Handling uncertainty about the likely treatment effect:

The roles of group sequential designs

and adaptive sample size re-estimation

Christopher Jennison

Department of Mathematical Sciences,

University of Bath, UK

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

ENAR 2007 Annual Meeting,

Atlanta, March 2007

Plan of talk

1. Survey of advice on ***how to specify delta***

2. Two strategies to keep sample size low:

   Start small, then increase sample size if necessary

   Plan for the worst case, then stop early if possible

3. Relation to adaptive and group sequential designs

4. Critical appraisal and comparison of approaches

5. Conclusions

Reference:

"Efficient group sequential designs when there are several effect sizes under consideration", Jennison and Turnbull, *Statistics in Medicine*, 2006.

Advice on “how to specify delta”

When a trial is designed, uncertainty about the true effect size is to be expected.

After all, the purpose of the study is to investigate this treatment effect.

Thus, it is often difficult to settle on an effect size at which to set a study’s power.

Example:

Comparing treatments A and B, we will observe responses

$$X_{Ai} \sim N(\mu_A,\, \sigma^2) \quad \text{and} \quad X_{Bi} \sim N(\mu_B,\, \sigma^2).$$

The “effect size” for the new treatment, A, is $\theta = \mu_A - \mu_B$.

It is desired to test H_0: $\theta \leq 0$ against $\theta > 0$ with type I error probability $\alpha = 0.025$ when $\theta = 0$ and power 0.9 at $\theta = \delta$, i.e.,

$$P_{\theta=\delta}\{\text{Reject } H_0\} \,=\, 1 - \beta \,=\, 0.9.$$

How should δ be chosen?

Specifying delta for the power requirement

The following sources provide a variety of advice on where a power requirement should be set:

Piantadosi (*Clinical trials: a methodological perspective*, 1997)

Pocock (*Clinical trials: a practical approach*, 1997)

Senn (*Statistical issues in drug development*, 1997)

FDA E9: Statistical principles for clinical trials, (*Federal Register*, 1998)

Shun et al. (*Statistics in Medicine*, 2001)

— and there are plenty more.

Specifying delta

Suggestions on where to set power include:

(A) The minimum clinically relevant or commercially viable effect — an effect that is "important to detect" (Pocock, p. 125 and 132, Senn, Piantadosi),

(B) The anticipated effect size (Shun et al., 2001),

(C) Either (A) or (B) above (ICH Guidance E9, Section 3.5),

(D) A "realistic" value (Pocock p. 128),

(E) A skeptic's value,

(F) An optimist's value,

(G) The true value (implicit in some proposals for adaptive designs).

What do you usually do?

Specifying delta

Proper consideration may lead to a sample size simply too high to be feasible.

You are not supposed to admit to using the rule:

(Z) *Work back from the sample size you can afford*

— but a pragmatic approach is sometimes unavoidable.

If the sample size for power at the effect size you wish to detect is very high, perhaps you can aim for this power but hope to manage this with fewer observations:

If the true effect size is higher than this minimum effect size,

If early data provide evidence for a conclusion (in either direction).

More on our example

The effect size is $\theta = \mu_A - \mu_B$.

Type I error probability:

$$P_{\theta=0}\{\text{Reject } H_0\} = \alpha = 0.025.$$

Power at effect size δ:

$$P_{\theta=\delta}\{\text{Reject } H_0\} = 1 - \beta = 0.9.$$

Suppose that setting power $= 0.9$ at $\theta = 20$ requires a sample size of 100 subjects per treatment in a fixed sample design.

It would be preferable to set power 0.9 at the lower effect size of $\theta = 14$ or 15, but the necessary sample sizes of 178 and 204 per treatment are unpalatable.

Two strategies

1. Start small, then ask for more

Design for the larger effect size $\theta = 20$ so the initial sample size is small.

At an interim analysis, consider the current estimate of θ and the likely outcome of the trial. If appropriate, seek permission to increase the sample size.

To incorporate such an "adaptive" increase in sample size, special methods must be used to protect the type I error rate.

2. Start large, then try to stop early

Choose a group sequential design with power at a smaller effect size $\theta = 14$, say.

The stopping boundary will allow an early decision when results favour the new treatment strongly — which should be the case when θ is as large as 20.

If results are disappointing, a lower "futility" boundary will permit early stopping to accept the null hypothesis.

Strategy 1: Adaptively increasing the sample size

Bauer & Köhne (*Biometrics*, 1994) proposed mid-course design changes to one or more of

Treatment definition

Choice of primary response variable

Sample size:

- to maintain power under an estimated nuisance parameter
- to change power in response to external information
- to change power for internal reasons

    a) secondary endpoint, e.g., safety

    b) primary endpoint, using interim estimate $\hat{\theta}$.

Bauer & Köhne's two-stage scheme

Investigators decide *at the design stage* to split the trial into two parts. Each part yields a one-sided P-value and these are combined at the end.

- Run part 1 as planned. This gives

$$P_1 \sim U(0, 1) \quad \text{under } H_0.$$

- Make design changes.

- Run part 2 with these changes, giving

$$P_2 \sim U(0, 1) \quad \text{under } H_0,$$

  conditionally on P_1 and other part 1 information.

- Combine P_1 and P_2 by Fisher's combination test:

$$-\log(P_1\, P_2) \sim \frac{1}{2}\,\chi^2_4 \quad \text{under } H_0.$$

A "start small and ask for more" strategy

Using Bauer & Köhne's two-stage scheme in our example:

1. Plan for a sample size of 100 per treatment to attain power 0.9 at $\theta = 20$.

   Run the study in 2 stages with $n_1 = 50$ observations per treatment in Stage 1.

2. Stage 1 produces estimated effect $\hat{\theta}_1$ and one-sided P-value P_1 for H_0: $\theta = 0$.

3. The sample size planned for Stage 2 is 50 per treatment — but this may be modified on seeing $\hat{\theta}_1$.

   E.g., if $\hat{\theta}_1 = 15$, investigators may seek to increase the Stage 2 sample size to ensure high conditional power if the true effect size is indeed equal to 15.

4. Stage 2 produces P-value P_2 for H_0: $\theta = 0$.

5. Overall, H_0 is rejected if $P_1 P_2 < k_\alpha$, where k_α is the lower α point of the null distribution for $P_1 P_2$, namely $e^{-(1/2)\chi^2_4}$.

Stopping at Stage 1 and a sample size rule for Stage 2

If $\hat{\theta}_1 \leq 0$, *Stop at Stage 1 and accept H_0.*

With this stopping for futility, we H_0 reject at Stage 2 if $P_1 P_2 < k'_\alpha = 0.00435$.

Thus, if $\hat{\theta}_1 > 22.9$, which implies $P_1 < 0.00435$ and hence $P_1 P_2 < 0.00435$,

Stop at Stage 1 and reject H_0.

Otherwise,

Find the sample size n_2 per treatment such that conditional power given P_1 will be 0.9 if the true effect size is equal to $\hat{\theta}_1$,

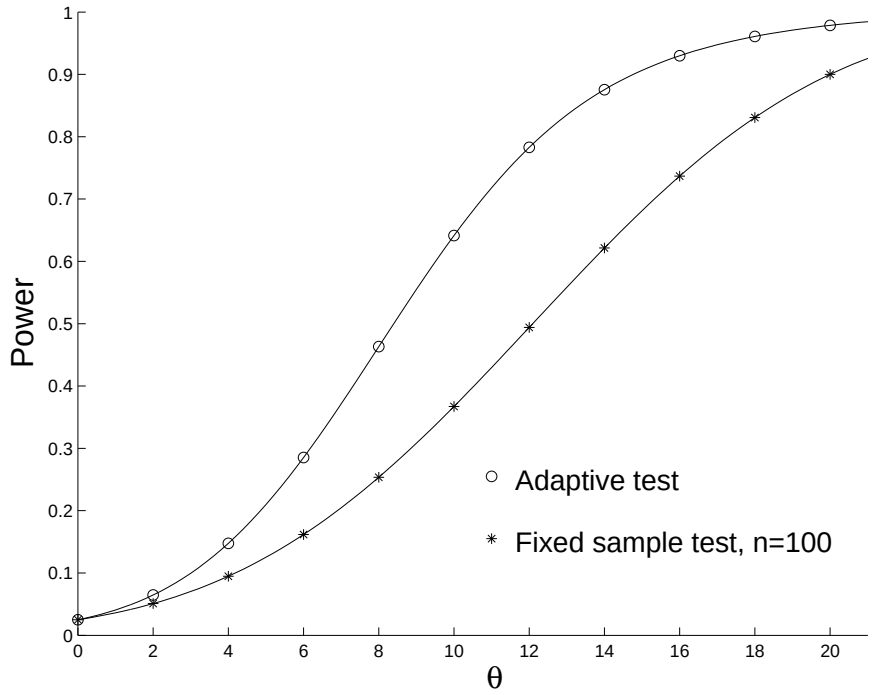
If $n_2 < 50$, increase it to 50; if $n_2 > 250$, decrease it to 250,

Run Stage 2 with n_2 observations per treatment.

After Stage 2,

Reject H_0 if $P_1 P_2 < k'_\alpha = 0.00435$.

Overall power for “start small, ask for more” procedure

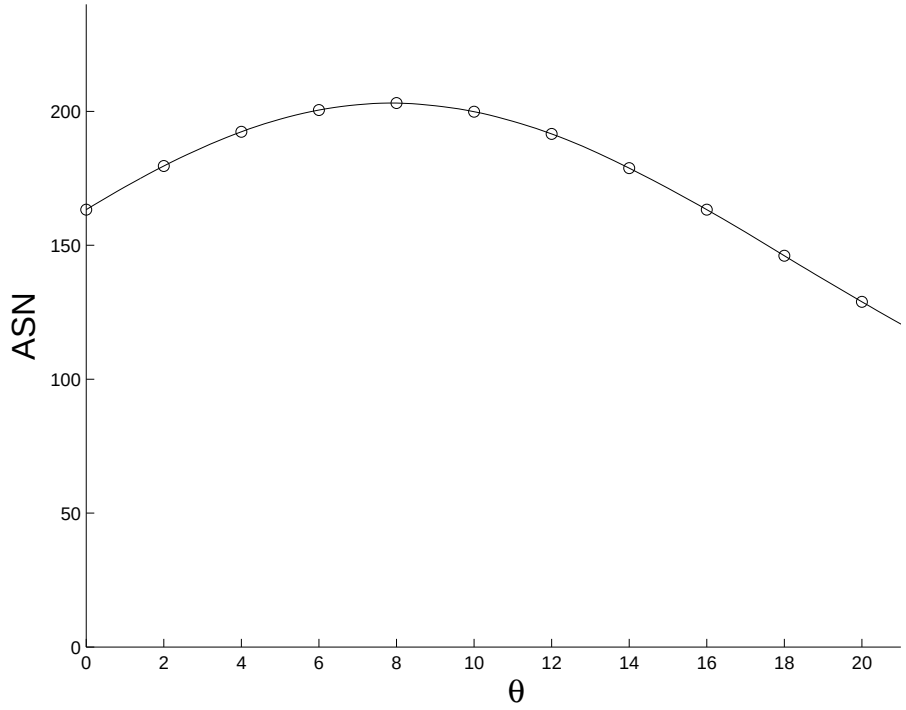


Power improves on that of a fixed sample test with 100 observations per treatment.

The adaptive design has power ≈ 0.9 for $\theta = 15$.

Values of n_2 are: 250 *for* $\hat{\theta}_1 < 11.5$, *then decreasing to* 50 *for* $\hat{\theta}_1 > 18.0$.

Average sample size for “start small, ask for more” procedure



Average sample size per treatment ranges between 130 and 200, depending on the true effect size, θ.

When proceeding to Stage 2, total sample size is from 100 to 300 per treatment.

Strategy 2: A Group Sequential Design

Recall, we wish to test H_0: $\theta \leq 0$ against $\theta > 0$ with:

Type I error probability $\alpha = 0.025$ when $\theta = 0$,

Power 0.9 at $\theta = \delta$.

In a Group Sequential Test, observations are taken in groups and a decision is taken after each group to:

Stop, reject H_0: $\theta \leq 0$ in favour of $\theta > 0$,

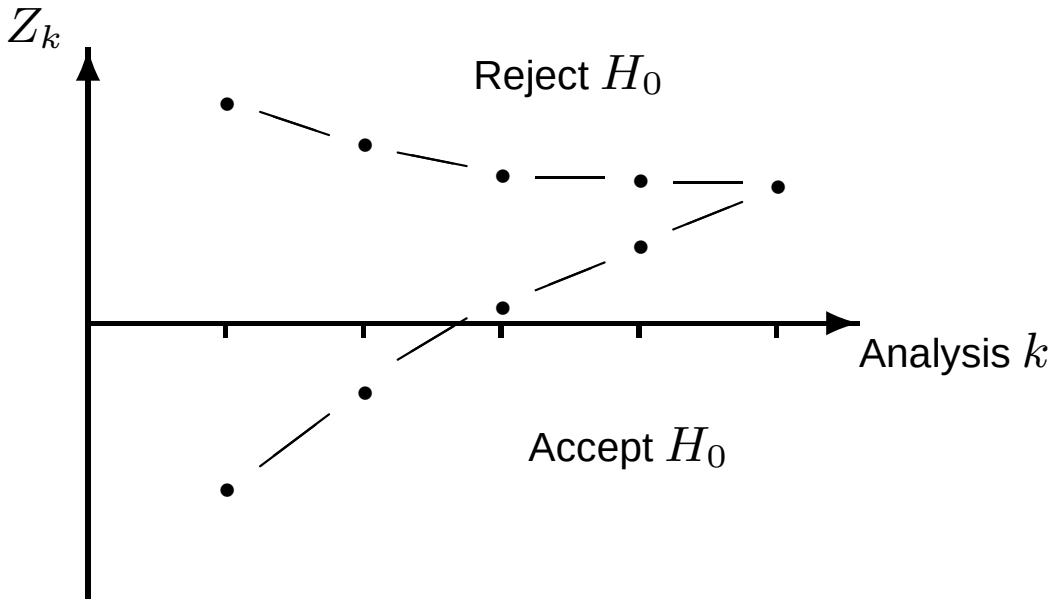
Continue to observe the next group of subjects, or

Stop, accept H_0.

Typically, group sizes are pre-specified. However, “error spending” designs are able to deal with unpredictable group sizes.

Group Sequential Tests

A group sequential test can be defined by a pair of boundaries for the sequence of Z-statistics.



Expected sample size can be around 50% to 70% of the fixed sample size with the same type I error rate and power.

Reference: "*Group Sequential Methods with Applications to Clinical Trials*", Jennison & Turnbull, 2000.

Group Sequential Tests

For our example, we wish to test H_0: $\theta \leq 0$ against $\theta > 0$, where $\theta = \mu_A - \mu_B$ is the effect size of a new treatment vs control.

The design must have

Type I error probability:

$$P_{\theta=0}\{\text{Reject } H_0\} = \alpha = 0.025,$$

Power at effect size δ:

$$P_{\theta=\delta}\{\text{Reject } H_0\} = 1 - \beta = 0.9.$$

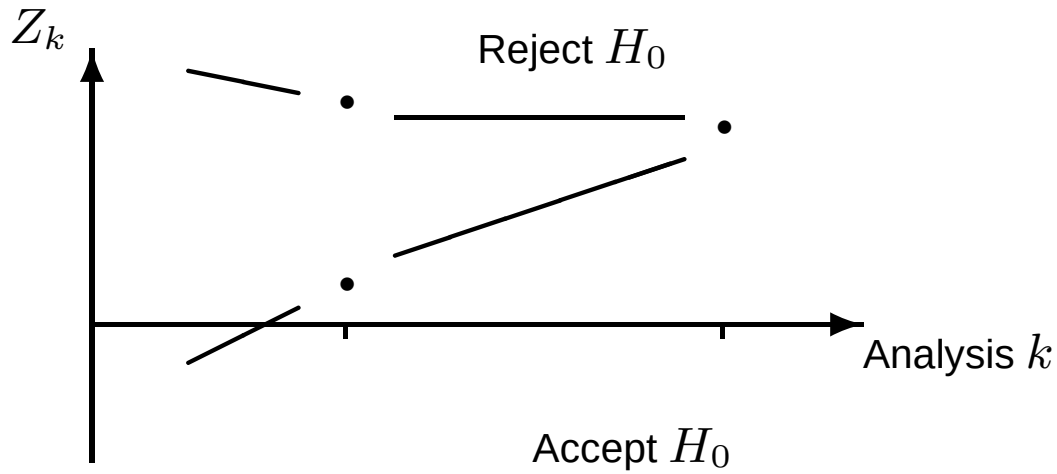
We can set up the test with power 0.9 at $\theta = 14$, knowing the study is likely to stop early if θ is much higher than this.

For comparability with the previous adaptive design, we consider a Group Sequential Test with just two groups.

Group Sequential Tests

I have chosen an error spending design, where both type I and type II error probabilities are spent linearly in sample size.

The maximum sample size is 229 observations per treatment and the first group contains 40% of these, i.e., 92 observations per treatment.

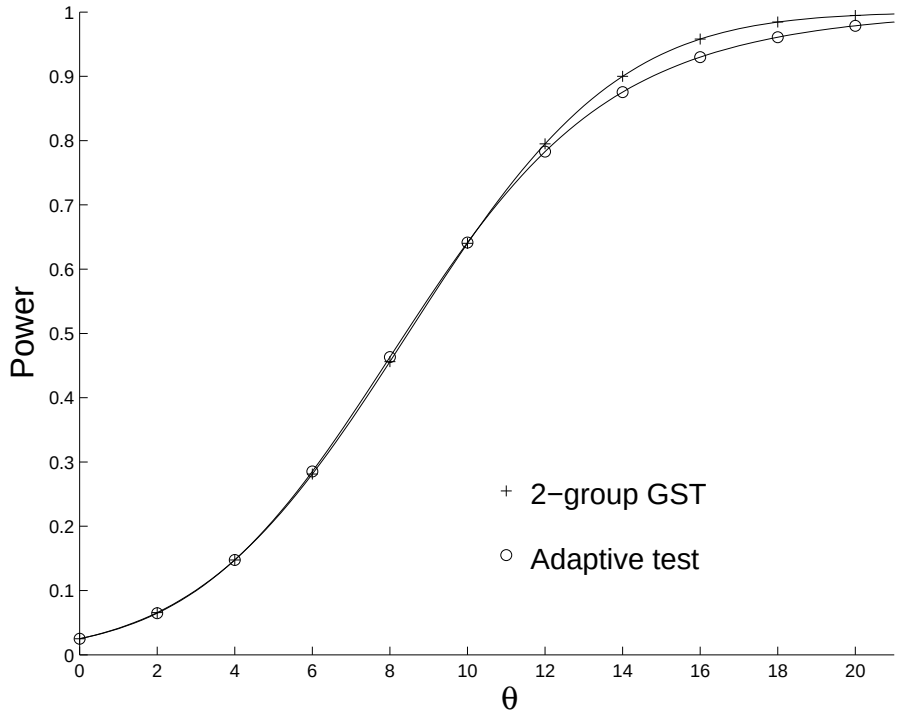


A fixed sample test with this power would require 204 observations per treatment.

Boundaries are: *for* Z_1, 0.42 *and* 2.33*; for* Z_2, 2.06*,*

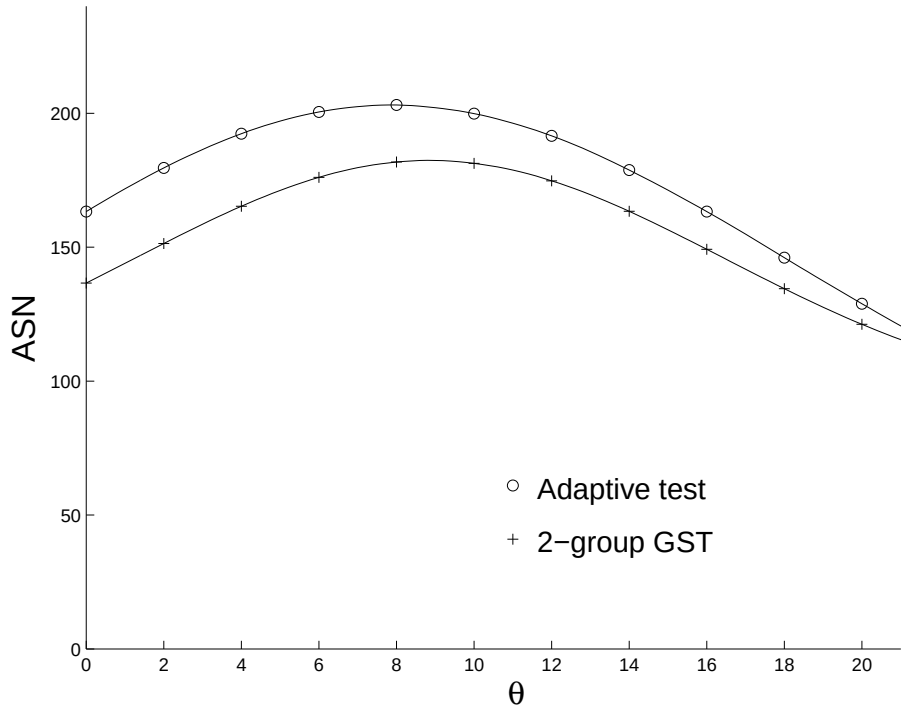
for $\hat{\theta}_1$, 2.7 *and* 14.7*; for* $\hat{\theta}_2$, 8.4*.*

Overall power: Group Sequential Test and Adaptive Design



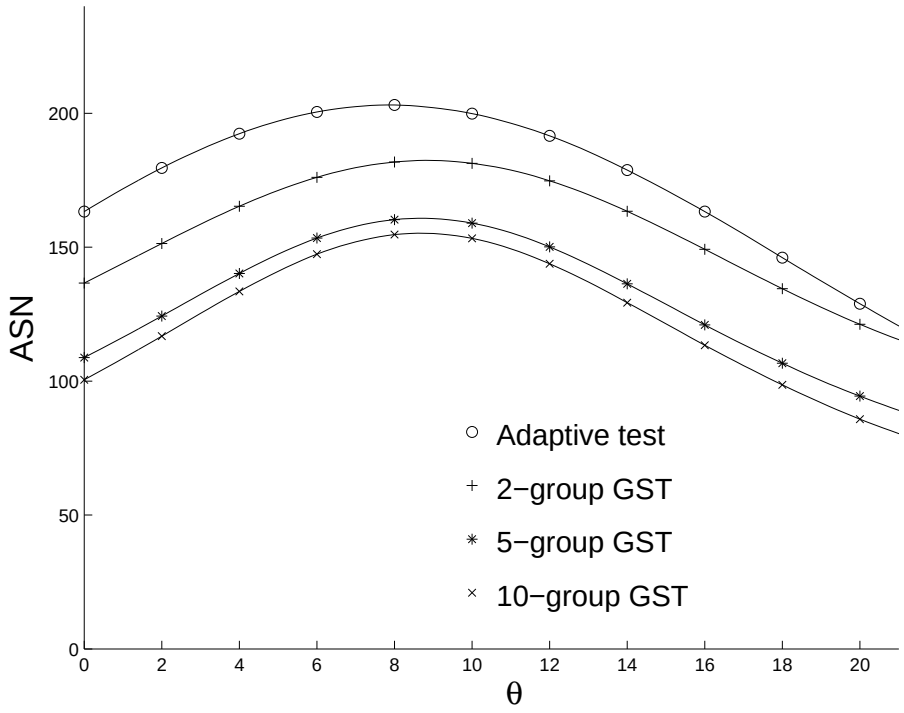
The power curve for the Group Sequential Test matches that of the Adaptive Design at lower effect sizes and exceeds it at higher ones.

Average Sample Size: Group Sequential and Adaptive Designs



The Group Sequential Test has lower expected sample size than the Adaptive Design by up to 15%, depending on the effect size θ — and remember the GST has greater power.

Average Sample Size: Group Sequential and Adaptive Designs



Adding further groups improves the efficiency of the Group Sequential Test.

Discussion

The example shows use of adaptive methods to facilitate revision of sample size in response to interim estimates of effect size is not an efficient strategy.

Other examples lead to similar conclusions.

However, adaptive methods have many other uses and these deserve attention.

One can formulate adaptive re-design methods and group sequential tests in very similar ways:

- Both allow stopping at an interim analysis to accept or reject H_0,
- Adaptive methods have the extra flexibility to specify future group sizes in response to observed data.

Why, then, do adaptive designs appear so inefficient?

Causes of inefficiency in adaptive designs

Reference: "Adaptive and nonadaptive group sequential tests",

Jennison and Turnbull, *Biometrika*, 2006.

1. Over-reliance on an interim estimate of effect size

In the adaptive method, we chose Stage 2 sample size to provide conditional power 0.9 assuming $\theta = \hat{\theta}_1$.

After 50 observations per treatment, the standard deviation of $\hat{\theta}_1$ is 8.7,

a 95% confidence interval for θ at this point would have width 34.2.

Yet, we are using $\hat{\theta}_1$ to make judgements of

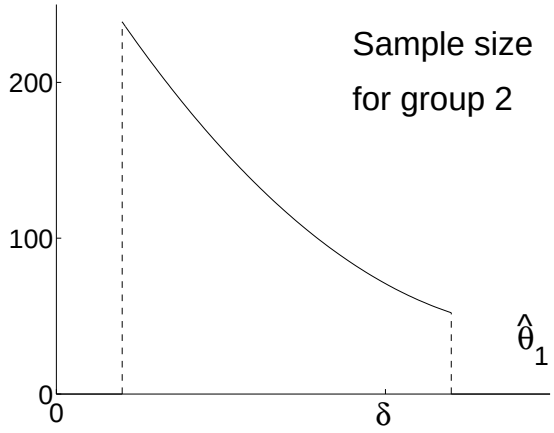
Stage 2 sample size of 250 for $\hat{\theta}_1 = 11$, vs

Stage 2 sample size of 50 for $\hat{\theta}_1 = 18.5$.

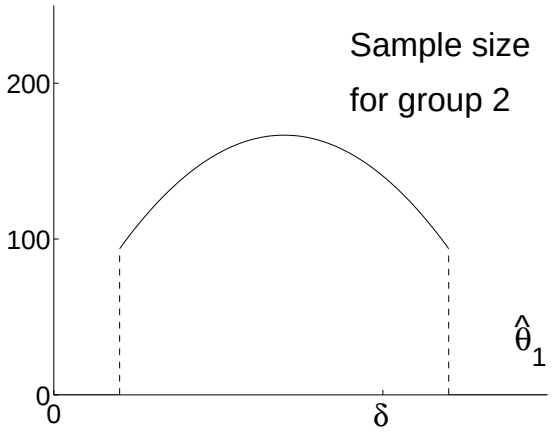
Causes of inefficiency in adaptive designs

2. Wrong shape for the sample size function

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



Typical sample size function for a conditional power adaptive design



Typical sample size function for an optimal adaptive test

The positive view

Group Sequential Tests:

Are efficient,

Are widely applicable,

Are well understood,

Are well supported by software and texts,

Provide an effective method to keep sample size in check while achieving power at the appropriate alternative.