Sample Size Re-estimation

in Clinical Trials

Christopher Jennison,

Dept of Mathematical Sciences,

University of Bath, UK

http://www.bath.ac.uk/$\sim$mascj

Amgen, Cambridge

14 November 2003

Plan of talk

1. Internal pilots in fixed sample studies

2. Group sequential tests

3. Example: 5-group test, normal response

4. Estimating σ^2 during a group sequential test

5. Example: survival data

6. Error spending tests

7. Changing the power requirement in mid-study

1. Internal pilots in fixed sample studies

The sample size needed to satisfy a power requirement often depends on an unknown nuisance parameter.

Examples are:

Variance, σ^2, of a normal response.

Binomial response: since variance depends on p, the sample size needed to detect a difference in probabilities $p_1 - p_2 = \delta$ depends on $(p_1 + p_2)/2$.

Survival data: information is governed by the number of observed deaths, and this depends on the overall failure rate and the degree of censoring.

“Over-interpretation of results from a small pilot study, positive or negative, may undermine support for the major investigation” (W. G. Cochran).

Internal pilots

Wittes & Brittain (1990, *Statistics in Medicine*) suggest an “internal” pilot.

> Let ϕ denote a nuisance parameter and suppose the sample size required under a given value of this parameter is $n(\phi)$.
>
> From a pre-study estimate, $\widehat{\phi}_0$, calculate an initial planned sample size of $n(\widehat{\phi}_0)$.
>
> At an interim stage, find a new estimate $\widehat{\phi}_1$ from the data obtained so far. Aim for the new target sample size of $n(\widehat{\phi}_1)$.

Variations on this are possible, e.g., only allow an increase over the original target sample size.

Features of Internal Pilot designs

Using Wittes and Brittain’s approach, the type I error rate is only slightly perturbed.

Results for normal data as σ^2 is estimated (Jennison & Turnbull, 2000, Ch. 14):

s_1^2 on 18 degrees of freedom: 0.05 $\rightarrow$ 0.050 – 0.057

s_1^2 on 38 degrees of freedom: 0.05 $\rightarrow$ 0.052 – 0.053

But, calculating s^2 may reveal the effect estimate, $\widehat{\theta}$:

this is undesirable as it breaks the blinding,

adjusting the sample size in the knowledge of $\widehat{\theta}$ can seriously inflate type I error rates — possibly to more than double its intended value.

Estimating σ^2 from pooled data

Note that with n observations per treatment

$$\textit{Total sum of squares} = (n-2)s^2 + \frac{n}{2}\,\widehat{\theta}^2.$$

Thus, knowledge of s^2 and a list of responses without treatment codes is enough to work out the value of $|\widehat{\theta}|$.

Gould & Shih (1998, *Statistics in Medicine*) propose use of the EM algorithm to fit two normal distributions to the pooled data. They claim the error in $\widehat{\sigma}^2$ is small but error in $|\widehat{\theta}|$ is high — so, effectively, results remain blinded.

Friede & Kieser (2002, *Statistics in Medicine*) criticise Gould and Shih's method because of:

deficiencies in the EM algorithm,

failure to allow for special randomisation procedures.

An alternative pooled data estimate of σ^2

Friede & Kieser suggest the alternative estimator

$$\widehat{\sigma}^2 \;=\; \frac{\textit{Total sum of squares}}{n-1}$$

as they find this has better properties than the Gould and Shih estimator.

Other options . . .

An independent party could calculate s^2 and reveal this (and only this) to the study's sponsors and the DSMB.

Combine the sample size exercise with an interim analysis of treatment effect — so $\widehat{\theta}$ (or at least $|\widehat{\theta}|$) will be made known anyway. It could also help to have a pre-set "sample size" rule ready to apply.

2. Group sequential tests

In clinical trials, animal trials and epidemiological studies there are reasons of

ethics

administration (accrual, compliance, . . .)

economics

to monitor progress and accumulating data.

Subjects should not be exposed to unsafe, ineffective or inferior treatments. National and international guidelines call for interim analyses to be performed — and reported.

It is now standard practice for medical studies to have a Data and Safety Monitoring Board to oversee the study and consider the option of early termination.

The need for special methods

There is a danger that multiple looks at data can lead to over-interpretation of interim results

Overall Type I error rate applying

repeated significance tests at

$\alpha = 5$*% to accumulating data*

Number of tests	Error rate
1	0.05
2	0.08
3	0.11
5	0.14
10	0.19
20	0.25
100	0.37
∞	1.00

Pocock (1983) *Clinical Trials* Table 10.1,
Armitage, *et al.* (1969), Table 2.

Pocock's repeated significance test

(1977, *Biometrika*)

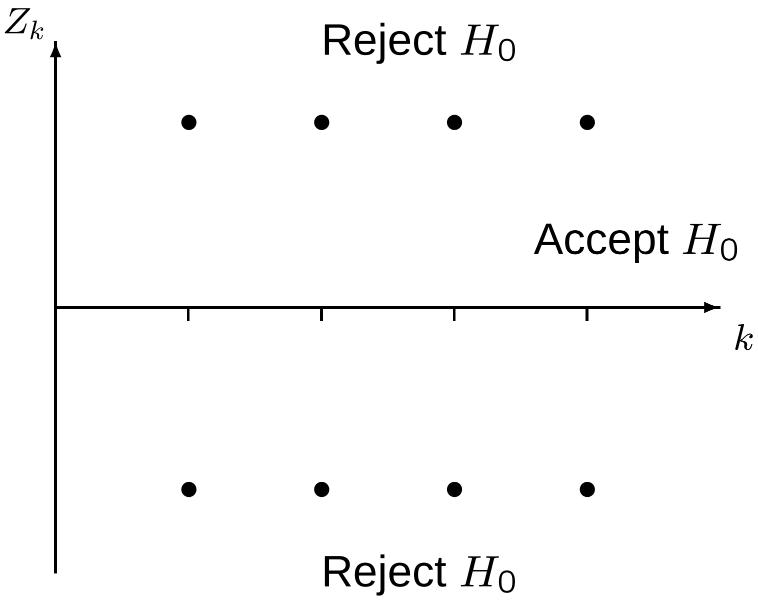
To test H_0: $\theta = 0$ against $\theta \neq 0$.

Fix a total number of analyses K. Use standardised test statistics Z_k, $k = 1, \ldots, K$.

Stop to reject H_0 at analysis k if

$$|Z_k| > c.$$

If H_0 has not been rejected by analysis K, stop and accept H_0.



Joint distribution of parameter estimates

Reference: Jennison & Turnbull (2000), Ch. 11

Suppose our main interest is in the parameter θ and let $\widehat{\theta}_k$ denote the estimate of θ based on data available at analysis k.

The information for θ at analysis k is

$$\mathcal{I}_k = \{\text{Var}(\widehat{\theta}_k)\}^{-1}, \quad k = 1, \ldots, K.$$

Canonical joint distribution of $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$

In many situations, $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$ are approximately multivariate normal,

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

and

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

Sequential distribution theory

The preceding results for the joint distribution of $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$ can be demonstrated directly for:

θ a single normal mean,

$\theta = \mu_A - \mu_B,$ the effect size in a comparison of two normal means.

The results also apply when θ is a parameter in:

a general normal linear,

a general model fitted by maximum likelihood (large sample theory).

There are related canonical distributions for z-statistics and score statistics.

Survival data

The canonical joint distributions also arise for

a) the estimates of a parameter in Cox's proportional hazards regression model

b) a sequence of log-rank statistics (score statistics) for comparing two survival curves

— and to z-statistics formed from these.

For survival data, observed information is roughly proportional to the number of failures seen.

Special types of group sequential test are needed to handle unpredictable and unevenly spaced information levels: see *error spending tests*.

Implementing a group sequential test

Think through a fixed sample version of the study.

Decide on the type of early stopping, number of analyses, and choice of stopping boundary: these will imply increasing the fixed sample size by a certain “inflation factor”.

In interim monitoring, compute the standardised statistic Z_k at each analysis and compare with critical values (calculated specifically for the observed information levels in the case of an error spending test).

On termination, one can obtain p-values and confidence intervals possessing the usual frequentist interpretations.

3. Example of a two treatment comparison, normal response, 2-sided test

Cholesterol reduction trial

Treatment A: new, experimental treatment

Treatment B: current treatment

Primary endpoint: reduction in serum cholesterol level over a four week period

Aim: To test for a treatment difference.

High power should be attained if the mean cholesterol reduction differs between treatments by 0.4 *mmol/l*.

How would we design a fixed-sample study?

Denote responses by

X_{Ai}, $i = 1, \ldots, n_A$, on treatment A,

X_{Bi}, $i = 1, \ldots, n_B$, on treatment B.

Suppose each

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

Problem: to test H_0: $\mu_A = \mu_B$ with

two-sided type I error probability $\alpha = 0.05$

and power 0.9 at $|\mu_A - \mu_B| = \delta = 0.4$.

We suppose σ^2 is known to be 0.5.

(Facey, 1992, *Controlled Clinical Trials*)

Fixed sample design

Standardised test statistic

$$Z = \frac{\bar{X}_A - \bar{X}_B}{\sqrt{\sigma^2/n_A + \sigma^2/n_B}}\,.$$

Under H_0, $Z \sim N(0,1)$ so reject H_0 if

$$|Z| > \Phi^{-1}(1-\alpha/2).$$

Let $\mu_A - \mu_B = \theta$. If $n_A = n_B = n$,

$$Z \sim N(\frac{\theta}{\sqrt{2\sigma^2/n}}, 1)$$

so, to attain desired power at $\theta = \delta$, aim for

$$\begin{aligned} n &= \{\Phi^{-1}(1-\alpha/2) + \Phi^{-1}(1-\beta)\}^2\, 2\sigma^2/\delta^2 \\ &= (1.960 + 1.282)^2 (2 \times 0.5)/0.4^2 = 65.67, \end{aligned}$$

i.e., 66 subjects on each treatment.

Group sequential design

Specify type of early termination:

stop early to reject H_0

Number of analyses:

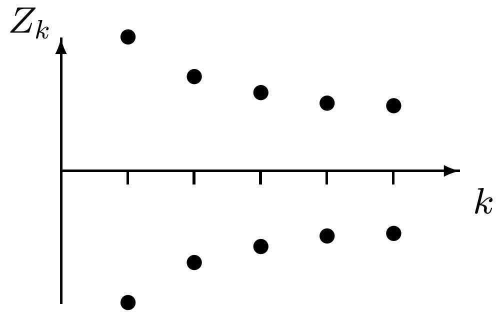
5 (fewer if we stop early)

Stopping boundary:

O'Brien & Fleming (1979, Biometrics).

Reject H_0 at analysis k, $k = 1, \ldots, 5$,

if $|Z_k| > c\sqrt{\{5/k\}}$,



where Z_k is the standardised statistic

based on data at analysis k.

Example: cholesterol reduction trial

O’Brien & Fleming design

From tables (JT, Table 2.3) or computer software

$$c = 2.040 \qquad \text{for } \alpha = 0.05$$

so reject H_0 at analysis k if

$$|Z_k| > 2.040\,\sqrt{5/k}.$$

Also, for specified power, inflate the fixed sample size by a factor (JT, Table 2.4)

$$IF = 1.026$$

to get the maximum sample size

$$1.026 \times 65.67 = 68.$$

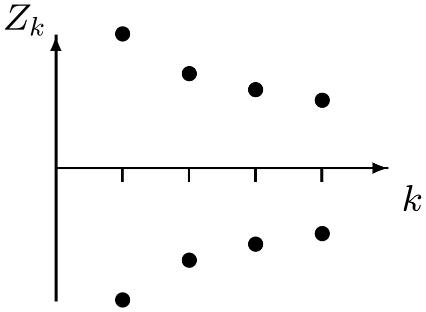
Divide this into 5 groups of 13 or 14 observations per treatment.

Some designs with K analyses

O'Brien & Fleming

Reject H_0 at analysis k if

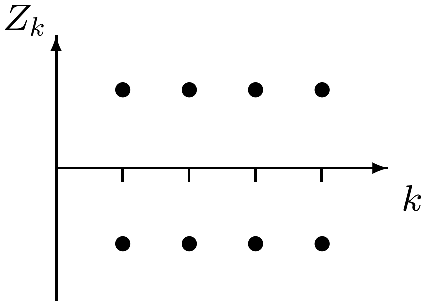
$$|Z_k| > c\sqrt{K/k}.$$



Pocock

Reject H_0 at analysis k if

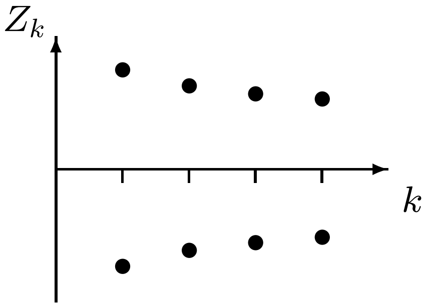
$$|Z_k| > c.$$



Wang & Tsiatis, shape Δ

Reject H_0 at analysis k if

$$|Z_k| > c\,(k/K)^{\Delta-1/2}.$$



($\Delta = 0$ gives *O'Brien & Fleming*, $\Delta = 0.5$ gives *Pocock*)

Example: cholesterol reduction trial

Properties of different designs

Sample sizes are per treatment.

Fixed sample size is 66.

K	Maximum sample size	Expected sample size $\theta = 0$	$\theta = 0.2$	$\theta = 0.4$
O'Brien & Fleming				
2	67	67	65	56
5	68	68	64	50
10	69	68	64	48
Wang & Tsiatis, $\Delta = 0.25$				
2	68	67	64	52
5	71	70	65	47
10	72	71	64	44
Pocock				
2	73	72	67	51
5	80	78	70	45
10	84	82	72	44

Implementing the OBF test

Divide the total sample size of 68 per treatment into 5 groups of roughly equal size, e.g.,

14 in groups 1 to 3,   13 in groups 4 and 5.

At analysis k, define

$$\bar{X}_A^{(k)} = \frac{1}{n_{Ak}} \sum_{i=1}^{n_{Ak}} X_{Ai}, \quad \bar{X}_B^{(k)} = \frac{1}{n_{Bk}} \sum_{i=1}^{n_{Bk}} X_{Bi}$$

and

$$Z_k = \frac{\bar{X}_A^{(k)} - \bar{X}_B^{(k)}}{\sqrt{\sigma^2(1/n_{Ak} + 1/n_{Bk})}}.$$

Stop to reject H_0 if

$$|Z_k| > 2.040\sqrt{5/k}, \quad k = 1, \ldots, 5.$$

Accept H_0 if $|Z_5| < 2.040$.

If n_{Ak} and n_{Bk} differ from their planned values — still use the above rule.

4. Updating a design as the nuisance parameter σ^2 is estimated

We can design as for the case of known variance but use an *estimate* of σ^2 initially.

If in doubt, err towards over-estimating σ^2 in order to safeguard the desired power.

At analysis k, estimate σ^2 by

$$s_k^2 = \frac{\sum(X_{Ai} - \bar{X}_A^{(k)})^2 + \sum(X_{Bi} - \bar{X}_B^{(k)})^2}{n_{Ak} + n_{Bk} - 2}\,.$$

In place of Z_k, define t-statistics

$$T_k = \frac{\bar{X}_A^{(k)} - \bar{X}_B^{(k)}}{\sqrt{s_k^2(1/n_{Ak} + 1/n_{Bk})}}\,,$$

then test at the *same significance level* used for Z_k when σ^2 is known.

Updating the target sample size

Recall, maximum sample size is set to be the fixed sample size multiplied by the Inflation Factor.

In a 5-group O'Brien & Fleming design for the cholesterol example this is

$$1.026 \times \{\Phi^{-1}(1-\alpha/2) + \Phi^{-1}(1-\beta)\}^2 \, 2\sigma^2/\delta^2$$

$$= 134.8 \times \sigma^2.$$

After choosing the first group sizes using an initial estimate of σ^2, at each analysis $k = 1, 2, \ldots$ we can re-estimate the target for n_{A5} and n_{B5} as

$$134.8 \times s_k^2$$

and modify future group sizes to achieve this.

Example: updating the sample size

Initially:

With initial estimate $\widehat{\sigma}^2 = 0.5$,

aim for $n_{A5} = n_{B5} = 134.8 \times 0.5 = 68$.

Plan 14 observations per treatment group.

Analysis 1:

With $n_{A1} = n_{B1} = 14$ and $s_1^2 = 0.80$,

aim for $n_{A5} = n_{B5} = 134.8 \times 0.80 = 108$.

For now, keep to 14 obs. per treatment group.

Analysis 2:

With $n_{A2} = n_{B2} = 28$ and $s_2^2 = 0.69$,

aim for $n_{A5} = n_{B5} = 134.8 \times 0.69 = 93$.

Now increase group size to 22 obs. per treatment.

Example: updating the sample size

Analysis 3:

With $n_{A3} = n_{B3} = 50$ and $s_3^2 = 0.65$,

aim for $n_{A5} = n_{B5} = 134.8 \times 0.65 = 88$.

Set next group size to 19 obs. per treatment.

Analysis 4:

With $n_{A4} = n_{B4} = 69$ and $s_4^2 = 0.72$,

aim for $n_{A5} = n_{B5} = 134.8 \times 0.72 = 97$.

Set final group size to 28 obs. per treatment.

Analysis 5:

With $n_{A5} = n_{B5} = 97$, suppose $s_5^2 = 0.74$,

so the target is $n_{A5} = n_{B5} = 134.8 \times 0.74 = 100$

— and the test may be slightly under-powered.

Remark on “re-estimating” sample size

The target information for $\theta = \mu_A - \mu_B$ is

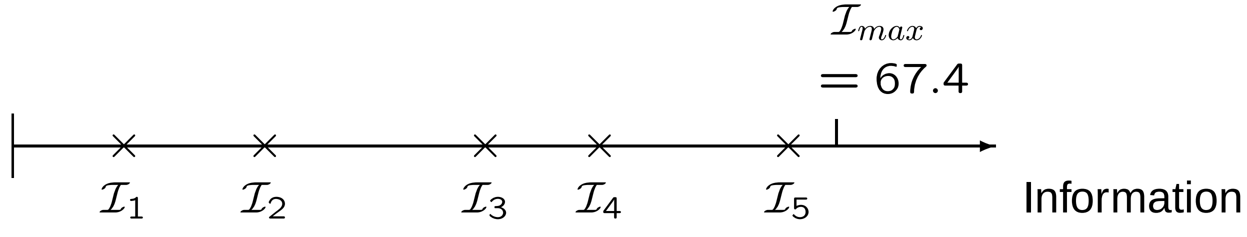
$$\begin{aligned}\mathcal{I}_{max} &= IF \times \{\Phi^{-1}(1-\alpha/2) + \Phi^{-1}(1-\beta)\}^2/\delta^2 \\ &= 1.026 \times (1.960 + 1.282)^2/0.4^2 = 67.4.\end{aligned}$$

The relation between information and sample size

$$\mathcal{I}_k = \left\{ \left(\frac{1}{n_{Ak}} + \frac{1}{n_{Bk}} \right) \sigma^2 \right\}^{-1}$$

involves the *unknown* σ^2. Hence, the initial uncertainty about the necessary sample size.

In effect, we proceed by monitoring observed information:



NB, state $\mathcal{I}_{max} = 67.4$ in the protocol, not $n = \ldots$

5. A survival data example

Oropharynx Clinical Trial Data

(Kalbfleisch & Prentice (1980) Appendix 1, Data Set II)

Patient survival was compared on experimental Treatment A and standard Treatment B.

		Number entered		Number of deaths	
k	Date	Trt A	Trt B	Trt A	Trt B
1	12/69	38	45	13	14
2	12/70	56	70	30	28
3	12/71	81	93	44	47
4	12/72	95	100	63	66
5	12/73	95	100	69	73

The logrank statistic

At stage k, observed number of deaths is d_k. Elapsed times between study entry and failure are $\tau_{1,k} < \tau_{2,k} < \ldots < \tau_{d_k,k}$ (assuming no ties).

Define

$r_{iA,k}$ and $r_{iB,k}$	numbers at risk on Treatments A and B at $\tau_{i,k}-$
$r_{ik} = r_{iA,k} + r_{iB,k}$	total number at risk at $\tau_{i,k}-$
O_k	observed number of deaths on Treatment B at stage k
$E_k = \sum_{i=1}^{d_k} \frac{r_{iB,k}}{r_{ik}}$	"expected" number of deaths on Treatment B at stage k.
$V_k = \sum_{i=1}^{d_k} \frac{r_{iA,k} r_{iB,k}}{r_{ik}^2}$	"variance" of O_k

The standardised logrank statistic at stage k is

$$Z_k = \frac{O_k - E_k}{\sqrt{V_k}}.$$

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,

$$Z_k \sim N(\theta\sqrt{\mathcal{I}_k},\, 1)$$

and

$$\text{Cov}(Z_{k_1}, Z_{k_2}) = \sqrt{(\mathcal{I}_{k_1}/\mathcal{I}_{k_2})}, \quad 1 \le k_1 \le k_2 \le K,$$

where $\mathcal{I}_k = V_k$.

For $\lambda \approx 1$, we have $\mathcal{I}_k \approx V_k \approx d_k/4$.

Design of the Oropharynx trial

One-sided test of H_0: $\theta \leq 0$ vs $\theta > 0$. Under the alternative $\lambda > 1$, i.e., Treatment A is better.

Require:

type I error probability $\alpha = 0.05$,

power $1 - \beta = 0.95$ at $\theta = 0.6$, i.e., $\lambda = 1.8$.

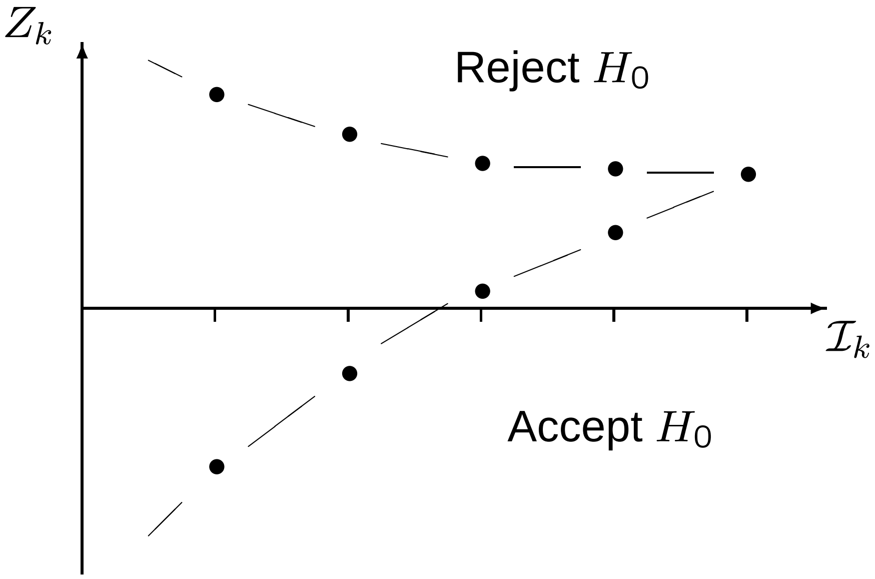
Information needed for a fixed sample study is

$$\mathcal{I}_f = \frac{\{\Phi^{-1}(\alpha) + \Phi^{-1}(\beta)\}^2}{0.6^2} = 30.06$$

Under the approximation $\mathcal{I} \approx d/4$ the total number of failures to be observed is $d_f = 4\,\mathcal{I}_f = 120.2$.

Design of the Oropharynx trial

For a one-sided test with up to 5 analyses, we could use a standard design created for equally spaced information levels.



However, increments in information between analyses will be unequal and unpredictable.

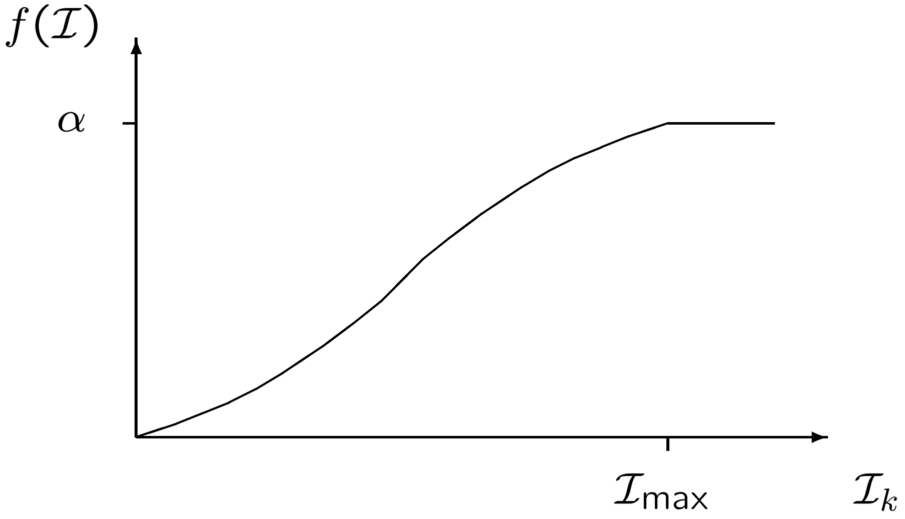
This leads to consideration of an “error spending” design.

6. Error spending tests

Lan & DeMets (1983, *Biometrika*) presented two-sided tests which “spend” type I error as a function of observed information.

Maximum information design:

Error spending function $f(\mathcal{I})$

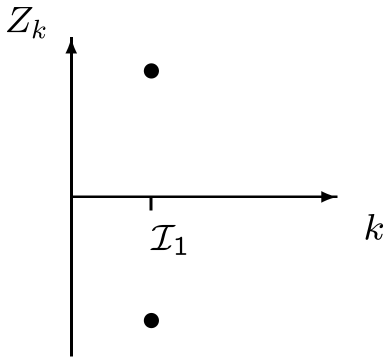


Set the boundary at analysis k to give cumulative Type I error $f(\mathcal{I}_k)$.

Accept H_0 if $\mathcal{I}_{\text{max}}$ is reached without rejecting H_0.

Error spending tests

Analysis 1:

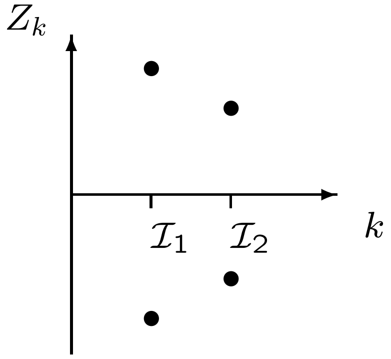


Observed information $\mathcal{I}_1$.

Reject H_0 if $|Z_1| > c_1$ where

$$Pr_{\theta=0}\{|Z_1| > c_1\} = f(\mathcal{I}_1).$$

Analysis 2:



Cumulative information $\mathcal{I}_2$.

Reject H_0 if $|Z_2| > c_2$ where

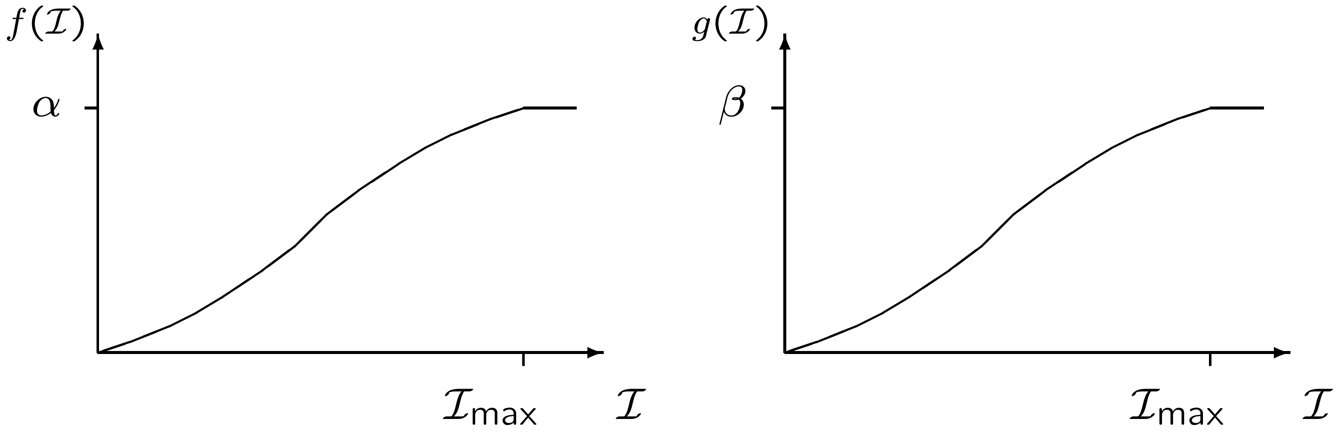
$$Pr_{\theta=0}\{|Z_1| < c_1, \, |Z_2| > c_2\}$$

$$= f(\mathcal{I}_2) - f(\mathcal{I}_1).$$

etc.

One-sided error spending tests

For a one-sided test, define $f(\mathcal{I})$ and $g(\mathcal{I})$ to specify how type I and type II error probabilities are spent as a function of observed information.



At analysis k, set boundary values $(a_k,\, b_k)$ so that

$$Pr_{\theta=0}\,\{\text{Reject } H_0 \text{ by analysis } k\} \;=\; f(\mathcal{I}_k),$$

$$Pr_{\theta=\delta}\,\{\text{Accept } H_0 \text{ by analysis } k\} \;=\; g(\mathcal{I}_k).$$

Power family of error spending tests:

$$f(\mathcal{I}) \text{ and } g(\mathcal{I}) \;\propto\; (\mathcal{I}/\mathcal{I}_{\max})^{\rho}.$$

One-sided error spending tests

1. Values $\{a_k, b_k\}$ are easily computed using iterative formulae of McPherson, Armitage & Rowe (1969, *JRSS, A*).

2. Computation of $(a_k,\, b_k)$ does **not** depend on future information levels, $\mathcal{I}_{k+1}, \mathcal{I}_{k+2}, \ldots$.

3. In a “maximum information design”, the study continues until the boundary is crossed or an analysis is reached with $\mathcal{I}_k \geq \mathcal{I}_{\mathsf{max}}$.

4. Special treatment of “over-running” or “under-running” protects the type I error rate.

5. The value of $\mathcal{I}_{\mathsf{max}}$ should be chosen so that boundaries converge at the final analysis under a typical sequence of information levels, e.g., $\mathcal{I}_k = (k/K)\,\mathcal{I}_{\mathsf{max}}, \quad k = 1, \ldots, K.$

A one-sided error spending design for the Oropharynx trial

Specification:

one-sided test of H_0: $\theta \leq 0$ vs $\theta > 0$,

type I error probability $\alpha = 0.05$,

power $1 - \beta = 0.95$ at $\theta = \ln(\lambda) = 0.6$.

At the design stage, assume $K = 5$ equally spaced information levels.

Use a power-family test with $\rho = 2$, i.e., spending error $\propto (\mathcal{I}/\mathcal{I}_{\text{max}})^2$.

Information of a fixed sample test is inflated by a factor $R(K, \alpha, \beta, \rho) = 1.101$ (JT, Table 7.6).

So, we require $\mathcal{I}_{\text{max}} = 1.101 \times 30.06 = 33.10$, which needs a total of $4 \times 33.10 = 132.4$ deaths.

Data and boundaries for the Oropharynx trial

We construct error spending boundaries for the information levels actually observed.

This gives boundary values $(a_1, b_1), \dots, (a_5, b_5)$ for the standardised statistics $Z_1, \dots, Z_5$.

k	*Number entered*	*Number of deaths*	$\mathcal{I}_k$	a_k	b_k	Z_k
1	83	27	5.43	-1.60	3.00	-1.04
2	126	58	12.58	-0.37	2.49	-1.00
3	174	91	21.11	0.63	2.13	-1.21
4	195	129	30.55	1.51	1.81	-0.73
5	195	142	33.28	1.73	1.73	-0.87

This rule would have led to termination at the 2nd analysis.

NB: A maximum information design is implicitly adaptive. More subjects must be recruited or existing subjects followed up further until the target $\mathcal{I}_{\text{max}}$ is reached.

7. Changing the power requirement in mid-study

Suppose a study is designed to attain power $1 - \beta$ at effect size $\theta = \delta$.

At an intermediate stage, results show:

$\widehat{\theta}_1$ is positive but smaller than the hoped for effect δ,

H_0 is unlikely to be rejected (low conditional power),

however, the magnitude of $\widehat{\theta}_1$ is clinically meaningful.

It appears that the original target effect size δ was over-optimistic — a larger sample size would have been better.

Can this trial be “rescued” ?

External changes

Suppose external information (e.g., concerning a competing treatment or changes in the manufacturer's circumstances) imply it is now worthwhile to find a smaller treatment effect than δ.

Alternately, the same change in objective may be motivated by, say, safety information internal to the current study.

Interim data have been seen, so the investigators do know the current estimate $\widehat{\theta}_1$.

Can the trial be enlarged without loss of credibility?

Methods for unplanned re-design

Bauer & Köhne (1994, *Biometrics*) provide a framework — but you must state at the outset that you are following this non-standard approach.

L. Fisher (1998, *Statistics in Medicine*) proposes the method of “variance spending” which can be introduced without prior planning.

Cui, Hung & Wang, (1999, *Biometrics*) present a method for re-specifying group sizes — and maximum sample size — in a group sequential design.

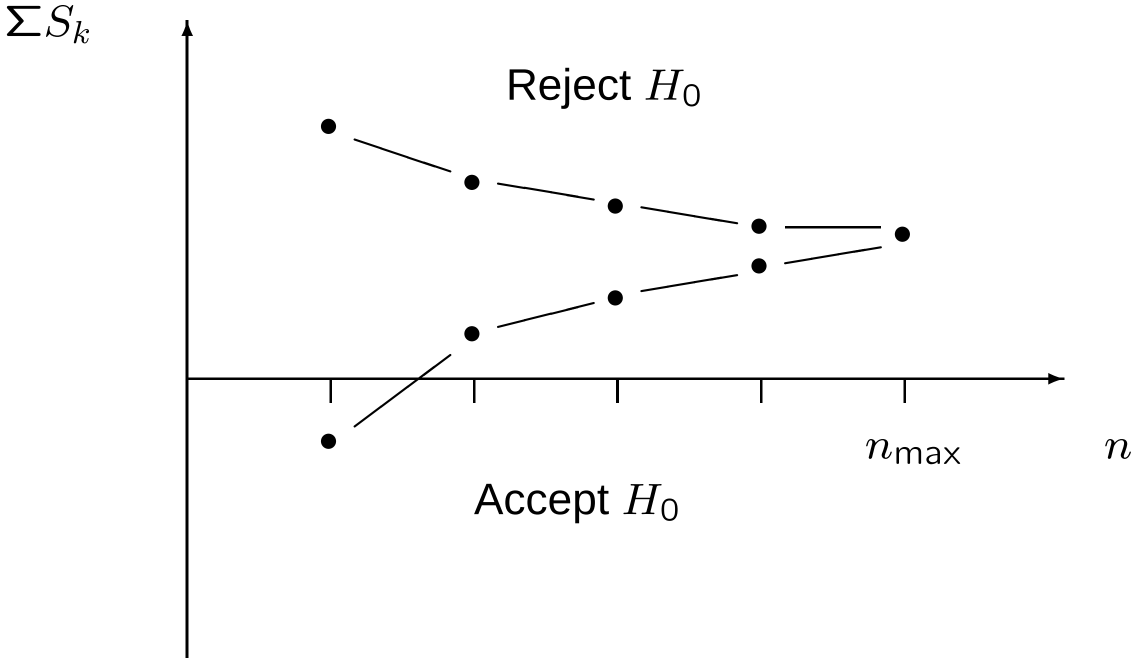
Müller & Schäfer (2001, *Biometrics*) give a general methodology based on preserving the conditional type I error probability. This has the flexibility to be used in error spending designs

Example: Müller-Schäfer adaptation to external information

Original error spending design:

To test H_0: $\theta = 0$ with type I error rate 0.025 and power 0.9 at $\theta = \delta$.

5 group error spending test, ρ-family with $\rho = 3$, early stopping to accept or reject H_0.



$$n_{\max} = 11.0/\delta^2, \quad \text{cf fixed sample size, } n_f = 10.5/\delta^2.$$

Design modification in response to external information

At analysis 2, suppose external factors prompt interest in lower θ values and we now aim for power 0.9 at $\delta/2$ rather than δ.

On observing $S_2 = s_2$ in the continuation region:

Calculate conditional type I error rate

$$\tilde{\alpha}(s_2) = P_{\theta=0}\,\{\text{Reject } H_0 \,|\, S_2 = s_2\}.$$

Set up a new design (ρ-family, $\rho = 3$) based on future observations with 3 further analyses and

type I error $\tilde{\alpha}(s_2)$, power 0.9 at $\delta/2$.

The future group sizes depend on s_2 through $\tilde{\alpha}(s_2)$.

Figure 1. Power functions of original error spending test and Müller-Schäfer adaptive test.

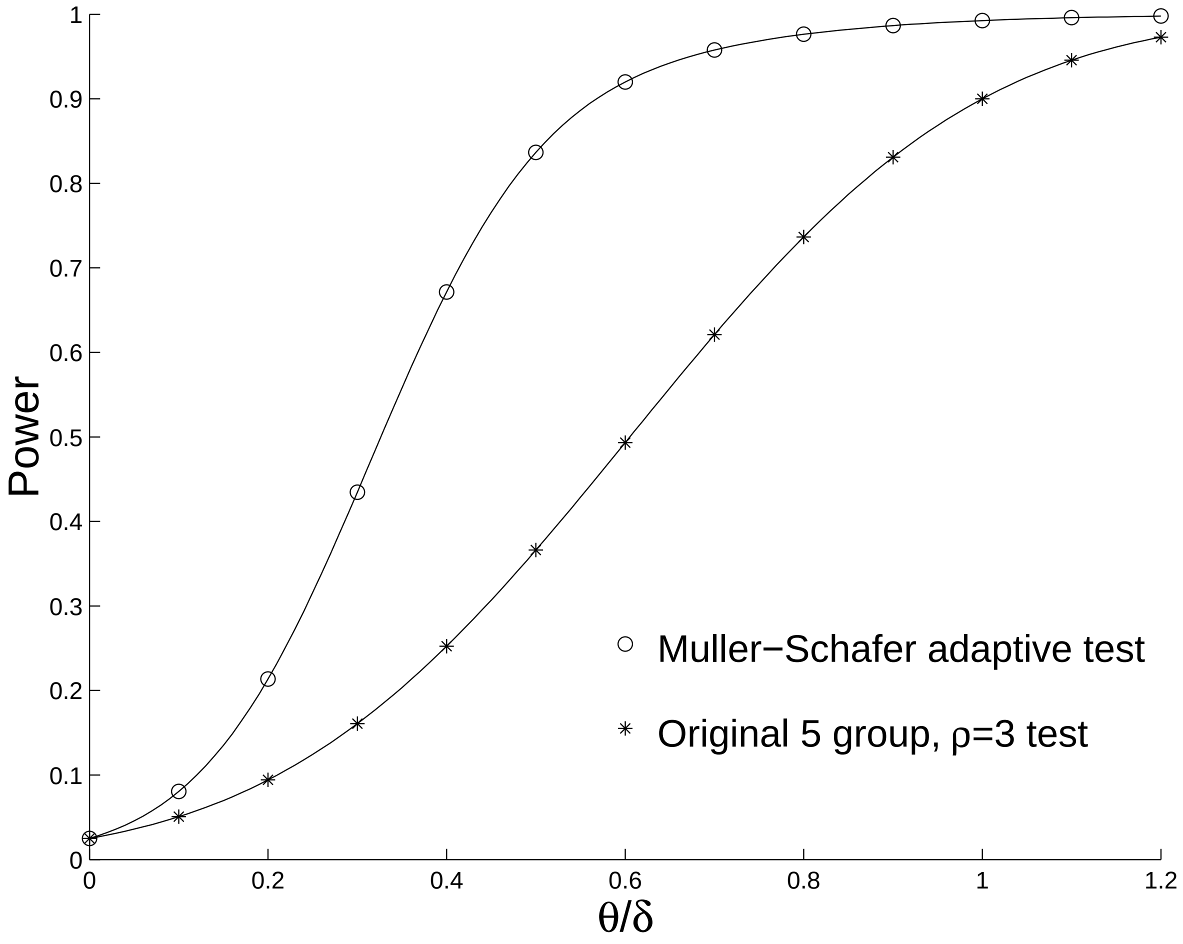
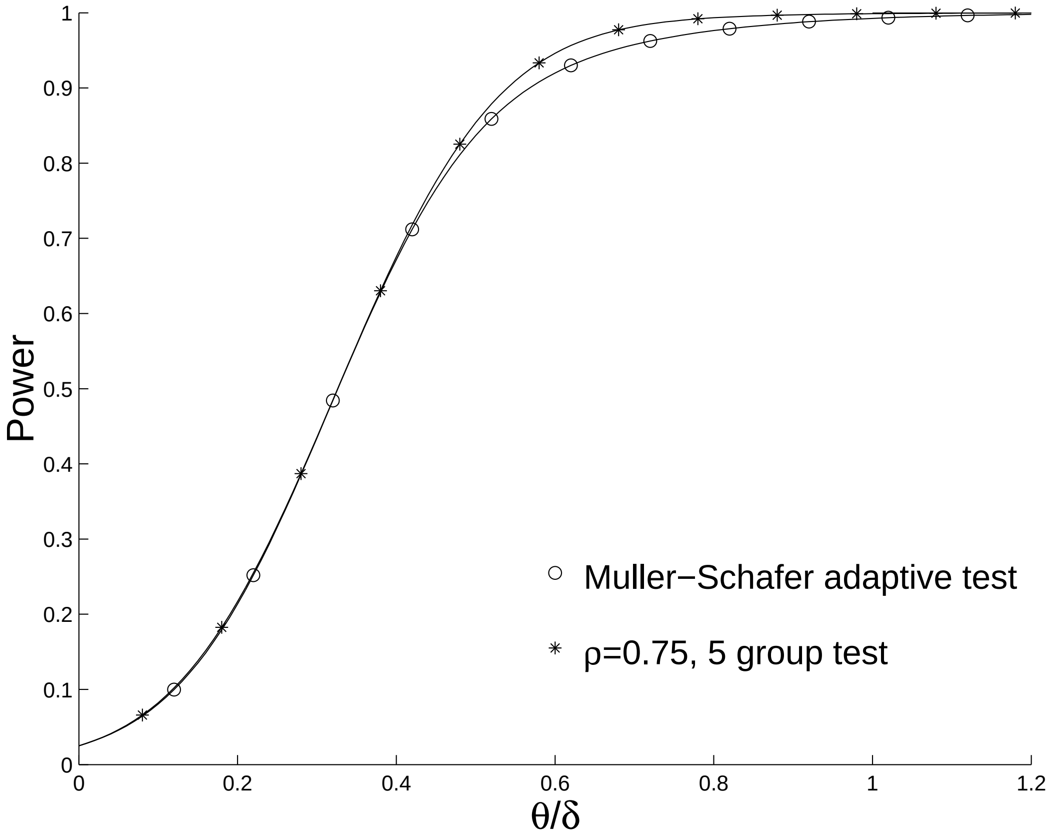
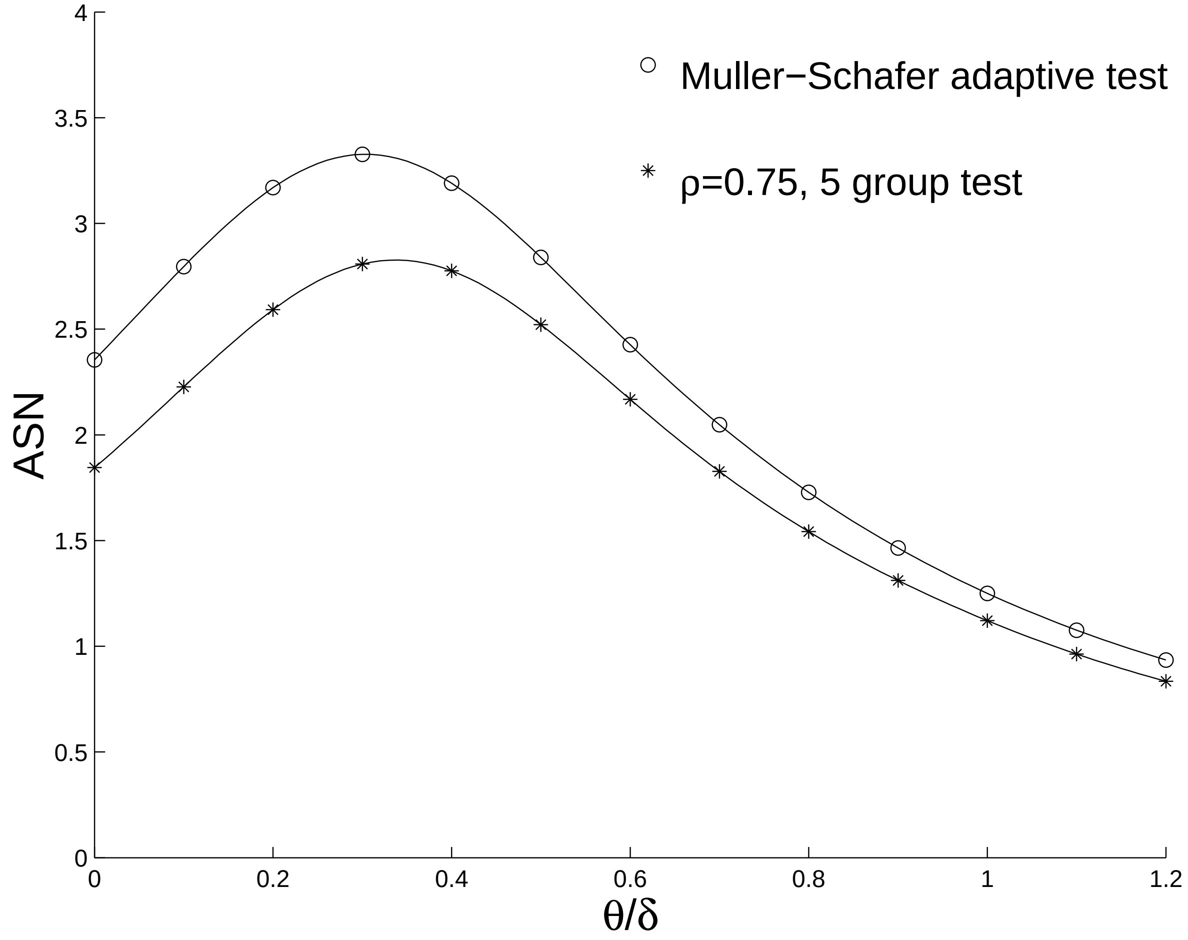


Figure 2. Power functions of Müller-Schäfer adaptive test and a non-adaptive 5 group test with power 0.9 at $\theta = 0.54\,\delta$.



The non-adaptive test is a ρ-family error spending test with $\rho = 0.75$ and interim analyses at 0.1, 0.2, 0.45 and 0.7 of the maximum sample size.

Figure 3. Average Sample Number (ASN) curves of Müller-Schäfer adaptive test and matched non-adaptive test.



ASN scale is in multiples of the original fixed sample size, n_f.

Note: not designing for the final objective from the outset incurs a penalty of a larger sample size.

Conclusion

Good sample size re-estimation methods are available to deal with nuisance parameters.

Care must be taken to avoid revealing the estimated effect size when it is meant to remain blinded. Considering sample size issues at interim analyses of a group sequential test can (at least partly) alleviate this problem.

Methods are available to modify sample size to meet a new power criterion mid-way through a study. However, these come at a price. Where possible, it is much better to identify the correct objective at the design stage.