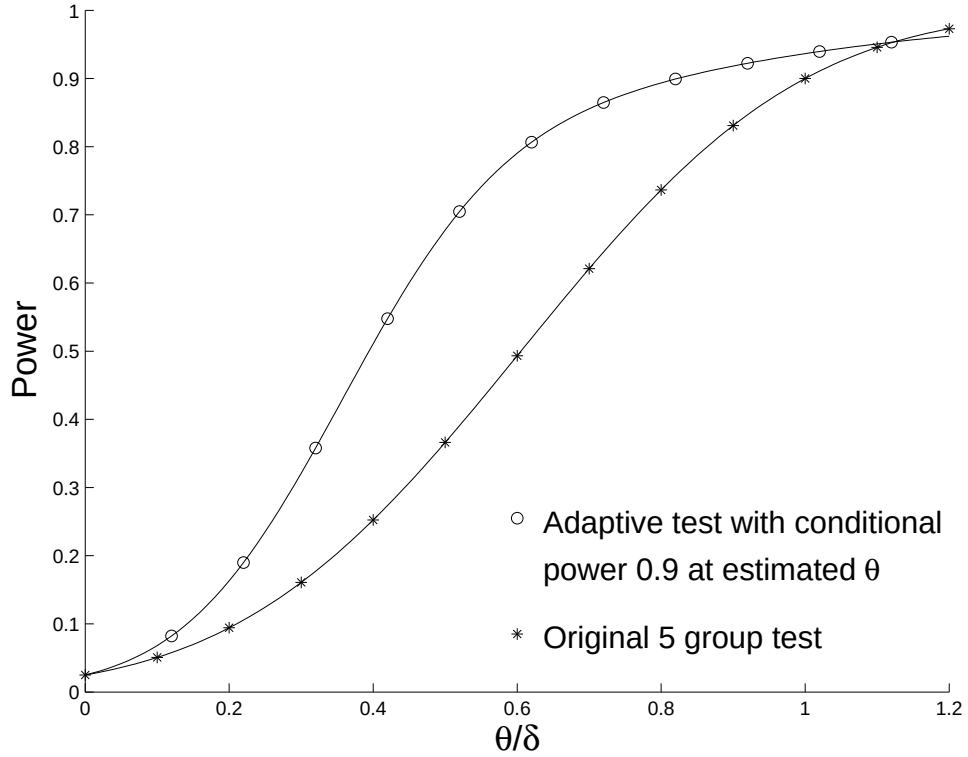
Over the range of values for $Z(2)$ in the continuation region $(-0.42, 2.97)$, the re-designed test has the following features:

$\hat{\theta}/\delta$	$z(2)$	*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	0.0365	6.00	0.1432
0.00	0.00	0.0036	0.0036	6.00	0.0036
−0.20	−0.42	0.0010	0.0002	6.00	0.0000

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

Results of re-design

Re-design has raised the power curve at all effect sizes.



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 purely from observing $\hat{\theta}_2$. A group sequential design 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 respond 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 were acceptable, how else might it have been achieved?

An alternative group sequential design

A design matching key features of the adaptive test is

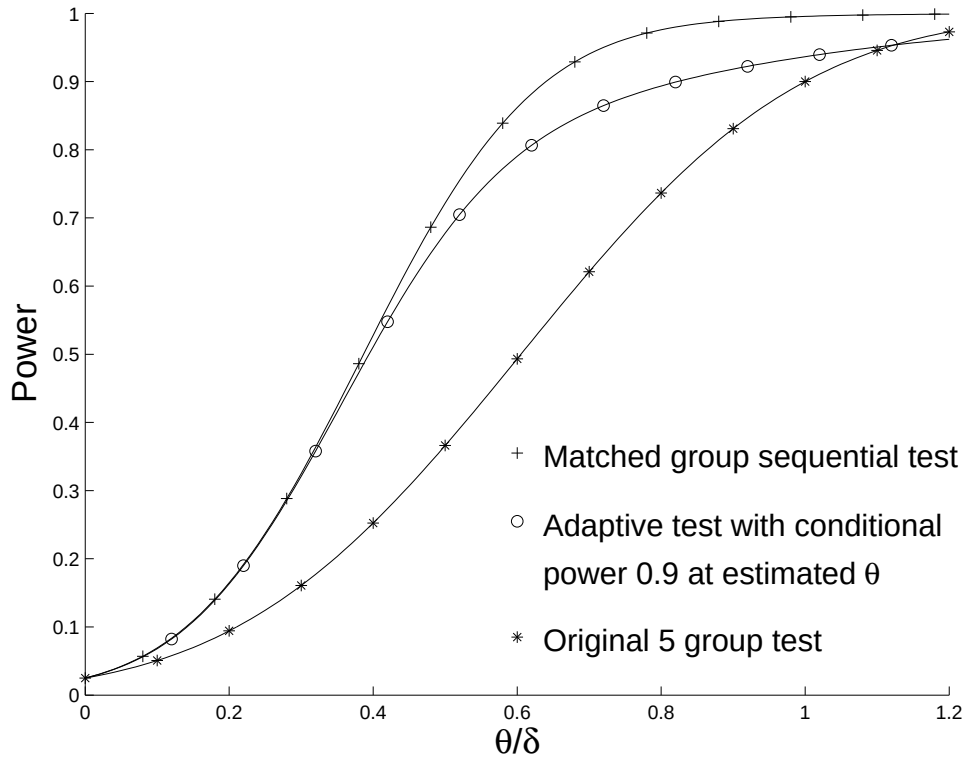
A design which spends error $\propto \mathcal{I}^{0.75}$ with power 0.9 at $\theta = 0.64\,\delta$,

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

$\mathcal{I}_{\max} = 3.21\,\mathcal{I}_f$ (compared to $4.2\,\mathcal{I}_f$ for the adaptive design).

A matched group sequential design

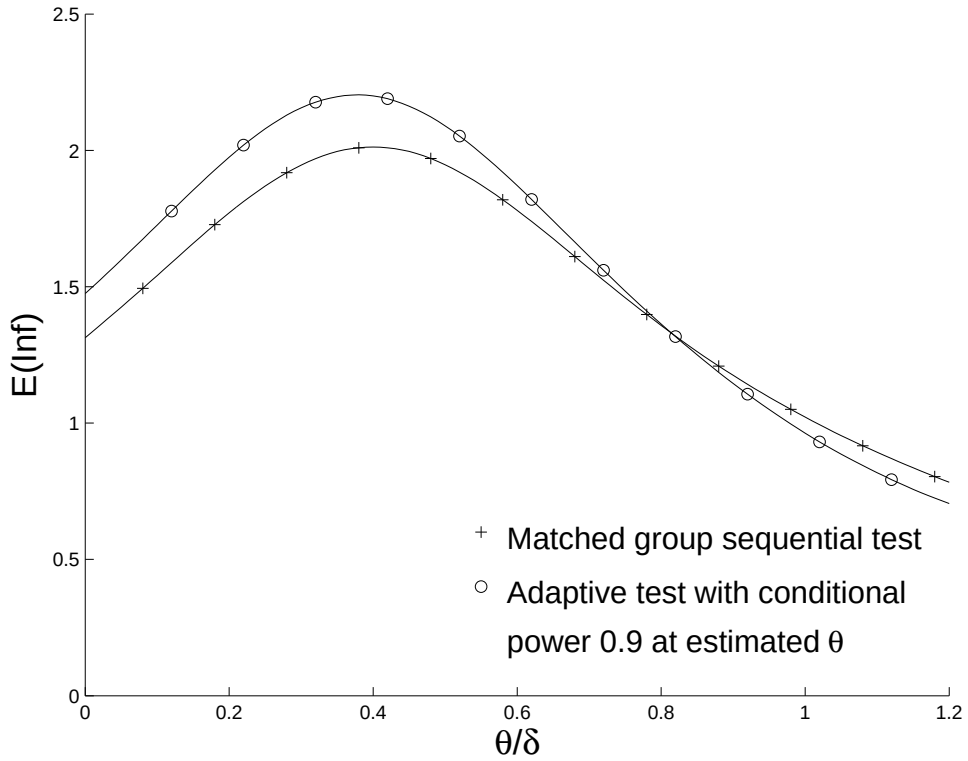
Power of the “matched” group sequential design is as high as that of the adaptive design at all effect sizes — and substantially higher at the largest θ values.



A matched group sequential design

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

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

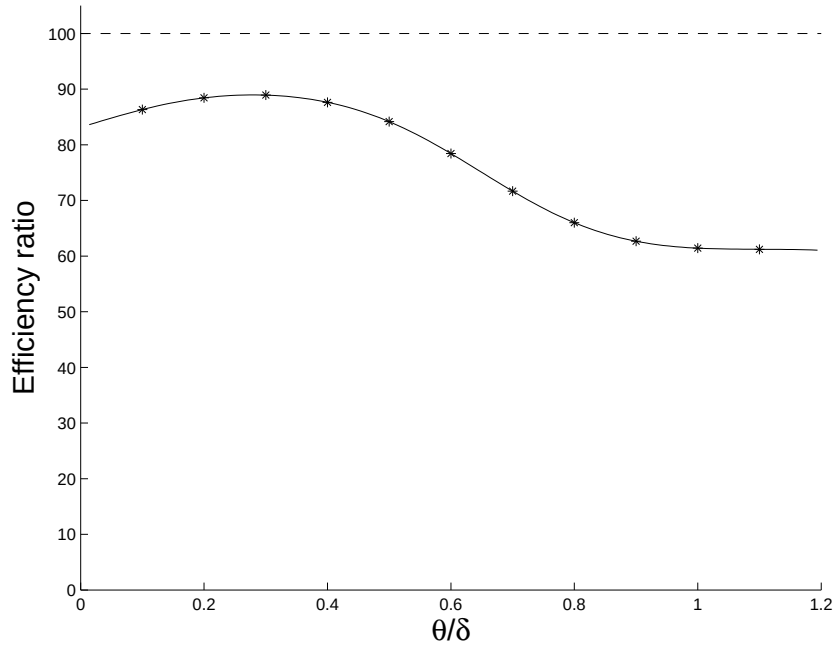


Efficiency ratio

We can use the Efficiency Ratio to combine information on attained power and expected information.

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



Recapitulation: Sample size re-estimation

(i) Nuisance parameters

There is a variety of methods to re-estimate the sample size needed to meet a specific power requirement under a given type I error probability.

(ii) In response to external information

It is good that we have adaptive methods that can do this when necessary.

But, earlier knowledge of the ultimate objective would be preferable.

(iii) In response to internal information

Adaptive methods can “rescue” an under-powered study.

Our example shows this can produce a poor design, with high average “sample size” for the power achieved: a standard group sequential design is preferable.

We have found this conclusion to hold quite generally.

Sample size adaptation in response to internal information

Just as in Example 2 above, we have found inefficiencies in a variety of proposed adaptive designs, including:

Bauer and Köhne (*Biometrics*, 1994)

Proschan and Hunsberger (*Biometrics*, 1995),

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

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

When adaptation makes smaller increases in sample size, the gain in power is smaller but efficiency loss is still present.

In general, adaptive designs have more freedom than group sequential tests since they can vary the next group size in response to current data.

Hence, the best adaptive designs *ought* to be superior to group sequential designs — so why do adaptive tests in the literature fare so poorly?

"Schmitz" designs

Adaptive group sequential designs (Schmitz, 1993)

These designs have stopping rules, just like group sequential tests, but they also have rules for choosing the next group size — or number of events — or increment in information — in response to current data.

Optimal group sequential tests

We can compute the test which minimises expected sample size averaged over a set of θ values, for given type I error and power and a fixed sequence of group sizes.

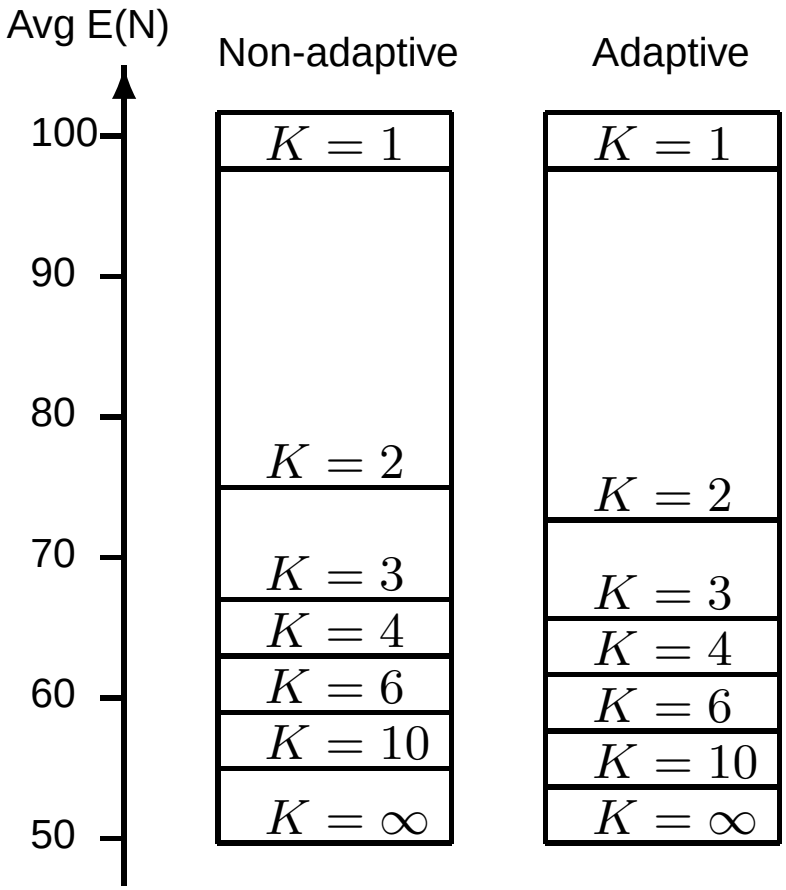
Optimal adaptive group sequential tests

We can also find the optimal adaptive, group sequential test with given type I error and power and a fixed number of groups with data-dependent sample sizes.

Since the class of adaptive tests includes non-adaptive tests as a special case, the optimal adaptive test is more efficient than the optimal group sequential test.

Sample size adaptation in response to internal information

Minimum sample sizes for adaptive and non-adaptive designs with K analyses.



Advantages of adaptive designs are small — but they are present.

Sources of inefficiency in flexible, adaptive designs

1. Use of non-sufficient statistics

Jennison and Turnbull (*Biometrika*, 2006) prove all admissible designs (adaptive or non-adaptive) are Bayes procedures. Hence, their decision rules and sample size rules must be functions of sufficient statistics.

Unequal weighting of observations in adaptive designs means these are not based on sufficient statistics. Thus, they cannot be optimal designs for any criteria.

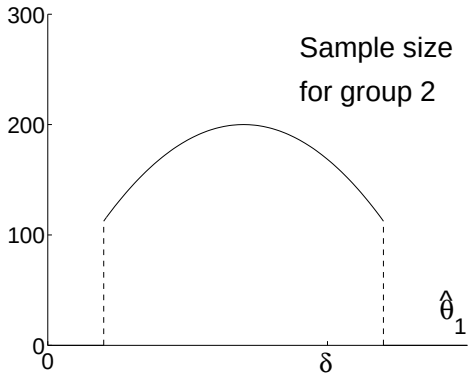
The potential benefits of adaptivity are slight and any departure from optimality can leave room for an efficient non-adaptive design, with the same number of analyses, to do better.

NB, this is stronger conclusion than that of Tsiatis and Mehta (*Biometrika*, 2003) who allow the comparator non-adaptive design to have additional analyses.

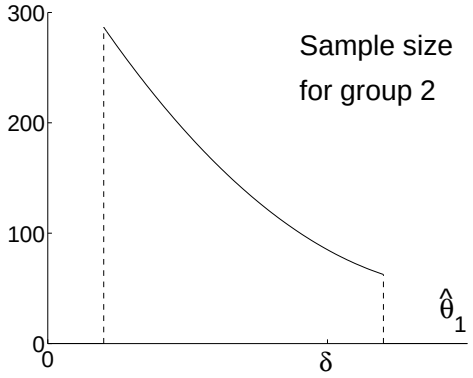
Sources of inefficiency in flexible, adaptive designs

2. Sub-optimal sample size modification rule

Typical sample size function for an optimal adaptive test



Typical sample size function for a conditional power adaptive design



“Conditional power” sample size modification rules differ qualitatively from those of optimal adaptive designs.

Final conclusions

(i) Nuisance parameters

There is a variety of methods to re-estimate the sample size needed to meet a specific power requirement under a given type I error probability.

(ii) In response to external information

It is good to have adaptive methods to do this when necessary.

But, it is preferable to know the ultimate objective at the outset.

(iii) In response to internal information

Adaptive methods can “rescue” an under-powered study.

There is an efficiency cost to such a rescue: it is much better to design the study with the correct power in the first place.

We do *not* recommend using this re-design feature to avoid tackling difficult questions about power at the design stage.