Adaptivity in Sequential Designs:

Old and New

Christopher Jennison,

Dept of Mathematical Sciences,

University of Bath, UK

http://www.bath.ac.uk/∼mascj

RSS & PSI, London,

6 October 2004

Plan of talk

1. Interim monitoring of clinical trials

   Adapting to observed data

2. Distribution theory, the role of “information”

3. Error-spending tests

   Adapting to unpredictable information

   Adapting to nuisance parameters

4. Most efficient group sequential tests

   Adapting optimally to observed data

5. More recent adaptive proposals

1. Interim monitoring of clinical trials

It is standard practice to monitor progress of clinical trials for reasons of *ethics*, *administration* (accrual, compliance) and *economics*.

Special methods are needed since multiple looks at accumulating data can lead to over-interpretation of interim results

Methods developed in manufacturing production were first transposed to clinical trials in the 1950s.

Traditional sequential methods assumed continuous monitoring of data, whereas it is only practical to analyse a clinical trial on a small number of occasions.

The major step forward was the advent of *Group Sequential* methods in the 1970s.

Pocock's repeated significance test (1977)

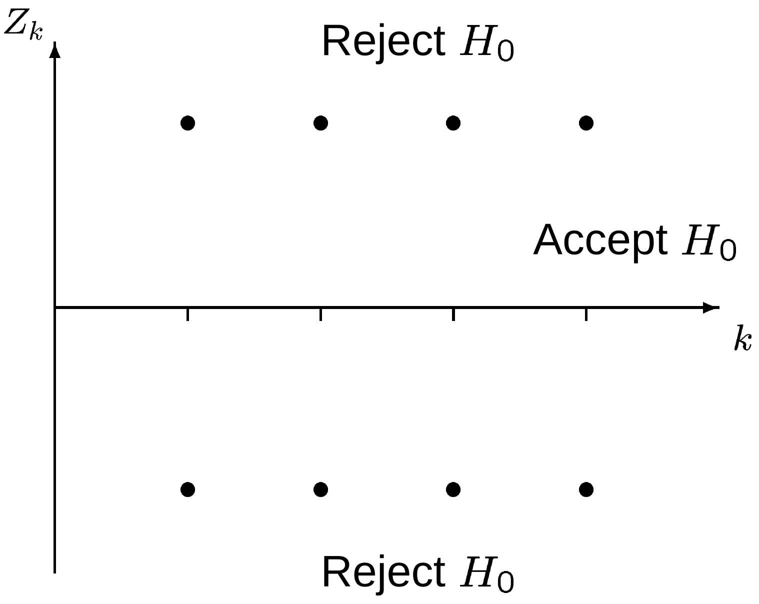
To test H_0: $\theta = 0$ against $\theta \neq 0$, where θ represents the treatment difference.

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.

Choose c to give overall type I error rate = α.



Types of hypothesis testing problems

Two-sided test:

testing H_0: $\theta = 0$ against $\theta \neq 0$.

One-sided test:

testing H_0: $\theta \leq 0$ against $\theta > 0$.

Equivalence tests:

one-sided — to show treatment A is as good as treatment B, within a margin δ.

two-sided — to show two treatment formulations are equal within an accepted tolerance.

Types of early stopping

1. Stopping **to reject** H_0: *No treatment difference*

- Allows progress from a positive outcome
- Avoids exposing further patients to the inferior treatment
- Appropriate if no further checks are needed on treatment safety or long-term effects.

2. Stopping **to accept** H_0: *No treatment difference*

- Stopping “ for futility” or “abandoning a lost cause”
- Saves time and effort when a study is unlikely to lead to a positive conclusion.

One-sided tests

To look for superiority of a new treatment, test

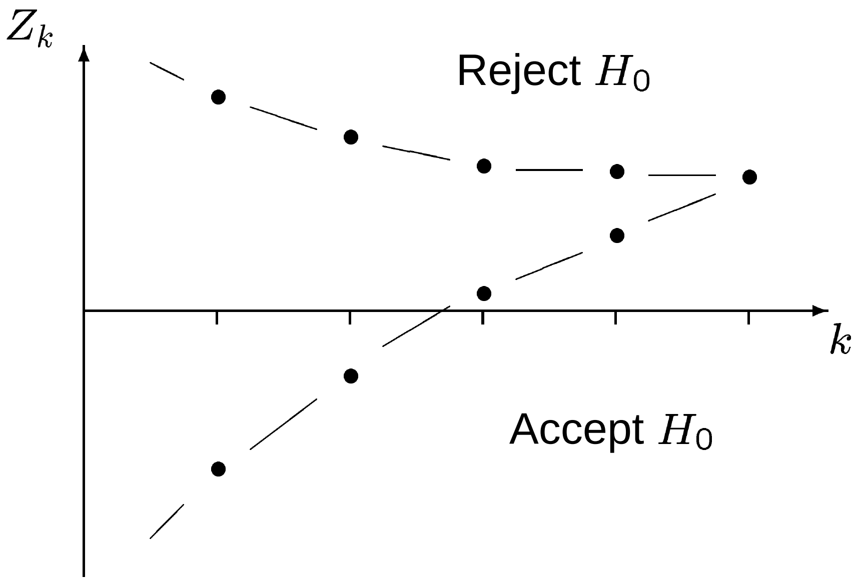
$$H_0\colon \theta \leq 0 \quad \text{against} \quad \theta > 0,$$

requiring

$$Pr\{\text{Reject } H_0 \,|\, \theta = 0\} = \alpha,$$

$$Pr\{\text{Reject } H_0 \,|\, \theta = \delta\} = 1 - \beta.$$

A typical boundary is:



E(Sample size) $\sim$ 50 to 70% of the fixed sample size

— ***adapting to data***, stopping when a decision is possible.

2. Joint distribution of parameter estimates

Let $\widehat{\theta}_k$ be the estimate of the parameter of interest, θ, based on data at analysis k.

The information for θ at analysis k is

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

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

In very 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

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

Canonical joint distribution of z-statistics

In a test of H_0: $\theta = 0$, the *standardised statistic* at analysis k is

$$Z_k = \frac{\widehat{\theta}_k}{\sqrt{\text{Var}(\widehat{\theta}_k)}} = \widehat{\theta}_k \sqrt{\mathcal{I}_k}.$$

For this,

$(Z_1, \ldots, Z_K)$ is multivariate normal,

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

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

Canonical joint distribution of score statistics

The *score statistics* $S_k = Z_k\sqrt{\mathcal{I}_k}$, are also multivariate normal with

$$S_k \sim N(\theta\,\mathcal{I}_k, \mathcal{I}_k), \quad k = 1, \ldots, K.$$

The score statistics possess the “independent increments” property,

$$\text{Cov}(S_k - S_{k-1}, S_{k'} - S_{k'-1}) = 0 \quad \text{for } k \neq k'.$$

It can be helpful to know the score statistics behave as Brownian motion with drift θ observed at times $\mathcal{I}_1, \ldots, \mathcal{I}_K$.

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).

So, we have the theory to support general comparisons, including adjustment for covariates if required.

Survival data

The canonical joint distributions also arise for:

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

> 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.

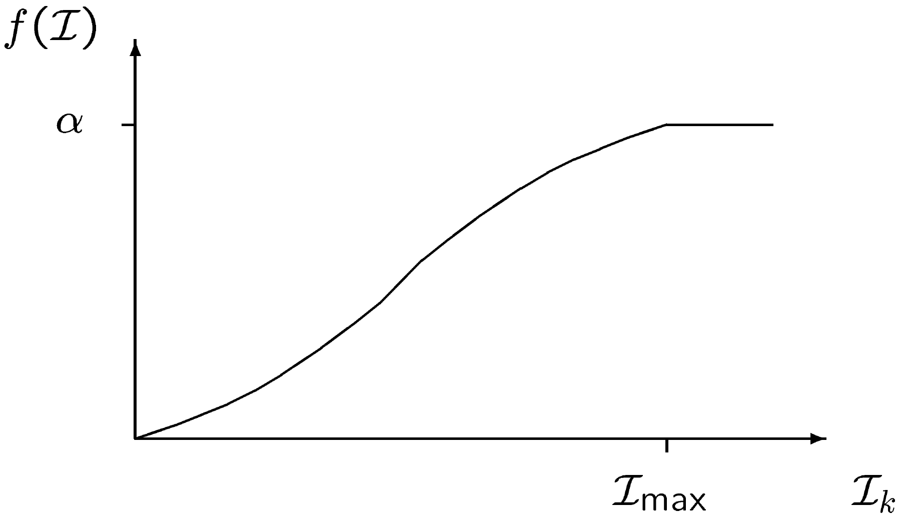
Special types of group sequential test are needed to handle unpredictable and unevenly spaced information levels.

3. Error spending tests

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

Maximum information design:

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

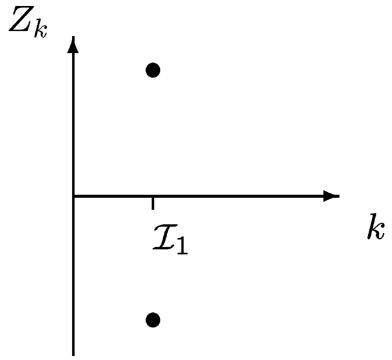


Set the boundary at analysis k to give cumulative type I error probability $f(\mathcal{I}_k)$.

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

Implementing error spending tests

Analysis 1:

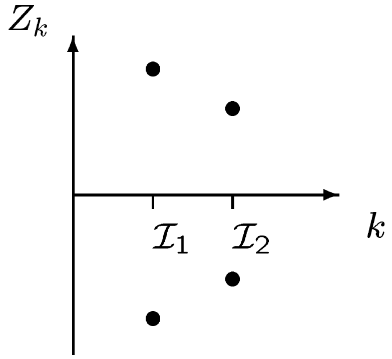


Observed information $\mathcal{I}_1$.

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

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

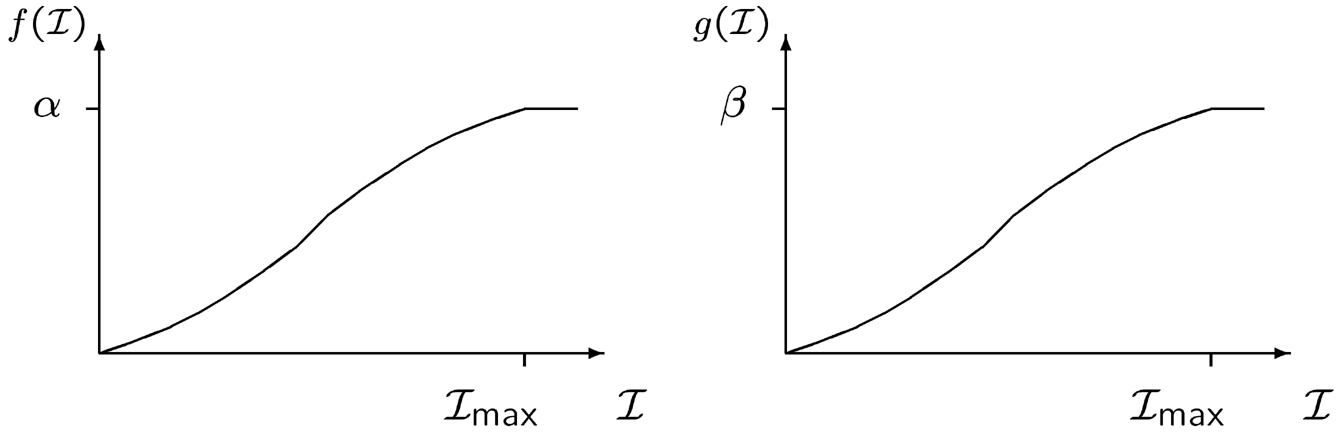
$$= f(\mathcal{I}_2) - f(\mathcal{I}_1).$$

etc.

Adapting to unpredictable information

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:

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

Implementing one-sided error spending tests

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

2. A “maximum information design” continues until a boundary is crossed or an analysis with $\mathcal{I}_k \geq \mathcal{I}_{\text{max}}$ is reached.

3. The value of $\mathcal{I}_{\text{max}}$ is chosen so that boundaries converge at the final analysis under a typical sequence of information levels, e.g.,

$$\mathcal{I}_k = (k/K)\,\mathcal{I}_{\text{max}}, \quad k = 1, \ldots, K.$$

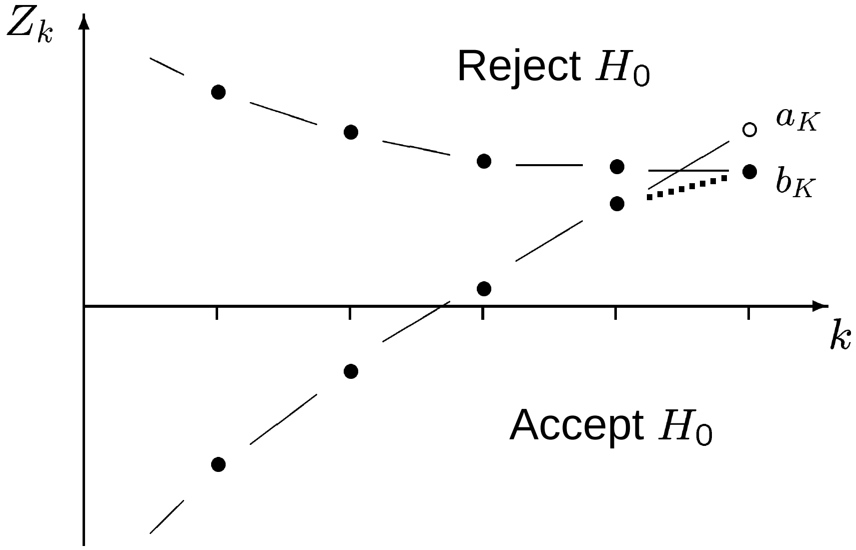
   For type I error rate α and power $1-\beta$ at $\theta = \delta$,

$$\mathcal{I}_{\text{max}} \;=\; R\,\frac{(z_\alpha + z_\beta)^2}{\delta^2},$$

   where R is the “inflation factor” for this design.

Over-running

If one reaches $\mathcal{I}_K > \mathcal{I}_{\text{max}}$, solving for a_K and b_K is liable to give $a_K > b_K$.



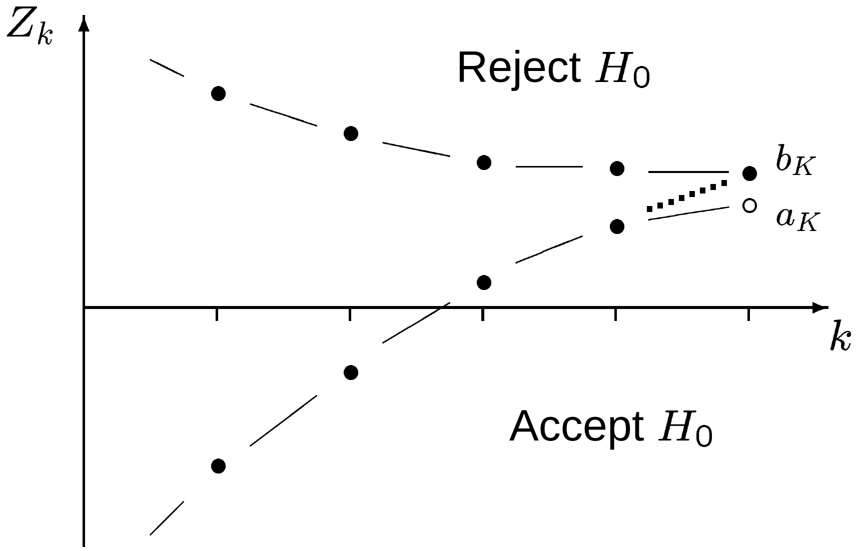
Keeping b_K as calculated guarantees type I error probability of exactly α.

So, reduce a_K to b_K — and gain extra power.

Over-running may also occur if $\mathcal{I}_K = \mathcal{I}_{\text{max}}$ but the information levels deviate from the equally spaced values (say) used in choosing $\mathcal{I}_{\text{max}}$.

Under-running

If a final information level $\mathcal{I}_K < \mathcal{I}_{\max}$ is imposed, solving for a_K and b_K is liable to give $a_K < b_K$.



Again, with b_K as calculated, the type I error probability is exactly α.

This time, increase a_K to b_K — and attained power will be a little below $1 - \beta$.

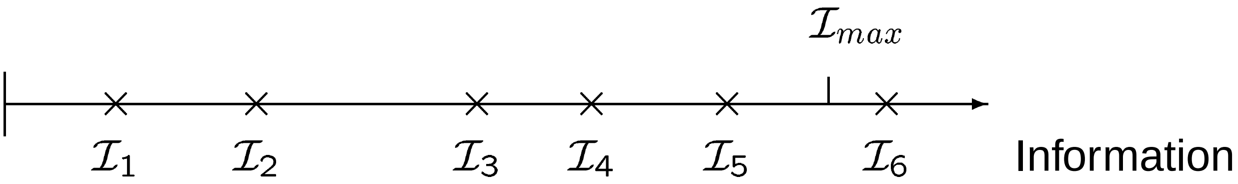
Error-spending designs and nuisance parameters

(1) Survival data, log-rank statistics

Information depends on the number of observed failures,

$$\mathcal{I}_k \;\approx\; \frac{1}{4}\{\text{Number of failures by analysis } k\}.$$

With analyses at fixed calendar times, continue until observed information reaches $\mathcal{I}_{\mathsf{max}}$.



If the overall failure rate is low or censoring is high, one may decide to extend the patient accrual period.

N.B. Changes affecting the sequence $\{\mathcal{I}_1, \mathcal{I}_2, \ldots\}$ can be based on observed information levels; they should **not** be influenced by the estimated treatment effect.

Error-spending designs and nuisance parameters

(2) Normal responses with unknown variance

In a two treatment comparison, a fixed sample test with type I error rate α and power $1-\beta$ at $\theta=\delta$ requires information

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

A group sequential design with inflation factor R needs maximum information $\mathcal{I}_{\text{max}} = R\mathcal{I}_f$.

The maximum required information is fixed — but the sample size needed to provide this level of information depends on the unknown variance σ^2.

Adapting to nuisance parameters

Adjusting sample size as variance is estimated

The information from n_A observations on treatment A and n_B on treatment B is

$$\mathcal{I} = \left\{ \left(\frac{1}{n_A} + \frac{1}{n_B} \right) \sigma^2 \right\}^{-1}.$$

Initially: Set maximum sample sizes to give information $\mathcal{I}_{\mathsf{max}}$ if σ^2 is equal to an initial estimate, σ_0^2.

As updated estimates of σ^2 *are obtained:* Adjust future group sizes so the final analysis has

$$\left\{ \left(\frac{1}{n_A} + \frac{1}{n_B} \right) \widehat{\sigma}^2 \right\}^{-1} = \mathcal{I}_{\mathsf{max}}.$$

NB, state $\mathcal{I}_{max}$ in the protocol, not initial targets for n_A and n_B.

At interim analyses, apply the error spending boundary based on observed (estimated) information.

4. Optimal group sequential tests

“Optimal” designs may be used directly — or they can serve as a benchmark for judging efficiency of designs with other desirable features.

Optimising a group sequential test:

Formulate the testing problem:

fix type I error rate α and power $1 - \beta$ at $\theta = \delta$,

fix number of analyses, K,

fix maximum sample size (information), if desired

Find the design which minimises average sample size (information) at one particular θ or averaged over several θ s.

Derivation of optimal group sequential tests

Create a Bayes decision problem with a prior on θ, sampling costs and costs for a wrong decision. Write a program to solve this Bayes problem by backwards induction (dynamic programming).

Search for a set of costs such that the Bayes test has the desired frequentist properties: type I error rate α and power $1-\beta$ at $\theta=\delta$.

This is essentially a Lagrangian method for solving a constrained optimisation problem — the key is that the unconstrained Bayes problem can be solved accurately and quickly.

Example of properties of optimal tests

One-sided tests, $\alpha = \beta = 0.05$, K analyses,

$\mathcal{I}_{max} = R\mathcal{I}_{fix}$, equal group sizes,

minimising $\{E_0(\mathcal{I}) + E_\delta(\mathcal{I})\}/2$.

Minimum values of $E_0(\mathcal{I})$ and $E_\delta(\mathcal{I})$,

stated as a percentage of $\mathcal{I}_{fix}$

	R					*Minimum*
K	1.01	1.05	1.1	1.2	1.3	*over R*
2	80.9	74.5	72.8	73.2	75.3	72.7 at 1.15
5	72.2	65.2	62.2	59.8	59.0	58.7 at 1.4
10	69.1	62.1	59.0	56.3	55.2	54.3 at 1.6
20	67.6	60.5	57.4	54.6	53.3	52.0 at 1.6

Note: $E(\mathcal{I}) \searrow$ as $K \nearrow$ but with diminishing returns,

$E(\mathcal{I}) \searrow$ as $R \nearrow$ up to a point.

Assessing families of group sequential tests

One-sided tests:

Pampallona & Tsiatis

Parametric family indexed by Δ,

boundaries for S_k involve $\mathcal{I}_k^{\Delta}$,

each Δ implies an "inflation factor" R such that

$$\mathcal{I}_{max} = R\mathcal{I}_{fix}.$$

Error spending, ρ-family

Error spent is proportional to $\mathcal{I}_k^{\rho}$,

ρ determines the inflation factor R.

Error spending, γ-family (Hwang et al, 1994)

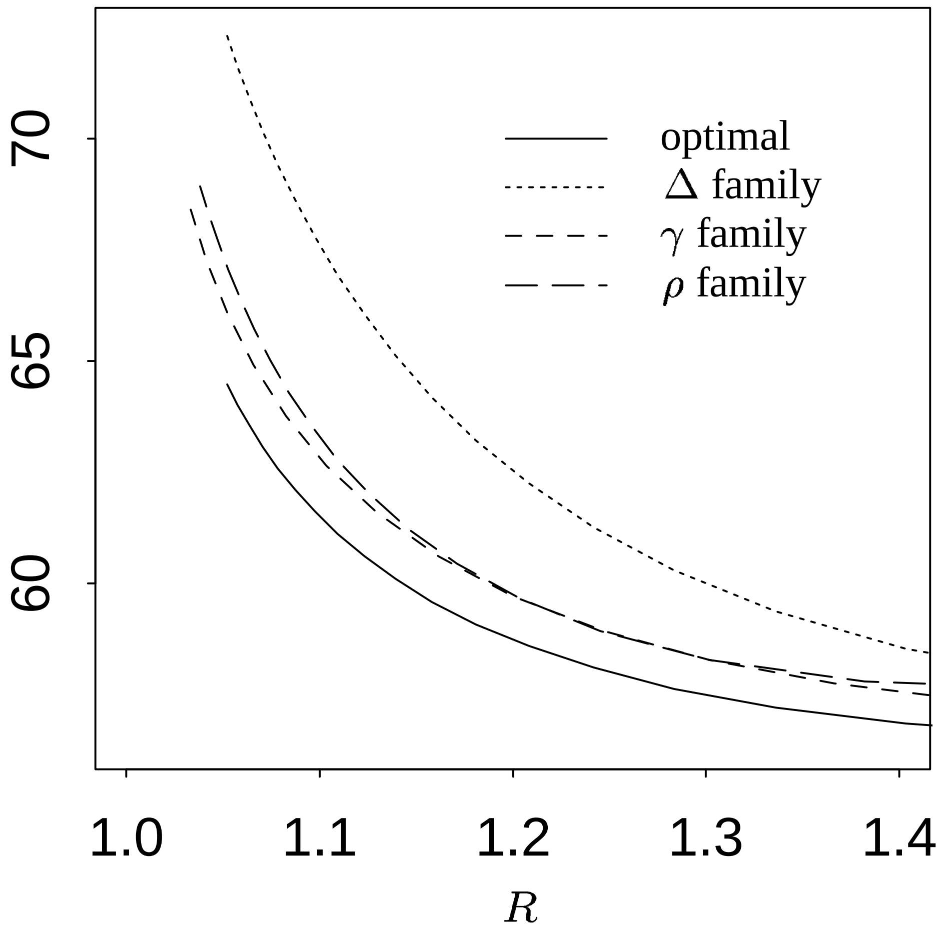
Error spent is proportional to

$$\frac{1 - e^{-\gamma\,\mathcal{I}_k/\mathcal{I}_{max}}}{1 - e^{-\gamma}}.$$

Families of tests

Tests with $K = 10$, $\alpha = 0.05$, $1 - \beta = 0.9$.

$\{E_0(\mathcal{I}) + E_\delta(\mathcal{I})\}/2$ as a percentage of $\mathcal{I}_{fix}$



Both error spending families are highly efficient but Pampallona & Tsiatis tests are sub-optimal.

Adapting optimally to observed data

Squeezing a little extra efficiency

Schmitz (1993) proposed group sequential tests in which group sizes are chosen adaptively:

Initially, fix $\mathcal{I}_1$,

observe $S_1 \sim N(\theta \mathcal{I}_1, \mathcal{I}_1)$,

then choose $\mathcal{I}_2$ as a function of S_1, observe S_2 where

$$S_2 - S_1 \sim N(\,\theta(\mathcal{I}_2 - \mathcal{I}_1), (\mathcal{I}_2 - \mathcal{I}_1)\,),$$

and so forth.

Specify sampling rule and stopping rule to achieve desired *overall* type I error rate and power.

Examples of “Schmitz” designs

To test H_0: $\theta = 0$ versus H_1: $\theta > 0$

with type I error rate $\alpha = 0.025$

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

Aim for low values of

$$\int E_\theta(N) f(\theta)\, d\theta,$$

where $f(\theta)$ is the density of a $N(\delta,\, \delta^2/4)$ distribution.

Constraints:

Maximum sample information $= 1.2 \times$ fixed sample information.

Maximum number of analyses $= K$.

Again, optimal designs can be found by solving related Bayes decision problems.

Examples of "Schmitz" designs

Optimal average $E(\mathcal{I})$ stated as a percentage of the fixed sample information.

K	*Optimal adaptive design (Schmitz)*	*Optimal non-adaptive, optimised group sizes*	*Optimal non-adaptive, equal group sizes*
2	72.5	73.2	74.8
3	64.8	65.6	66.1
4	61.2	62.4	62.7
6	58.0	59.4	59.8
8	56.6	58.0	58.3
10	55.9	57.2	57.5

Varying group sizes *adaptively* makes for a complex procedure and the efficiency gains are slight.

Adapting super-optimally to observed data

5. Recent adaptive methods

Bauer (1989) and Bauer & Köhne (1994) proposed mid-course design changes to one or more of

Treatment definition

Choice of primary response variable

Sample size:

— in order 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, i.e., $\widehat{\theta}$.

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.

- 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:

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

B & K: Major design changes before part 2

With major changes, the two parts are rather like separate studies in a drug development process, such as:

Phase IIb

Compare several doses and select the best.
Use a rapidly available endpoint (e.g., tumour response).

Phase III

Compare selected dose against control.
Use a long-term endpoint (e.g., survival).

Applying Fisher's combination test for P_1 and P_2 gives a meta-analysis of the two stages with a pre-specified rule.

Note: Each stage has its own null hypothesis and the overall H_0 is the intersection of these.

B & K: Minor design changes before part 2

With only minor changes, the form of responses in part 2 stays close to the original plan.

Bauer & Köhne’s method provides a way to handle this.

Or, an overall *score statistic* could be used:

Typically, one would derive separate score statistics from part 1 and part 2 patients, then add these together. For survival data, this is equivalent to stratification by “part”.

Given score statistics, one can use an error spending test with a maximum information design.

B & K: Nuisance parameters

Example. Normal response with unknown variance, σ^2.

Aiming for type I error rate α and power $1-\beta$ at $\theta = \delta$, the necessary sample size depends on σ^2.

One can choose the second part's sample size to meet this power requirement assuming variance is equal to s_1^2, the estimate from part 1.

Taking P_1 and P_2 from t-tests, these are independent $U(0,\,1)$ under H_0 — exactly.

Other methods:

Many "internal pilot" designs are available.

Error spending designs using estimated information (based on s^2) can be used.

The two-stage design of Stein (1945) attains both type I error and power precisely!

B & K: External factors or internal, secondary information

At an interim stage suppose, for reasons not concerning the primary endpoint, investigators wish to achieve power $1-\beta$ at $\theta=\tilde{\delta}$ rather than $\theta=\delta$ ($\tilde{\delta}<\delta$).

If this happens after part 1 of a B & K design, the part 2 sample size can be increased, e.g., to give conditional power $1-\beta$ at $\theta=\tilde{\delta}$.

Unplanned re-design

Recent work shows the same can be done within a fixed sample or group sequential design by

preserving conditional type I error rate under $\theta=0$,

ensuring conditional power $1-\beta$ at $\theta=\tilde{\delta}$

— see Denne (2001) or Müller & Schäfer (2001).

B & K: Responding to $\widehat{\theta}$, an estimate of the primary endpoint

We have seen this in methods where design changes are made to attain a certain conditional power.

Elsewhere, motivation may be:

- to rescue an under-powered study,
- a “wait and see” approach to choosing a study’s power requirement,
- trying to be efficient.

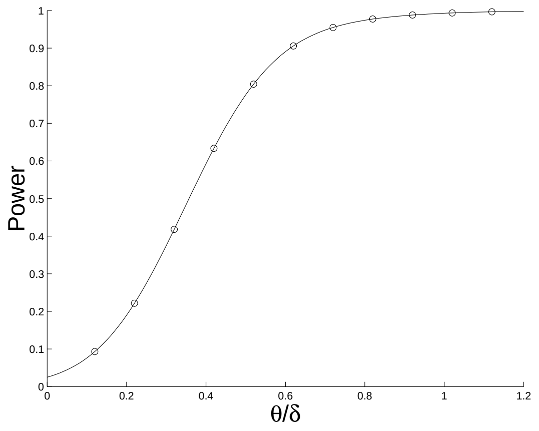
Many methods have been proposed and one can use the conditional type I error approach for unplanned re-design.

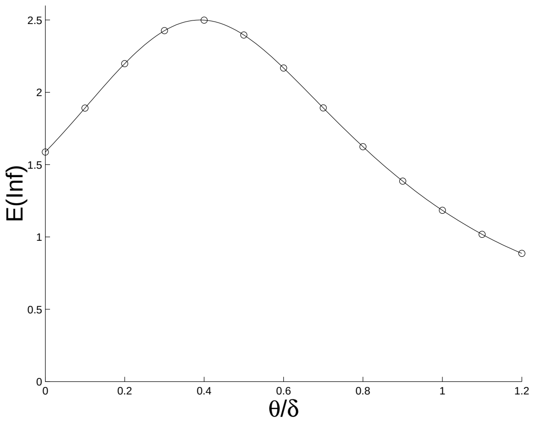
It is good to be able to rescue a poorly designed study.

Group sequential tests base the decision for early stopping on $\widehat{\theta}$. Optimal tests do this optimally!

Re-design based on $\widehat{\theta}$.

Any adaptive scheme has an *overall* power function and expected sample size or $E(\mathit{Inf})$ function.





These combine conditional properties in just the right way.

Before using an adaptive test check that its $E(\mathit{Inf})$ function is acceptably low for the attained power.

Improving on a fixed sample test is a minimal step. Compare with well chosen group sequential tests.

The ρ-family of error spending tests provides close to optimal designs. ***Why look further?***

Conclusions

Error Spending tests using Information Monitoring can adapt to

- unpredictable information levels,
- nuisance parameters,
- observed data, i.e., efficient stopping rules.

Methods preserving the conditional type I error rate allow re-design in response to external developments or internal evidence from secondary endpoints.

Recently proposed adaptive methods are appropriate when re-design is on a large scale.

They facilitate re-sizing for nuisance parameters.

They support re-sizing to rescue an under-powered study.

They allow an on-going approach to study design.

They will not improve on the efficiency of “standard” Group Sequential Tests — and can be substantially inferior.