Sample Size Re-estimation in Clinical Trials

Christopher Jennison

Department of Mathematical Sciences,

University of Bath, UK

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

PSI One Day Meeting:

Sample size re-estimation — dealing with those unknowns

London, 2 November 2016

Outline of talk

1. Sample size formulae — what is it that we don't know?

   Normal, binary and survival endpoints

2. Dealing with nuisance parameters

   Normal: The variance,

   Binary: Probability of success on the control arm,

   Survival: Overall hazard rate

3. The unknown treatment effect

   Is there a problem?

   Group sequential or adaptive designs?

4. Conclusions

1. Sample size formulae

(i) A two treatment comparison with normal response

Consider a Phase 3 trial comparing a new treatment with a control, where the primary endpoint follows a normal distribution.

Denote responses by

$Y_{Bi}, \ i = 1, 2, \ldots,$ on the new treatment,

$Y_{Ai}, \ i = 1, 2, \ldots,$ on the control arm.

A common variance σ^2 is assumed for both treatment and control, so we have

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

The treatment effect is

$$\theta = \mu_B - \mu_A.$$

A two treatment comparison with normal response

We wish to test H_0: $\theta \leq 0$ against the alternative $\theta > 0$.

We set the (one-sided) type I error rate to be $\alpha = 0.025$.

We specify an effect size δ we wish to detect with high probability.

In order to achieve power $1 - \beta$ when $\theta = \delta$, we need a sample size of approximately

$$n = \frac{2\left(z_\alpha + z_\beta\right)^2 \sigma^2}{\delta^2} \tag{1}$$

in both the treatment and control groups. (A more precise answer uses a t-distribution instead of the standard normal distribution.)

The sample size formula (1) depends crucially on σ^2 and δ.

Practical considerations will favour choices of σ^2 and δ that lead to an affordable trial with a feasible target sample size.

(ii) A two treatment comparison with binary response

Consider a trial with a binary outcome, e.g., success or failure of the treatment.

Denote responses by

$$Y_{Bi}, \ \ i = 1, 2, \ldots, \ \text{ on the new treatment,}$$

$$Y_{Ai}, \ \ i = 1, 2, \ldots, \ \text{ on the control arm,}$$

and success probabilities by p_A and p_B, so

$$Y_{Bi} = 1 \quad \text{with probability } p_B,$$
$$Y_{Ai} = 1 \quad \text{with probability } p_A.$$

The treatment effect is

$$\theta = p_B - p_A$$

and we wish to test H_0: $\theta \leq 0$ against $\theta > 0$.

A two treatment comparison with binary response

The success probabilities of treatments A and B are p_A and p_B.

With $\theta = p_B - p_A$, we wish to test H_0: $\theta \leq 0$ against $\theta > 0$.

The (one-sided) type I error rate is $\alpha = 0.025$ and we aim to achieve power $1 - \beta$ at a specified effect size $\theta = \delta$.

To achieve this power, we need a sample size in each treatment group of

$$n = \frac{2\left(z_\alpha + z_\beta\right)^2 \tilde{p}\left(1 - \tilde{p}\right)}{\delta^2}, \tag{2}$$

where $\tilde{p} = (p_A + p_B)/2$.

Thus, the sample size depends on the specified treatment effect δ and the "nuisance parameter" $\tilde{p} = (p_A + p_B)/2$.

(iii) Two treatment comparison with a survival outcome

Consider a Phase 3 trial comparing a new treatment with a control, where the primary endpoint is overall survival.

Survival times are assumed to follow a proportional hazards model with hazard rates

$$h_A(t) \qquad \text{on the control arm,}$$

$$h_B(t) = \lambda h_A(t) \quad \text{on the new treatment.}$$

Let $\theta = \log(\lambda)$.

If the new treatment is successful, $\lambda < 1$ and $\theta < 0$.

Thus, we wish to test H_0: $\theta \geq 0$ against $\theta < 0$.

A two treatment comparison with a survival outcome

We test H_0: $\theta \geq 0$ against $\theta < 0$ with one-sided type I error rate $\alpha = 0.025$ and we want a high probability of detecting an effect size $\theta = \delta$, i.e., a hazard ratio $\lambda = e^{\delta}$, where $\delta < 0$.

The distribution of the logrank statistic depends on the number of observed events (deaths in this case).

If the total number of events is n, the unstandardised logrank statistic is distributed, approximately, as

$$N(\theta\, n/4,\, n/4).$$

To achieve power $1 - \beta$ when $\theta = \delta$, we need a sample size and follow-up time that will yield a total of

$$n \;=\; \frac{2\,(z_\alpha + z_\beta)^2\, 4}{\delta^2} \tag{3}$$

events in the treatment and control groups together.

2. Dealing with unknowns in the sample size formula

The sample size formula, $n = 2\,(z_\alpha + z_\beta)^2\,\sigma^2/\delta^2$, for the normal case contains the response variance, σ^2.

Formula (2) for the binary case contains the average success rate $\tilde{p}$.

We call σ^2 and $\tilde{p}$ "nuisance parameters".

For survival data, the required number of events in (3) depends on: accrual rate, follow-up, baseline hazard rate and censoring.

The internal pilot study approach: Wittes & Brittain (*Statist. in Med.*, 1990) proposed a strategy to achieve desired power.

> Let ϕ denote a nuisance parameter in the sample size formula.
>
> Design the trial using an initial estimate, ϕ_0.
>
> At an interim analysis, estimate ϕ from the current data and re-calculate the sample size using this new estimate.

Example: Sample size re-estimation for a variance

A trial is to compare two cholesterol reducing drugs, A and B.

The primary endpoint is the fall in serum cholesterol over 4 weeks.

The one-sided type I error rate is set at $\alpha = 0.025$ and a power of $1 - \beta = 0.9$ is desired to detect an improvement of 0.4 mmol/L in Treatment B vs Treatment A.

Assuming the fall in cholesterol is normally distributed, if the response variance is σ^2, power 0.9 to detect an effect size $\delta = 0.4$ is achieved by a sample size per treatment of

$$n = \frac{2\left(z_{0.025} + z_{0.1}\right)^2 \sigma^2}{0.4^2}.$$

The initial estimate $\sigma_0^2 = 0.5$ gives a sample size per treatment of

$$n_0 = \frac{2\left(1.960 + 1.281\right)^2 0.5}{0.4^2} = 65.7 \approx 66.$$

Example: Sample size re-estimation for a variance

Following the Wittes & Brittain approach, an interim analysis is conducted after observing 33 patients per treatment.

Suppose this yields an estimated variance $\hat{\sigma}_1^2 = 0.62$.

We re-calculate the sample size per treatment as

$$n_1 = \frac{2\,(1.960 + 1.281)^2\,0.62}{0.4^2} = 81.4 \approx 82$$

and increase the total sample size to 82 per treatment arm.

We analyse the final data as if from a fixed sample size study.

Questions:

Does this procedure achieve the overall power of 0.9 when the treatment effect is 0.4?

Is the type I error rate controlled at 0.025?

Example: Sample size re-estimation for a variance

In our example, suppose the sample size rule is to calculate

$$n_1 = \frac{2\,(1.960 + 1.281)^2\,\hat{\sigma}_1^2}{0.4^2}$$

and take the maximum of n_1 and 66 as the new sample size per treatment arm.

If the true variance is $\sigma^2 = 0.6$ and this rule is applied,

The overall power is 0.899,

The type I error probability is 0.0256.

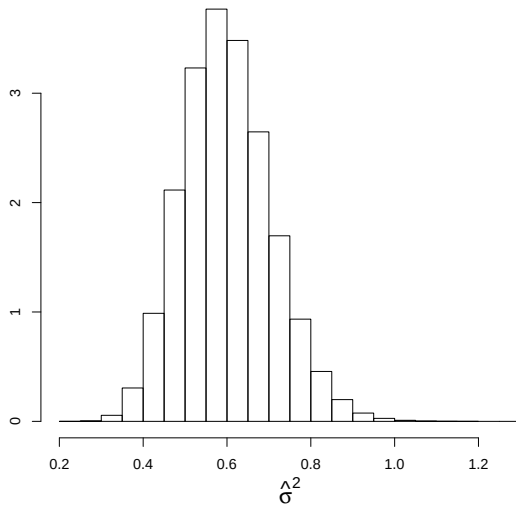
This small inflation of the type I error rate is quite typical — the type I error rate can be as high as 0.03 or 0.04 when the interim analysis has fewer observations.

Example: Sample size re-estimation for a variance

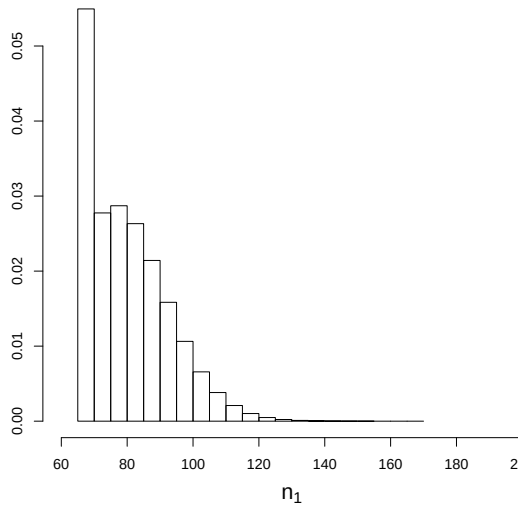
Even with 64 degrees of freedom to estimate σ^2 at the interim analysis, the estimate $\hat{\sigma}^2$ is very variable — and, hence, so is the total sample size n_1.

However, over-estimates compensate for under-estimates in achieving the desired power.

Histogram of $\hat{\sigma}^2$



Histogram of n_1



Similar variability in total sample size and similar type I error rate inflation are seen for the case of binary response.

Further comments on sample size re-estimation for σ^2

Information monitoring and group sequential tests:

This method of sample size re-estimation aims for a final estimate of the treatment effect θ with a certain variance.

Since "Information" is the reciprocal of $\text{Var}(\hat{\theta})$, we are aiming for a target information level.

The same method can be applied in a group sequential test (GST).

Mehta & Tsiatis (*Drug Information Journal*, 2001) implement an error spending GST where type I error probability is spent as a function of observed information.

A target information level is specified and, unless early stopping occurs, the trial continues until this target is reached — with additional recruitment to increase information if necessary.

Further comments on sample size re-estimation for σ^2

Blinded sample size re-estimation:

A number of papers address the question of re-estimating sample size when the treatment labels for each observation are not revealed (in the interests of maintaining complete blinding).

One can fit a mixture of two normal distributions to the pooled data: see papers by Gould & Shih, Friede & Kieser.

Alternatively, one can estimate σ^2 directly from the pooled data — even though this estimate will include a contribution from the difference in means of the two treatments.

Friede & Miller (*Applied Statistics*, 2012) observe that sample size modification based on this form of blinded estimate of σ^2 produces little or no type I error rate inflation.

Using a combination test with sample size re-estimation

A combination test:

Reference: Bauer & Köhne (*Biometrics*, 1994)

Define the null hypothesis H_0 (with a one-sided alternative).

We shall test H_0: $\theta \leq 0$ vs $\theta > 0$, with type I error probability α.

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

Stage 1

Observe the P-value P_1 for testing H_0.

After seeing Stage 1 data, design Stage 2 and fix the test statistic.

Stage 2

Observe the P-value P_2 for testing H_0.

A combination test

Suppose $\theta = 0$:

Then, P_1 has the usual $U(0,1)$ distribution.

Also, $P_2 \sim U(0,1)$ conditionally on the Stage 1 data and the resulting Stage 2 design.

Since the conditional distribution of P_2 is the same for all Stage 1 data, P_2 is independent of the Stage 1 data (including P_1).

Thus, when $\theta = 0$, P_1 and P_2 are independent $U(0,1)$ variables.

Bauer and Köhne proposed a test based on $P_1\,P_2$, using the fact that $P_1\,P_2 \sim \exp(-\chi^2_4/2)$ when P_1 and $P_2 \sim U(0,1)$.

Alternatively, one can define a test in terms of the Z-statistics

$$Z_1 = \Phi^{-1}(1-P_1) \quad \text{and} \quad Z_2 = \Phi^{-1}(1-P_2),$$

which have independent $N(0,1)$ distributions under $\theta = 0$.

The "inverse normal" combination test

To use the inverse normal" combination test, we stipulate that this test will be used and specify weights w_1 and w_2, where

$$w_1^2 + w_2^2 = 1.$$

The trial is conducted in two stages, with the design of Stage 2 set after seeing Stage 1 data, as described above.

The stage-wise P-values P_1 and P_2 are calculated, and from these we obtain $Z_1 = \Phi^{-1}(1 - P_1)$ and $Z_2 = \Phi^{-1}(1 - P_2)$.

The overall combination test rejects H_0 if

$$w_1 Z_1 + w_2 Z_2 > z_\alpha.$$

Since Z_1 and Z_2 are independent $N(0, 1)$ variables under $\theta = 0$,

$$w_1 Z_1 + w_2 Z_2 \sim N(0, 1)$$

and, hence, the combination test has type I error probability α.

Comments on the inverse normal combination test

The inverse normal combination test rejects H_0 if

$$w_1 Z_1 + w_2 Z_2 > z_\alpha,$$

where w_1 and w_2 are pre-specified weights satisfying $w_1^2 + w_2^2 = 1$.

It may be tempting to adjust the weights w_1 and w_2 to reflect the sample sizes actually seen in Stages 1 and 2.

However, this would undermine the mechanism by which the type I is protected — and type I error inflation may ensue.

Combination tests provide a method to control type I error probability precisely in adaptive trial designs.

Together, combination tests and multiple testing procedures underpin a wide variety of adaptive designs, including seamless Phase 2/3 designs and enrichment trials.

Example: Sample size re-estimation for a variance

In the setting of our previous example, we proceed as follows.

First stipulate that an inverse normal combination test will be used with weights $w_1 = w_2 = 1/\sqrt{2}$.

Stage 1

Take $n_0 = 33$ observations per treatment arm,

Calculate the t-statistic

$$t_1 = \frac{\widehat{\theta}_1}{\sqrt{2\,\widehat{\sigma}_1^2\,/\,33}}$$

and find the corresponding P-value $P_1 = P\{T_{64} > t_1\}$.

Compute $n_1 = 2\,(z_\alpha + z_\beta)^2\,\widehat{\sigma}_1^2\,/\,\delta^2$ and set the Stage 2 sample size per treatment to be $n_2 = \max(n_1, 66) - 33$.

Example: Sample size re-estimation for a variance

Stage 2

Take n_2 observations per treatment arm,

Calculate the t-statistic based solely on Stage 2 data

$$t_2 \;=\; \frac{\widehat{\theta}_2}{\sqrt{2\,\widehat{\sigma}_2^2\,/\,n_2}}$$

and find the P-value $P_2 = P\{T_{2n_2-2} > t_2\}$.

The overall combination test

Find $Z_1 = \Phi^{-1}(1-P_1)$ and $Z_2 = \Phi^{-1}(1-P_2)$.

Using the pre-specified weights, reject H_0: $\theta \leq 0$ if

$$(1/\sqrt{2})\,Z_1 + (1/\sqrt{2})\,Z_2 \;>\; z_\alpha.$$

Properties of the trial design using a combination test

In the above design, we take an initial sample of $n_0 = 33$ per treatment, then calculate

$$n_1 \;=\; 2\,(1.960 + 1.281)^2\,\hat{\sigma}_1^2/0.4^2,$$

and take a further $\{\max(n_1, 66) - 33\}$ patients per treatment arm.

If the true variance is $\sigma^2 = 0.6$ the Wittes & Brittain method gave

Type I error probability = **0.0256**,

Overall power = **0.899**

Using the inverse normal combination test, we have

Type I error probability = **0.0250**,

Overall power = **0.896**.

Sample size re-estimation with a binary response

Example: Treatment for heart failure

A new treatment is to be compared to the current standard.

The primary endpoint is

Re-admission to hospital (or death) within 30 days.

The current treatment has a re-admission rate of 25%.

Testing for superiority

It is hoped the new treatment will reduce re-admissions to 20%.

Denote re-admission probabilities by p_t on the new treatment and p_c on the control.

To establish superiority of the new treatment, we test H_0: $p_t \geq p_c$ against $p_t < p_c$ — hoping to reject H_0.

Binary example: Treatment for heart failure

Setting $\theta = p_c - p_t$, we wish to test H_0: $\theta \leq 0$ against $\theta > 0$ with

Type I error rate: $\alpha = 0.025$ at $\theta = 0$,

Power: $1 - \beta = 0.9$ when $\theta = \delta = 0.25 - 0.2 = 0.05$.

From (2), we achieve this power with a sample size of

$$n = \frac{2\,(z_\alpha + z_\beta)^2\,\tilde{p}\,(1-\tilde{p})}{\delta^2},$$

in each treatment group, where $\tilde{p} = (p_c + p_t)/2$.

From historical data, we expect $p_c = 0.25$.

With $p_t = 0.2$, this gives $\tilde{p} = (0.25 + 0.2)/2 = 0.225$, and the required sample size test per treatment arm is

$$n_0 = 1466.$$

Binary example: Treatment for heart failure

The sample size formula depends on p_c, as well as $\theta = p_c - p_t$.

Desired power may not be achieved if previous data are not representative of the new study: for example, hospitals involved may have a different case mix and admit more seriously ill patients.

Suppose investigators decide to conduct an interim analysis at which an increase in sample size may be agreed.

The trial design

A Bauer & Köhne two-stage design is specified.

Data from before and after the interim analysis will be combined using an inverse normal combination test with $w_1 = w_2 = 1/\sqrt{2}$.

The initial calculation gave a target sample size of $n_0 = 1466$ per treatment arm: we recruit 730 patients per treatment in Stage 1.

Binary example: Treatment for heart failure

Stage 1, with 730 subjects per treatment, yields $\hat{p}_c = 0.253$ and $\hat{p}_t = 0.219$, so $\widehat{\theta} = 0.034$ with standard error 0.0222.

A test of H_0: $\theta \leq 0$ has $Z_1 = 0.034/0.0222 = 1.531$.

The overall test will reject H_0 if

$$(1/\sqrt{2})\, Z_1 + (1/\sqrt{2})\, Z_2 \;>\; z_\alpha \;=\; 1.96.$$

Since $Z_1/\sqrt{2} = 1.083$, results thus far are promising. However, a positive outcome is by no means certain.

Investigators learn that trials of competing treatments have been unsuccessful.

It is decided to increase the second stage sample size to give higher probability of a positive outcome under the original alternative, $\theta = 0.05$ — and under smaller effect sizes.

Binary example: Planning the Stage 2 sample size

Implications of several sample size choices are summarised below.

p_c	p_t	θ	Stage 2 sample size	Conditional power
0.25	0.22	0.03	750	0.55
			1000	0.63
			1250	0.70
0.25	0.21	0.04	750	0.73
			1000	0.81
			1250	0.87
0.25	0.20	0.05	750	0.86
			1000	0.92
			1250	0.96

Investigators increase Stage 2 sample size to 1000 per treatment.

Binary example: Treatment for heart failure

With 1000 subjects per treatment, Stage 2 data (alone) yield $\hat{p}_c = 0.248$ and $\hat{p}_t = 0.223$.

Thus, $\widehat{\theta} = 0.025$ with standard error 0.0190.

A test of H_0: $\theta \leq 0$ based on Stage 2 data has Z-statistic $0.025/0.0190 = 1.318$.

In the overall test,

$$Z_1/\sqrt{2} + Z_2/\sqrt{2} \;=\; (1.531 + 1.318)/\sqrt{2} \;=\; 2.013 \;>\; z_\alpha.$$

Thus, the null hypothesis H_0: $\theta \leq 0$ is rejected and the new treatment is recommended for use.

We have dealt with unknown "nuisance parameters" — and we have seen other reasons for a change in sample size.

3. The unknown treatment effect

In the last example, we saw the opportunity to increase sample size in order to increase power under smaller treatment effects.

It seems quite reasonable to do this in response to external information that was not available when the trial was designed.

Should one do this in response to an interim estimate, $\widehat{\theta}$, of the treatment effect?

In the sample size formulae

$$n = \frac{2\left(z_\alpha + z_\beta\right)^2 \sigma^2}{\delta^2} \qquad \text{(normal case)},$$

$$n = \frac{2\left(z_\alpha + z_\beta\right)^2 \tilde{p}\left(1-\tilde{p}\right)}{\delta^2} \qquad \text{(binary case)},$$

is δ an "unknown"?

Choosing the effect size δ for the sample size formula

We denote by θ the effect size of a new treatment, i.e., the difference in mean response between the new treatment and the control.

If we wish the trial to have power $1 - \beta$ when $\theta = \delta$, we put the effect size δ in the sample size formula.

Dispute can arise over the choice of δ.

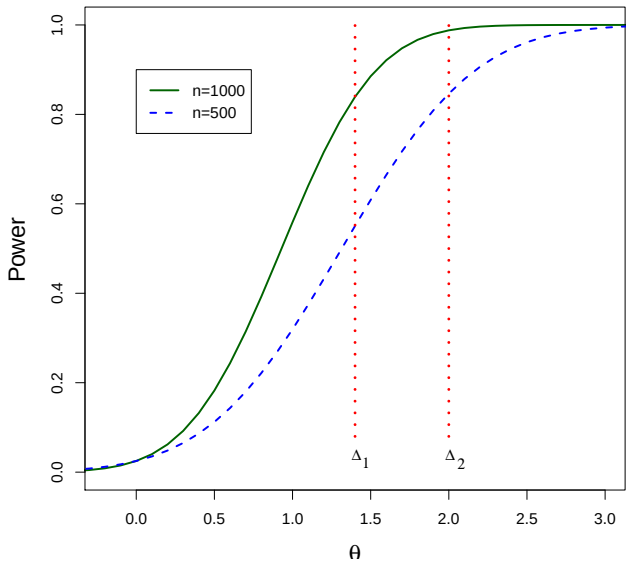
For example, should investigators use:

The minimum effect of interest Δ_1, or

The anticipated effect size Δ_2 ?

An example: Specifying the effect size δ

Suppose sample sizes of 500 and 1000 give these power curves:



With 1000 subjects, there is good power at the minimum clinically significant effect, Δ_1.

With only 500 subjects, a high power is achieved at the more optimistic Δ_2 — but there is not a lot of power at Δ_1.

If $\theta = \Delta_2$, a sample size of 1000 is unnecessarily high.

An example: Specifying the effect size δ

A sample size of 500 would be sufficient if $\theta = \Delta_2$.

However, if $\theta = \Delta_1$ we would like to have the power provided by a sample size of around 1000.

An adaptive strategy: Start small then ask for more

Start with a planned sample size of 500,

Look at the results of the first 250 observations,

If appropriate, increase the sample size to 1000.

The group sequential approach

Start with a maximum sample size of 1000,

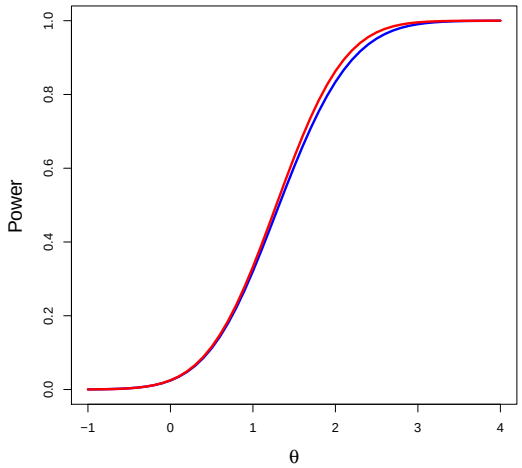
Conduct one or more interim analyses,

Stop early if there is enough evidence to do so.

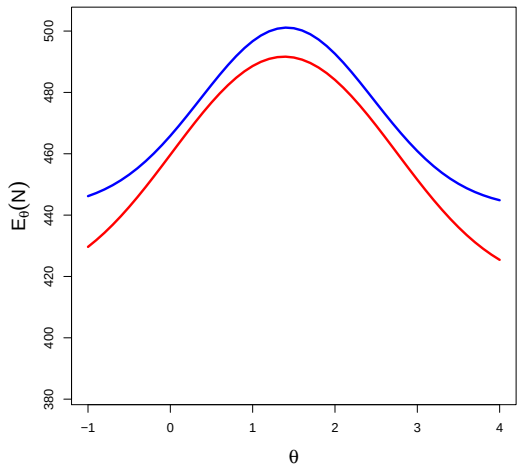
Comparing different types of trial design

All designs have overall power and $E_\theta(N)$ curves.

Power curve



$E_\theta(N)$ curves



Designs with similar power curves can be compared in terms of their average sample size functions, $E_\theta(N)$.

Even if investigators are uncertain about the likely treatment effect, they can usually specify values of θ under which early stopping is most desirable.

Adaptive or group sequential designs?

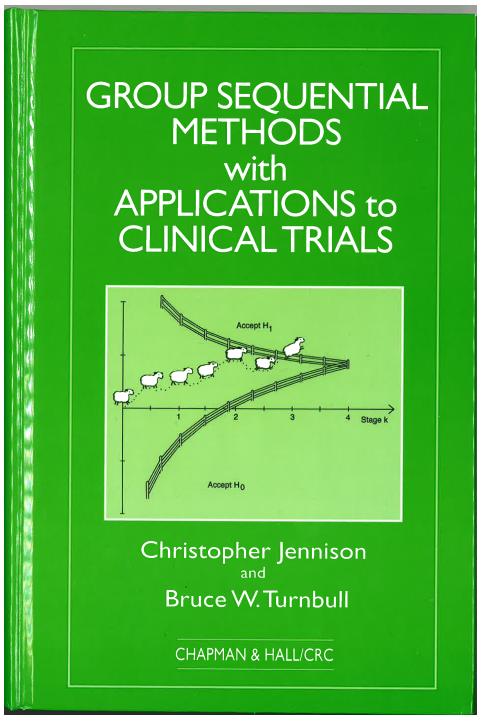
Jennison & Turnbull have studied optimal versions of adaptive and non-adaptive sequential designs (e.g., *Statist. in Med.*, 2003 and 2006, *Biometrika*, 2006). They report:

The set of group sequential tests (GSTs) is a subset of the set of adaptive designs,

Adaptive designs are, at best, a little more efficient than GSTs with the same number of analyses, reducing average sample size by 1% or 2% for the same power,

Many published adaptive designs are considerably less efficient than a well chosen GST.

And advice is available on how to create good group sequential designs:



What to look for in a trial design

If you are considering a trial design with sample size re-estimation in response to an interim estimate of the treatment effect, then:

Look at the power function and $E_\theta(N)$ curve,

Compare with $E_\theta(N)$ for a standard GST, e.g., from the ρ-family of error spending tests (J & T, Ch. 7).

You should be wary of a sample size rule that treats an interim $\widehat{\theta}$ as an accurate estimate of the true θ:

In the Heart Failure example, we wanted power 0.9 to differentiate between $\theta = 0$ and $\theta = 0.05$,

The interim estimate $\widehat{\theta} = 0.034$ had a standard error of 0.0222, giving a 95% CI for θ of $(-0.01, 0.08)$,

This scale of standard error of $\widehat{\theta}$ is typical.

Some comments on the "Promising Zone" approach

Mehta & Pocock (*Statist. in Med.*, 2010) proposed a particular form of sample size re-estimation in their paper:

"Adaptive increase in sample size when interim results are promising: A practical guide with examples"

In their Example 1, response is measured 26 weeks after treatment, causing problems for standard group sequential tests.

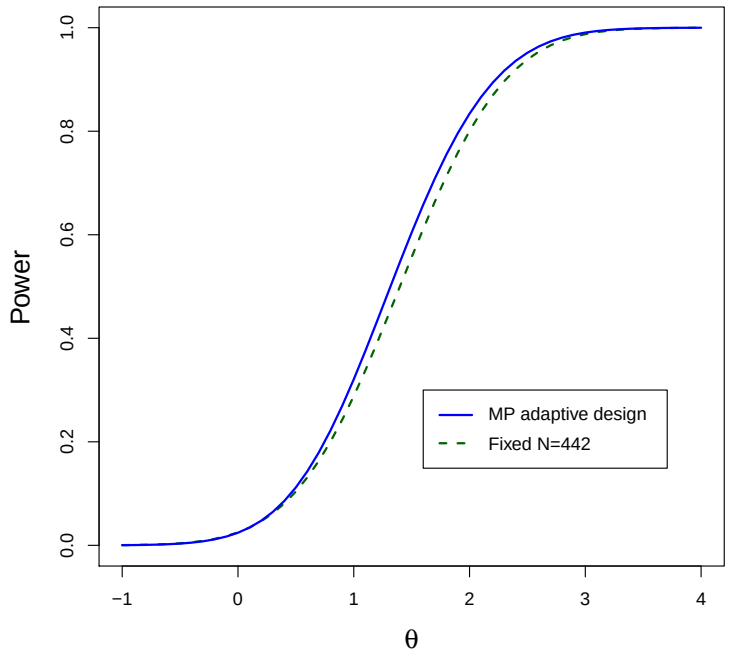
At the interim analysis, there is a large number of "pipeline" patients who have been treated but are yet to produce a response.

Jennison & Turnbull focus on this example in their (*Statist. in Med.*, 2015) paper

"Adaptive sample size modification in clinical trials: start small then ask for more?"

Properties of the Mehta-Pocock design

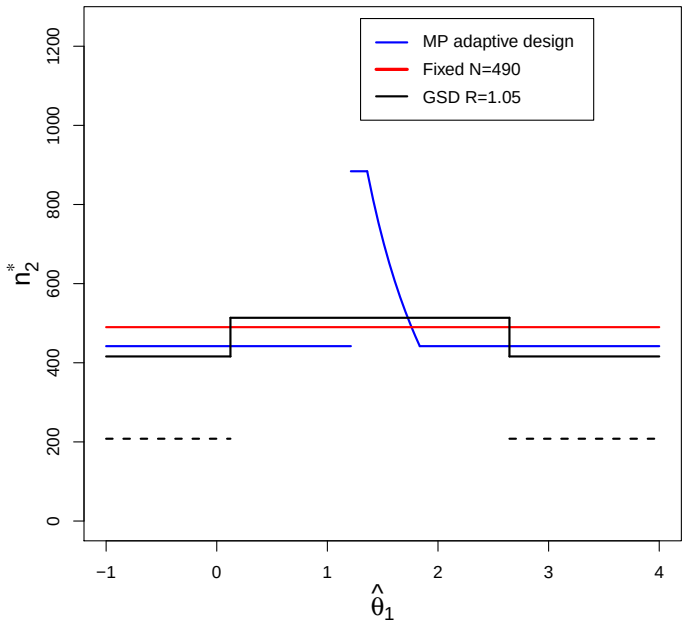
M & P use a result of Chen, DeMets & Lan (*Statist. in Med.*, 2004) that allows an increase in sample size (in certain situations) to be followed by a standard, fixed sample size analysis at the end of the trial.



J & T note that the limited opportunity for increasing sample size leads to only a small increase in overall power.

Alternatives to the MP design for their Example 1

J & T explore other ways of achieving the power of MP's design.

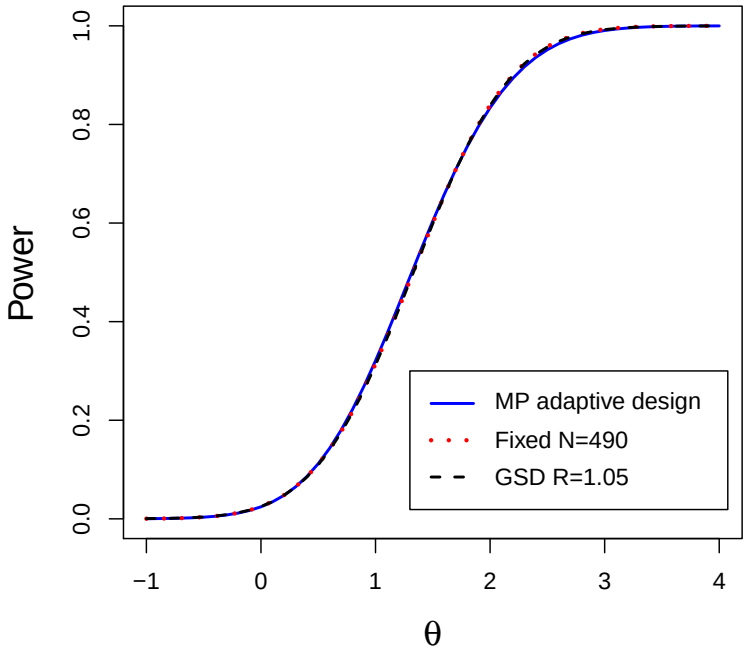


1. A fixed sample size design with 490 observations (cf, the minimum of 442 for MP)

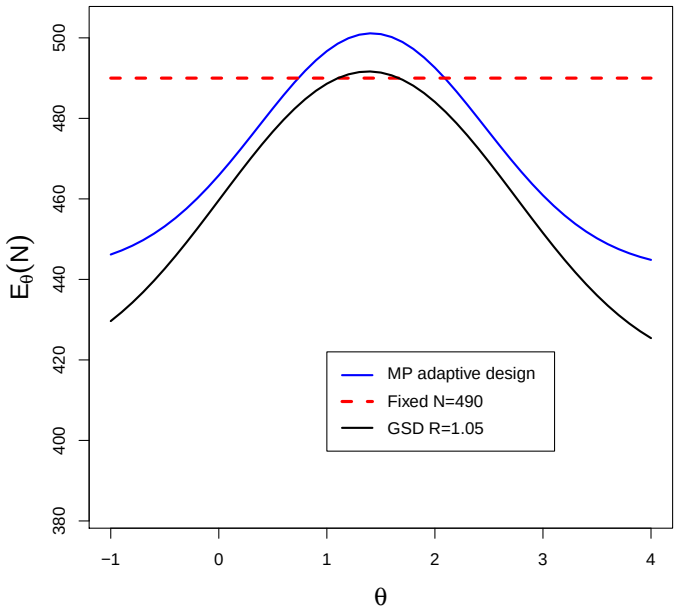
2. A group sequential test that stops after with a sample size of 416 or 514. If the GST stops at the first analysis, responses from the 208 pipeline subjects are not used — but these patients are counted in $E_\theta(N)$.

Alternatives to the MP design for their Example 1

Power curve



$E_\theta(N)$ curves



All three designs have essentially the same power curve.

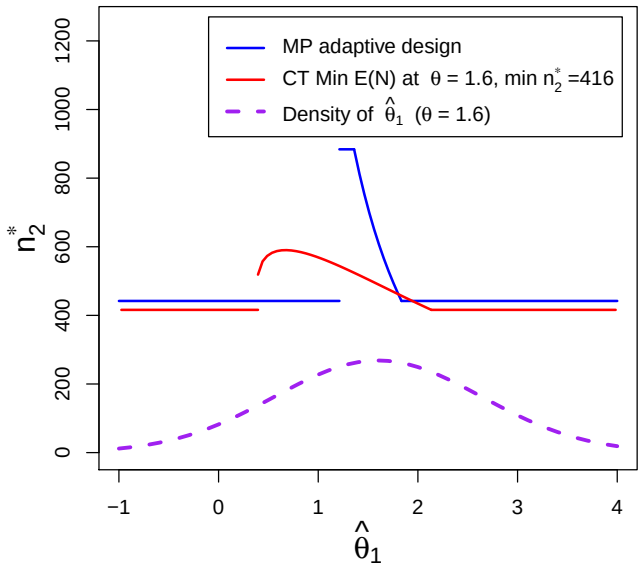
The fixed sample design has lower $E_\theta(N)$ than the MP design over the θ values of most interest.

The GST has uniformly lower $E_\theta(N)$ than the MP design.

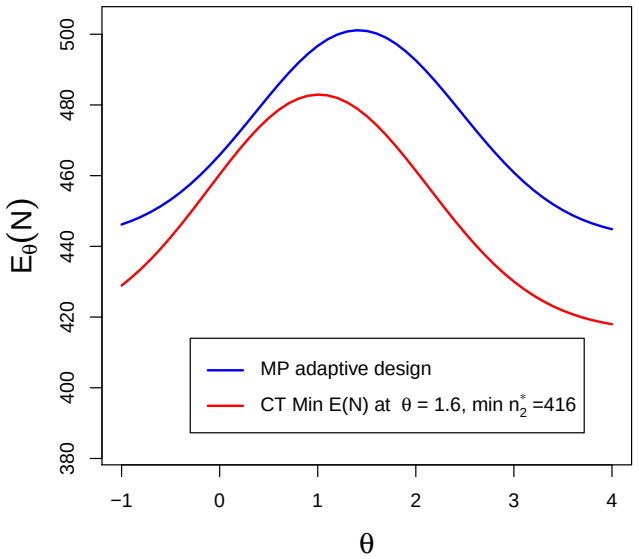
Alternatives to the MP design for their Example 1

J & T go on to develop a method in the adaptive framework ("start small and ask for more") that lowers the $E_\theta(N)$ curve while maintaining power.

Sample size rules



$E_\theta(N)$ curves



J & T use an inverse normal combination test.

They also employ a "rate of exchange" between sample size and power to ensure a consistent approach to the choice of sample size.

4. Conclusions

Sample size re-estimation in response to information about **nuisance parameters** can help in achieving a desired power curve.

In doing this, the use of combination tests avoids any inflation of the type I error rate.

Group sequential tests provide a mechanism for responding to information about the **treatment effect** — by stopping the trial at an interim analysis.

GSTs are tried and tested, and special forms of GST have been developed to deal with unequal group sizes and delayed response (see Hampson & Jennison, J. Roy. Statist. Soc., B, 2013).

Some adaptive designs match the performance of good GSTs.

However, some other adaptive designs do not.