Interim Monitoring of Clinical Trials:

Decision Theory, Dynamic Programming

and Optimal Stopping

Christopher Jennison

Department of Mathematical Sciences,

University of Bath, UK

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

University of Kent

December 2009

Plan of talk

1. Monitoring clinical trials

2. Sequential distribution theory

3. An optimal stopping problem

4. Numerical evaluation of stopping boundaries

5. Finding optimal group sequential designs

6. Related problems:

    Adaptive choice of group sizes

    Testing for either superiority or non-inferiority

    Trials with delayed response

1. Monitoring clinical trials

A clinical trial is run to compare a new treatment with an existing treatment or placebo.

As the trial progresses, a Data and Safety Monitoring Board (DSMB) monitors patient recruitment, treatment administration, and the responses observed at interim points.

The DSMB can take actions in view of safety variables or secondary endpoints, for example, to drop a treatment arm with a high dose level if this appears unsafe.

Response on the primary endpoint may indicate early termination of the study is desirable, for either a positive or negative conclusion.

The need for special methods

Multiple looks at accumulating data can lead to over-interpretation of interim results.

Armitage et al. (*JRSS, A*, 1969) report the overall type I error rate when applying repeated significance tests at $\alpha = 0.05$ to accumulating data:

Number of tests	*Error rate*
1	0.05
2	0.08
3	0.11
5	0.14
10	0.19

Clearly, a different approach is needed to avoid inflation of the type I error rate.

Formulating the problem

Let θ denote the "effect size", a measure of the improvement in the new treatment over the standard.

We shall test the null hypothesis H_0: $\theta \leq 0$ against the alternative $\theta > 0$.

Then, rejecting H_0 allows us to conclude the new treatment is better than the standard.

We allow type I error probability α for rejecting H_0 when it is actually true.

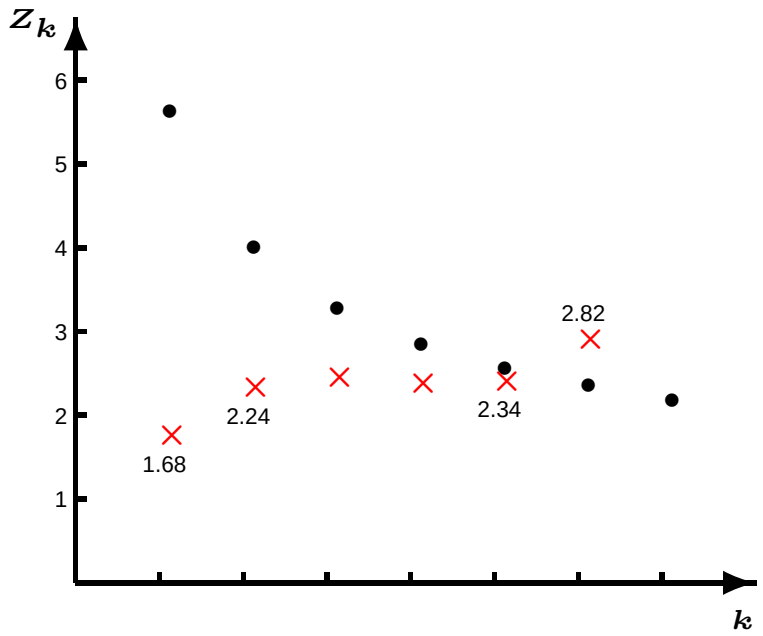
We specify power $1 - \beta$ for the probability of (correctly) rejecting H_0 when $\theta = \delta$. Here, δ is, typically, the minimal clinically significant treatment difference.

The trial design, including the method of analysis and stopping rule, must be set up to attain these error rates.

An early example: The BHAT trial

DeMets et al. (*Controlled Clinical Trials*, 1984) report on the Beta-Blocker Heart Attack Trial, which compared propanolol with placebo.

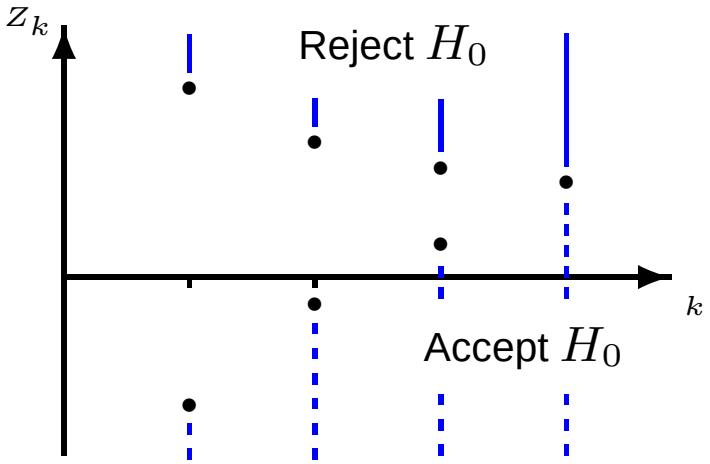
An “O’Brien and Fleming” stopping boundary was defined with overall type I error probability 0.025.



The trial stopped after the 6th of 7 planned analyses.

Group sequential tests: Stopping for futility

Adding a lower boundary allows stopping when there is little chance of a positive conclusion.



Rosner & Tsiatis (*Statistics in Medicine*, 1989) carried out retrospective analyses of 72 cancer studies of the U.S. Eastern Co-operative Oncology Group.

If group sequential stopping rules had been applied, early stopping (mostly for futility, i.e., to accept H_0) could have occurred in around 80% of cases.

Requirements for clinical trial designs

Regulatory bodies recommend group sequential designs to protect subjects in a clinical trial and produce results as efficiently as possible.

We need designs which:

Achieve specified type I error rate and power,

Stop early, on average, under key parameter values,

Can be applied to a variety of response types.

We shall present distribution theory which shows that a common set of methods can be applied to many data types.

To define efficient tests, we shall formulate and solve an optimal stopping problem.

2. Sequential distribution theory

Our interest is in the parameter for the treatment effect, θ.

Let $\hat{\theta}_k$ denote the estimate of θ based on data at analysis k.

The information for θ at analysis k is

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

Canonical joint distribution of $\hat{\boldsymbol{\theta}}_{\mathbf{1}}, \ldots, \hat{\boldsymbol{\theta}}_{\boldsymbol{K}}$

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

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

and

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

Sequential distribution theory

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

θ *a single normal mean,*

$\theta = \mu_A - \mu_B,$ *comparing two normal means.*

The results also apply when θ is a parameter in:

a general normal linear model,

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

a Cox proportional hazards regression model for survival data.

Thus, theory supports general comparisons, including:

crossover studies, analysis of longitudinal data, covariate adjustment.

Explanation of the canonical joint distribution

The special correlation structure applies to all efficient, or asymptotically efficient, unbiased estimators.

Proof

Suppose $\text{Var}(\hat{\theta}_2) \neq \text{Cov}(\hat{\theta}_1, \hat{\theta}_2)$.

The estimator at the second analysis, $\tilde{\theta}_2 = \hat{\theta}_2 + \epsilon\,(\hat{\theta}_2 - \hat{\theta}_1),$ has expectation θ and variance

$$\text{Var}(\hat{\theta}_2) + 2\,\epsilon\,\text{Cov}(\hat{\theta}_2, \hat{\theta}_2 - \hat{\theta}_1) + \epsilon^2\,\text{Var}(\hat{\theta}_2 - \hat{\theta}_1).$$

For small ϵ of opposite sign to $\text{Cov}(\hat{\theta}_2, \hat{\theta}_2 - \hat{\theta}_1) = \{\text{Var}(\hat{\theta}_2) - \text{Cov}(\hat{\theta}_1, \hat{\theta}_2)\}$, we have

$$\text{Var}(\tilde{\theta}_2) < \text{Var}(\hat{\theta}_2),$$

contradicting the assumption that $\hat{\theta}_2$ is efficient.

Canonical joint distribution of z-statistics

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

$$Z_k \;=\; \frac{\hat{\theta}_k}{\sqrt{\mathsf{Var}(\hat{\theta}_k)}} \;=\; \hat{\theta}_k \sqrt{\mathcal{I}_k}.$$

For this, the distribution theory for $\hat{\theta}_1, \ldots, \hat{\theta}_K$ implies

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

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

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

Canonical joint distribution of score statistics

The general theory also implies that *score statistics*, $S_k = Z_k\sqrt{\mathcal{I}_k}$, are 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 that the score statistics behave as Brownian motion with drift θ observed at times $\mathcal{I}_1, \ldots, \mathcal{I}_K$.

3. An optimal stopping problem

Consider a trial designed to test H_0: $\theta \leq 0$ vs $\theta > 0$, with:

Type I error rate α,

Power $1 - \beta$ at $\theta = \delta$,

Up to K analyses.

A fixed sample test needs information

$$\mathcal{I}_{fix} = \{\Phi^{-1}(\alpha) + \Phi^{-1}(\beta)\}^2/\delta^2.$$

We set the maximum information to be

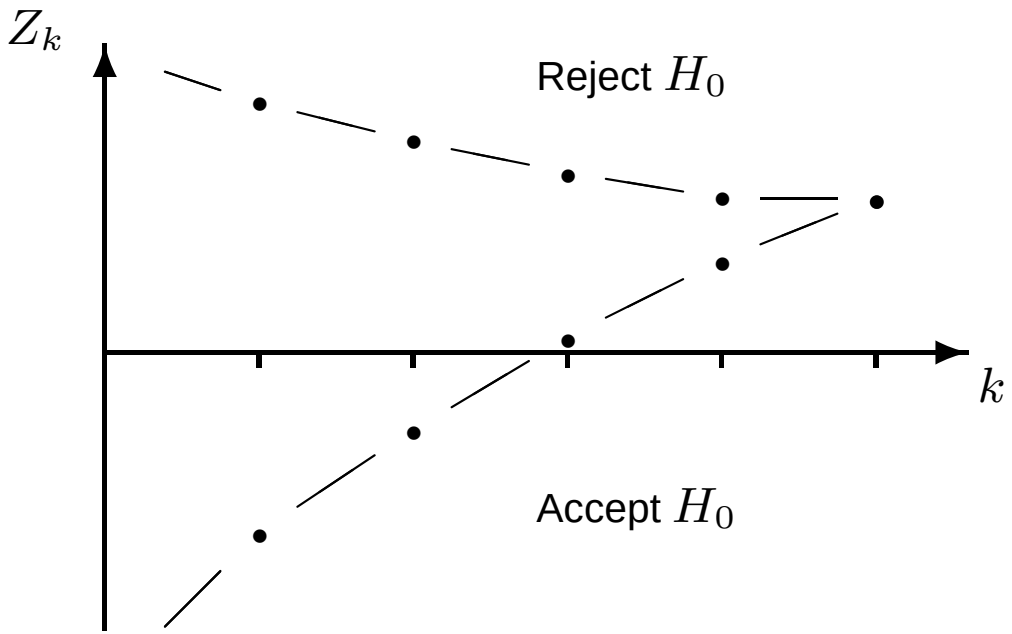
$$\mathcal{I}_{max} = R\,\mathcal{I}_{fix},$$

where $R > 1$, with equal increments between analyses, so

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

Optimal group sequential tests

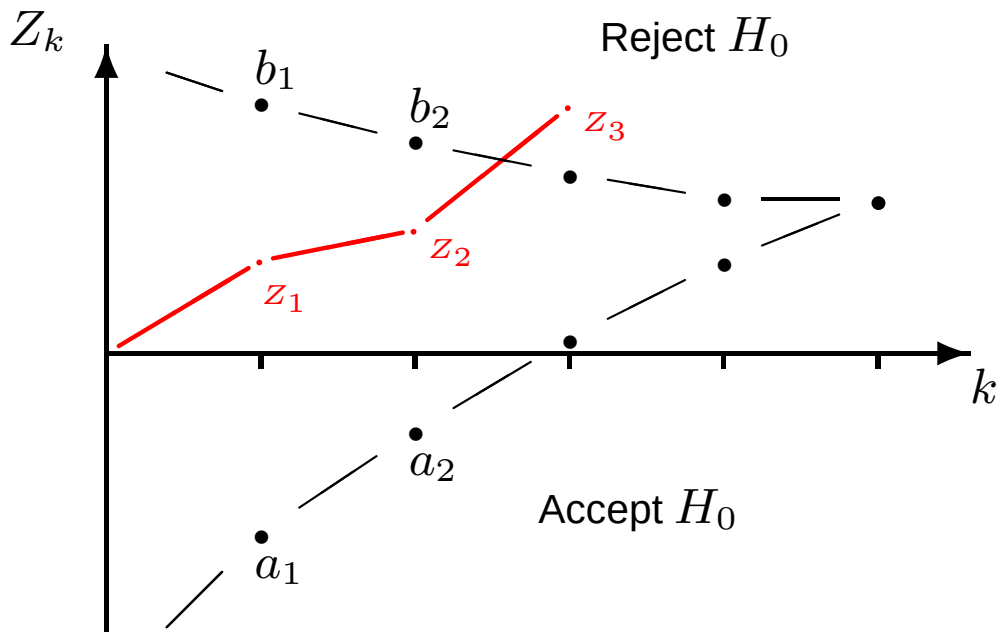
The error rates impose two constraints on the $2K-1$ boundary points — leaving a high dimensional space of possible boundaries.



We shall look for a boundary with an optimality property, specifically, minimising

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

4. Computations for group sequential tests



We need to be able to calculate the probabilities of basic events such as

$$a_1 < Z_1 < b_1, \quad a_2 < Z_2 < b_2, \quad Z_3 > b_3.$$

Combining such probabilities gives key properties, such as $Pr_\theta\{\text{Reject } H_0\}$, etc.

Computations for group sequential tests

For a one-sided test with K analyses, define the events

$$\mathcal{A}_k = \{a_1 < Z_1 < b_1,\ \ldots\ , a_{k-1} < Z_{k-1} < b_{k-1},\ Z_k < a_k\}$$

and

$$R_k = \{a_1 < Z_1 < b_1,\ \ldots\ , a_{k-1} < Z_{k-1} < b_{k-1},\ Z_k > b_k\}.$$

Then

$$\begin{aligned} Pr\{\text{Accept } H_0\} &= Pr\{\mathcal{A}_1\} + \ldots + Pr\{\mathcal{A}_K\}, \\ Pr\{\text{Reject } H_0\} &= Pr\{\mathcal{R}_1\} + \ldots + Pr\{\mathcal{R}_K\} \end{aligned}$$

and the observed information on termination is

$$\begin{aligned} E\{\mathcal{I}\} &= (Pr\{\mathcal{A}_1\} + Pr\{\mathcal{R}_1\})\,\mathcal{I}_1 + \ldots \\ &\quad + (Pr\{\mathcal{A}_K\} + Pr\{\mathcal{R}_K\})\,\mathcal{I}_K. \end{aligned}$$

Recursive formulae

Armitage, McPherson & Rowe (*JRSS, A*, 1969) present recursive formulae for densities of score statistics at interim analyses.

On the Z-statistic scale:

The density $f_1(z_1)$ of Z_1 is that of a $N(\theta\sqrt{\mathcal{I}_1},\, 1)$ variate.

The joint distribution of the Z_ks implies that

$$Z_2|Z_1 \;\sim\; N(\theta(\mathcal{I}_2-\mathcal{I}_1)/\sqrt{\mathcal{I}_2} + Z_1\sqrt{(\mathcal{I}_1/\mathcal{I}_2)},\, (\mathcal{I}_2-\mathcal{I}_1)/\mathcal{I}_2).$$

Denote this conditional density by $f_2(z_2|z_1)$.

Since analysis 2 is only reached if $a_1 < Z_1 < b_1$, the sub-density for Z_2 is

$$f_2(z_2) \;=\; \int_{a_1}^{b_1} f_1(z_1)\, f_2(z_2|z_1)\, dz_1.$$

Recursive formulae

In the general recursive step, the sub-density for Z_k at analysis k is

$$f_k(z_k) \;=\; \int_{a_{k-1}}^{b_{k-1}} f_{k-1}(z_{k-1})\, f_k(z_k|z_{k-1})\, dz_{k-1},$$

where $f_k(z_k|z_{k-1})$ is the density of the distribution

$$N(\theta(\mathcal{I}_k - \mathcal{I}_{k-1})/\sqrt{\mathcal{I}_k} + Z_{k-1}\sqrt{(\mathcal{I}_{k-1}/\mathcal{I}_k)}, (\mathcal{I}_k - \mathcal{I}_{k-1})/\mathcal{I}_k).$$

Numerical quadrature can be used to evaluate each of f_1, f_2, etc., in succession on a grid of points. Then, for example, one can compute

$$Pr\{a_1 < Z_1 < b_1,\, a_2 < Z_2 < b_2,\, Z_3 > b_3\} \;=$$

$$\int_{a_2}^{b_2} f_2(z_2)\, \Phi\left(\frac{\theta(\mathcal{I}_3 - \mathcal{I}_2) + z_2\sqrt{\mathcal{I}_2} - b_3\sqrt{\mathcal{I}_3}}{\sqrt{(\mathcal{I}_3 - \mathcal{I}_2)}}\right) dz_2.$$

Direct numerical integration

Alternatively, we can write probabilities as nested integrals, e.g.,

$$Pr\{a_1 < Z_1 < b_1,\, a_2 < Z_2 < b_2,\, Z_3 > b_3\} =$$

$$\int_{a_1}^{b_1} \int_{a_2}^{b_2} \int_{b_3}^{\infty} f_1(z_1)\; f_2(z_2|z_1)\; f_3(z_3|z_2)\; dz_3\; dz_2\; dz_1.$$

Applying numerical integration, we replace each integral by a sum of the form

$$\int_a^b f(z)\, dz \;=\; \sum_{i=1}^{n} w(i)\, f(z(i)),$$

where $z(1), \ldots, z(n)$ is a grid of points from a to b.

Direct numerical integration

Thus, we have

$$Pr\{a_1 < Z_1 < b_1,\, a_2 < Z_2 < b_2,\, Z_3 > b_3\} \approx$$

$$\sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \sum_{i_3=1}^{n_3} w_1(i_1)\, f_1(z_1(i_1))\; w_2(i_2) f_2(z_2(i_2)|z_1(i_1))$$

$$w_3(i_3)\; f_3(z_3(i_3)|z_2(i_2)).$$

Multiple integrations and summations will arise, e.g., for an outcome at analysis k,

$$\sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \ldots \sum_{i_k=1}^{n_k} w_1(i_1)\, f_1(z_1(i_1))\; w_2(i_2) f_2(z_2(i_2)|z_1(i_1))$$

$$\ldots\; w_k(i_k)\; f_k(z_k(i_k)|z_{k-1}(i_{k-1})).$$

Direct numerical integration

In the multiple summation

$$\sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \cdots \sum_{i_k=1}^{n_k} w_1(i_1)\, f_1(z_1(i_1))\; w_2(i_2) f_2(z_2(i_2)|z_1(i_1))$$

$$\cdots\; w_k(i_k)\; f_k(z_k(i_k)|z_{k-1}(i_{k-1})),$$

the structure of the k nested summations is such that the computation required is of the order of $k-1$ double summations.

Using Simpson's rule with 100 to 200 grid points per integral can give accuracy to 5 or 6 decimal places.

For details of efficient sets of grid points, see Ch. 19 of *Group Sequential Methods with Applications to Clinical Trials* by Jennison and Turnbull (2000).

5. Finding optimal group sequential tests

Recall we wish to find a group sequential test of H_0: $\theta \leq 0$ vs $\theta > 0$ with

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

$$Pr_{\theta=\delta}\{\text{Accept } H_0\} = \beta,$$

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

Minimum possible value of $\{E_0(\mathcal{I}) + E_\delta(\mathcal{I})\}/2.$

We deal with the constraints on error rates by introducing Lagrangian multipliers, creating the *unconstrained problem* of minimising

$$\{E_0(\mathcal{I}) + E_\delta(\mathcal{I})\}/2 \; + \; \lambda_1 \, Pr_{\theta=0}\{\text{Reject } H_0\} \; + \; \lambda_2 \, Pr_{\theta=\delta}\{\text{Accept } H_0\}.$$

We shall find a pair of multipliers (λ_1, λ_2) such that the solution has type I and II error rates α and β, then this design will solve the *constrained problem* too.

Bayesian interpretation of the Lagrangian approach

Suppose we put a prior distribution on θ with $Pr\{\theta = 0\} = Pr\{\theta = \delta\} = 0.5$ and specify costs of

1	per unit of information observed,
$2\,\lambda_1$	for rejecting H_0 when $\theta = 0$,
$2\,\lambda_2$	for accepting H_0 when $\theta = \delta$.

Then, the total Bayes risk is

$$\{E_0(\mathcal{I}) + E_\delta(\mathcal{I})\}/2 + \lambda_1\, Pr_{\theta=0}\{\text{Reject } H_0\} + \lambda_2\, Pr_{\theta=\delta}\{\text{Accept } H_0\},$$

just as in the Lagrangian problem.

The advantage of the Bayes interpretation is that it is easier to see how to solve the problem by techniques of “Dynamic Programming” or “Backwards Induction”.

Solution by Dynamic Programming

Denote the posterior distribution of θ given $Z_k = z_k$ at analysis k by

$$p^{(k)}(\theta|z_k), \quad \theta = 0,\, \delta.$$

At the final analysis, K

There is no further sampling cost, so compare decisions

$$\text{Reject } H_0\text{:} \qquad E(\text{Cost}) \;=\; 2\,\lambda_1\, p^{(K)}(0|z_K),$$

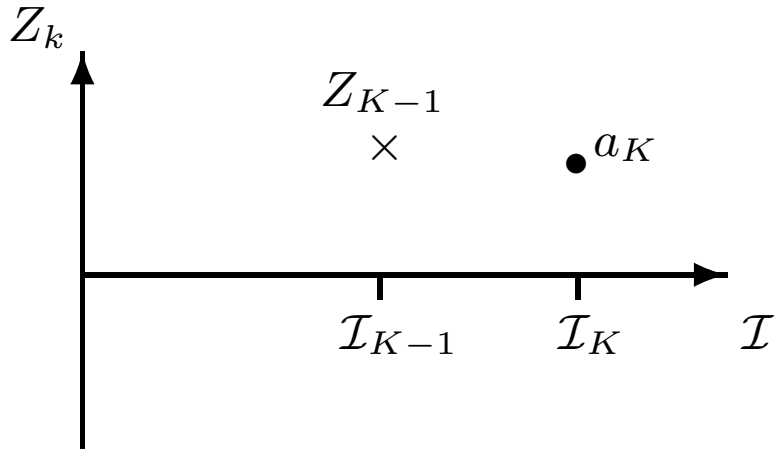
$$\text{Accept } H_0\text{:} \qquad E(\text{Cost}) \;=\; 2\,\lambda_2\, p^{(K)}(\delta|z_K).$$

The boundary point a_K is the value of z_K where these expected losses are equal.

The optimum decision rule is to reject H_0 for $Z_K > a_K$.

Dynamic Programming

At analysis $K-1$



If the trial stops at this analysis, there is no further cost of sampling and the expected additional cost is

Reject H_0: $\quad 2\,\lambda_1\,p^{(K-1)}(0|z_K)$,

Accept H_0: $\quad 2\,\lambda_2\,p^{(K-1)}(\delta|z_K)$.

At analysis $K-1$

If the trial continues to analysis K, the expected additional cost is

$$1 \times (\mathcal{I}_K - \mathcal{I}_{K-1})$$

$$+ 2\,\lambda_1\, p^{(K-1)}(0|z_{K-1})\, Pr_{\theta=0}\{Z_K > a_K | Z_{K-1} = z_{K-1}\}$$

$$+ 2\,\lambda_2\, p^{(K-1)}(\delta|z_{K-1})\, Pr_{\theta=\delta}\{Z_K < a_K | Z_{K-1} = z_{K-1}\}.$$

We can now define the optimal boundary points:

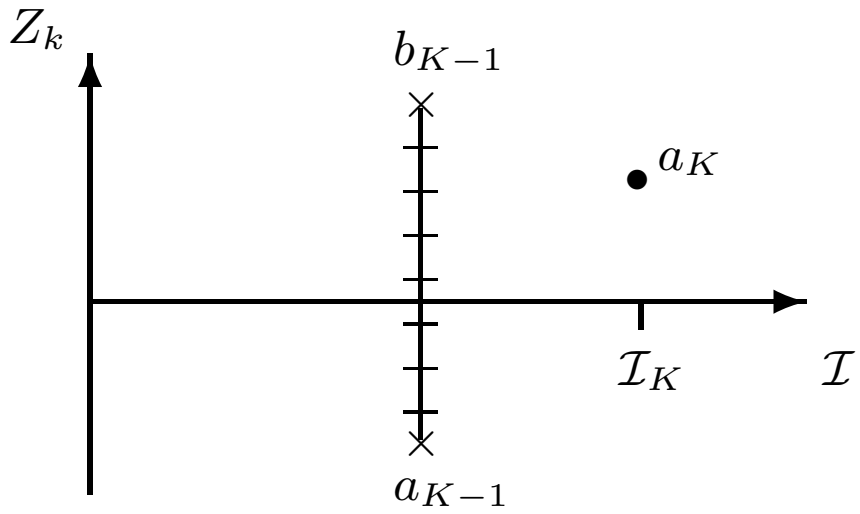
Set b_{K-1} to be the value of z_{K-1} where

$$E(\text{Cost of continuing}) \;=\; E(\text{Cost of stopping to reject } H_0),$$

Set a_{K-1} to be the value of z_{K-1} where

$$E(\text{Cost of continuing}) \;=\; E(\text{Cost of stopping to accept } H_0).$$

At analysis $K-1$



Before leaving analysis $K-1$, we set up a grid of points for use in numerical integration over the range a_{K-1} to b_{K-1}.

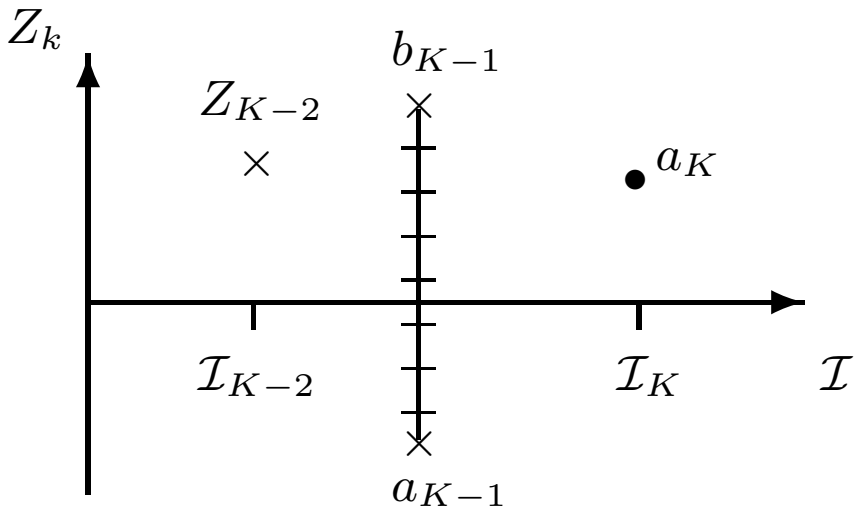
For each point, we sum over the posterior distribution of θ to calculate

$$\beta^{(K-1)}(z_{K-1}) \;=\; E(\text{Additional cost when continuing} \,|\, Z_{K-1} = z_{K-1}).$$

We are now ready to move back to analysis $K-2$.

Dynamic Programming

At analysis $K-2$



If the trial stops, the expected additional cost is

Reject H_0: $\quad 2\,\lambda_1\,p^{(K-2)}(0|z_K)$,

Accept H_0: $\quad 2\,\lambda_2\,p^{(K-2)}(\delta|z_K)$.

At analysis $K-2$

If the trial continues to analysis $K-1$, the expected additional cost is

$$1 \times (\mathcal{I}_{K-1} - \mathcal{I}_{K-2})$$

$$+ 2\,\lambda_1\, p^{(K-2)}(0|z_{K-2})\, Pr_{\theta=0}\{Z_{K-1} > b_{K-1}|Z_{K-2} = z_{K-2}\}$$

$$+ 2\,\lambda_2\, p^{(K-2)}(\delta|z_{K-2})\, Pr_{\theta=\delta}\{Z_{K-1} < a_{K-1}|Z_{K-2} = z_{K-2}\}$$

$$+ \int_{a_{K-1}}^{b_{K-1}} \{p^{(K-2)}(0|z_{K-2})\, f_0^{(K-1)}(z_{K-2}, z_{K-1}) +$$

$$p^{(K-2)}(\delta|z_{K-2}))\, f_\delta^{(K-1)}(z_{K-2}, z_{K-1})\}\, \beta^{(K-1)}(z_{K-1}) dz_{K-1},$$

where $f_\theta^{(K-1)}(z_{K-2}, z_{K-1})$ is the conditional density under θ of Z_{K-1} given $Z_{K-2} = z_{K-2}$.

At analysis $K-2$

Comparing costs for stopping and continuing at values of z_{K-2}, we can now define the optimal boundary points a_{K-2} and b_{K-2}.

We then set up a grid of points for use in numerical integration over the range a_{K-2} to b_{K-2}.

For each point, we calculate

$$\beta^{(K-2)}(z_{K-2}) \;=\; E(\text{Additional cost when continuing} \,|\, Z_{K-2} = z_{K-2}).$$

The process now moves back to analysis $K-3$, and so on all the way back to analysis 1.

Note: We have solved an "optimal stopping problem".

Solving the original problem

For any given $(\lambda_1,\ \lambda_2)$ we can find the Bayes optimal design and compute its type I and II error rates.

We now add another layer above this to search for a pair $(\lambda_1,\ \lambda_2)$ for which type I and type II error rates of the optimal design equal α and β respectively.

The resulting design will be the optimal group sequential test, with the specified frequentist error rates, for our original problem.

Notes

1. Since the output of the Dynamic Programming routine will be fed into another numerical algorithm, results should be of high accuracy. They should also possess the continuity properties, etc., that the higher level search algorithm expects to see.

2. The method provides an explicit demonstration that good frequentist procedures should be similar to Bayes procedures.

Properties of optimal designs

One-sided tests, $\alpha = 0.025$, $1 - \beta = 0.9$, 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 $\{E_0(\mathcal{I}) + E_\delta(\mathcal{I})\}/2$, as a percentage of $\mathcal{I}_{fix}$

	R					
K	1.01	1.05	1.1	1.2	1.3	*Minimum over R*
2	80.8	74.7	73.2	73.7	75.8	73.0 at R=1.13
3	76.2	69.3	66.6	65.1	65.2	65.0 at R=1.23
5	72.2	65.2	62.2	59.8	59.0	58.8 at R=1.38
10	69.2	62.2	59.0	56.3	55.1	54.2 at R=1.6
20	67.8	60.6	57.5	54.6	53.3	51.7 at R=1.8

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

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

Role of optimal designs

The methods we have described can be applied with a variety of optimality criteria.

We can minimise general criteria of the form $\sum_i w_i E_{\theta_i}(\mathcal{I})$.

Or, we can optimise

$$\int f(\theta)\, E_\theta(\mathcal{I})\, d\theta$$

for a normal density f.

As well as being available for direct use, optimal procedures serve as benchmarks for other methods which may have additional useful features.

They provide calibration for simple parametric boundaries or “error spending tests” which can handle uncertain information sequences.

6. Related problems

(i) Adaptive choice of group sizes in a group sequential test

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

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

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

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

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

Continue to choose $\mathcal{I}_3$ and observe S_3, etc, etc.

The whole procedure, sampling rule and stopping rule, should achieve desired *overall* type I error rate and power, and minimise a sample size criterion.

Examples of "Schmitz" designs

Consider designs which test H_0: $\theta = 0$ versus H_1: $\theta > 0$ with

Type I error rate $\alpha = 0.025$,

Power $1 - \beta = 0.9$ *at* $\theta = \delta$.

We wish to minimise

$$\int E_\theta(\mathcal{I})\, f(\theta)\, d\theta,$$

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

Maximum information $= 1.2 \times \mathcal{I}_{fix}$,

Maximum number of analyses $= K$.

Jennison & Turnbull (*Biometrika*, 2006) define and solve Bayes decision problems to find optimal Schmitz designs.

Examples of “Schmitz” designs

Optimal average $E(\mathcal{I})$ 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
5	59.2	60.5	60.9
10	55.9	57.2	57.5

The Schmitz procedures are complex and their efficiency gains are slight.

These results and the nature of the optimal adaptive designs shed light on recent proposals for sample size re-estimation in “adaptive” clinical trials.

Related problems

(ii) Testing for either superiority or non-inferiority

When there is already an accepted treatment for a condition, it is not appropriate to test a new treatment against placebo.

A trial using the standard treatment as an active control has two positive outcomes:

Showing the new treatment is *superior* to the current standard,

Showing the new treatment is *non-inferior* to the standard.

Investigators may start a trial intending to show superiority, then decide to adapt to a new goal of non-inferiority if results are not as good as expected.

Having two hypotheses is not an issue as the two tests are nested:

Superiority — Null hypothesis: $\theta \leq 0$,

Non-inferiority — Null hypothesis: $\theta \leq -d$.

Testing for superiority and non-inferiority

Differing sample size requirements

Wang, Hung, Tsong & Cui (*Statistics in Medicine*, 2001) note the non-inferiority margin d is often smaller than the effect size δ at which power for declaring superiority is specified.

Thus, a larger sample size is needed to test adequately for non-inferiority.

If early data indicate that the key issue is to test for non-inferiority, one may wish to increase the overall sample size.

Adaptive re-design

Wang et al. propose a group sequential test with group size determined by the power for superiority.

They then use the adaptive method of Cui, Hung & Wang (*Biometrics*, 1999) to increase group sizes if interest shifts to proving non-inferiority.

Testing for superiority and non-inferiority

A non-adaptive group sequential approach

One can embed testing for both superiority and non-inferiority in a group sequential design.

Early stopping may be appropriate:

to reject $H_{0,S}$: $\theta \leq 0$ (establishing superiority),

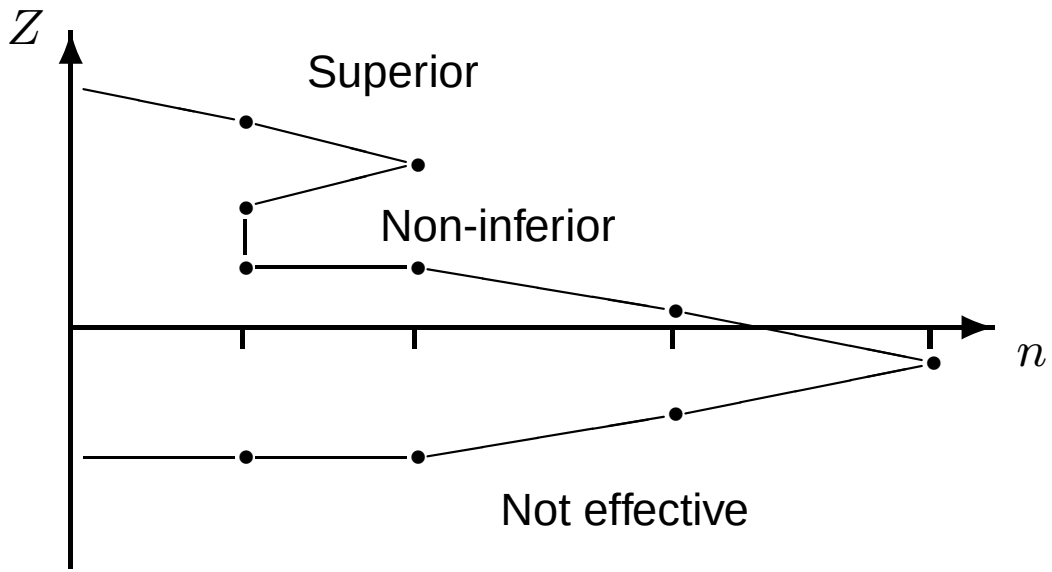
to accept $H_{0,NI}$: $\theta \leq -d$ (failing even to show non-inferiority),

to declare non-inferiority only.

If power for declaring superiority is set at a higher effect size, δ, than the margin of non-inferiority, d, the stopping rule for declaring superiority will be more aggressive.

Testing for superiority and non-inferiority

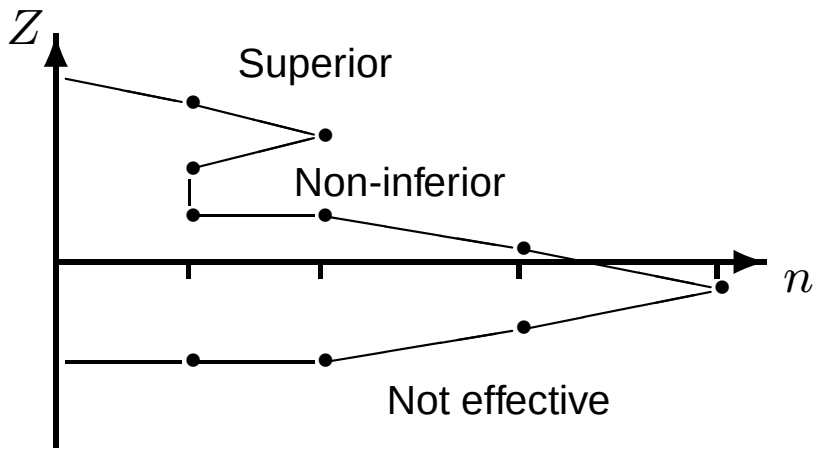
A group sequential design to test for either *Superiority* or *Non-inferiority* can have the general form:



In work with Fredrik Öhrn (to appear in *Statistics in Medicine*), we derived designs which minimise expected sample size while satisfying four error rate constraints.

Again, we define Bayes decision problems, solve these by Dynamic Programming, and search for costs such that the optimal procedure has specified error rates.

Testing for superiority and non-inferiority



The asymmetry of these designs is important when fixed sample sizes needed for superiority and non-inferiority goals are different.

Optimal designs can be used in their own right.

They also provide a benchmark against which to judge other proposals.

The design leads to larger sample sizes when the issue is to test between inferiority and non-inferiority: there is little further benefit in choosing group sizes adaptively.

Related problems

(iii) Group sequential tests for a delayed response

In many trials, response is measured after some period of time, e.g., change in a measurement from baseline to 4 months after treatment.

There can be further delays in validating and analysing responses.

Thus, after a group sequential test stops, one should expect additional data from "pipeline" subjects who have entered the study but not yet responded.

There has been some limited work on "adjusting" the final analysis of a trial to incorporate responses obtained after termination.

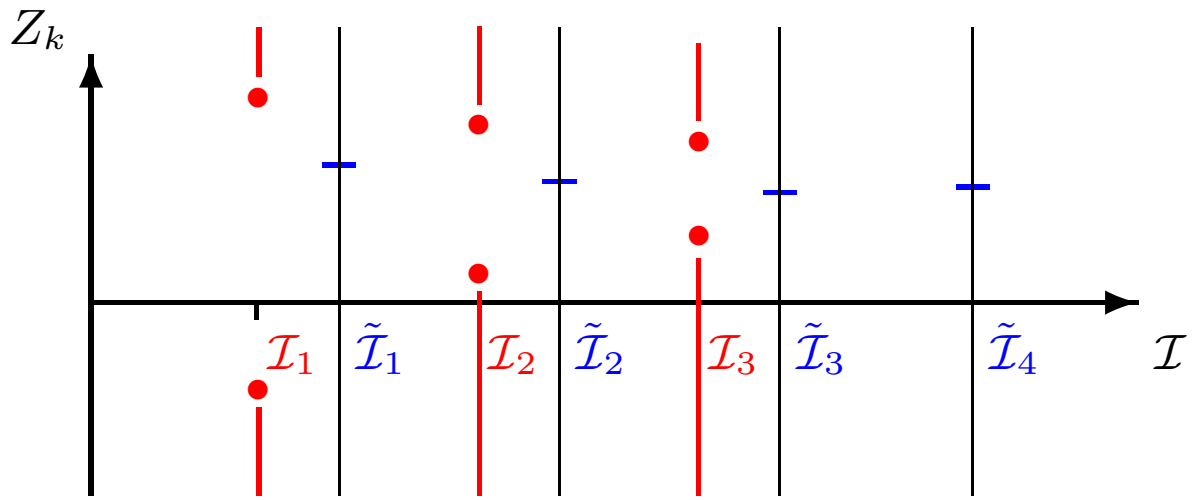
In recent work with Lisa Hampson, we have considered the derivation of group sequential designs which recognise there will be a delay in observing data.

Group sequential tests for a delayed response

A formal procedure terminates the trial in two stages:

1. Cease recruitment of new patients,

2. Wait for responses from all existing subjects, then make a final decision.



Exiting the boundary upwards or downwards at $\mathcal{I}_k$ indicates the likely decision.

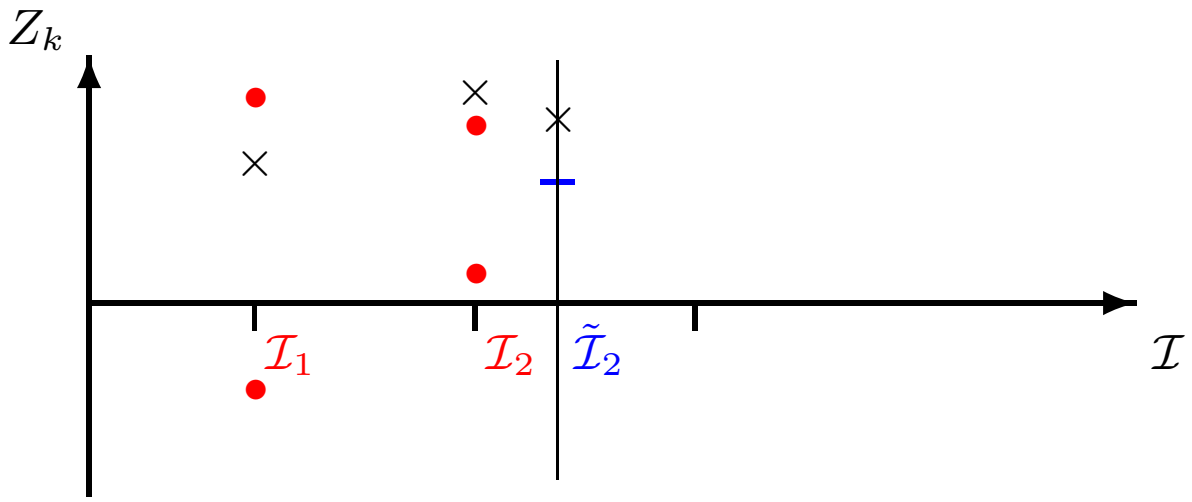
But, the full data at $\tilde{\mathcal{I}}_k$, including delayed responses, determines the final outcome.

Group sequential tests for a delayed response

For a particular sequence of observed responses, we apply boundary points at a sequence of information levels of the form

$$\mathcal{I}_1, \ldots, \mathcal{I}_k, \tilde{\mathcal{I}}_k.$$

Here, recruitment ceases at the kth analysis and the final decision is made with extra "pipeline" data bringing the information up to $\tilde{\mathcal{I}}_k$.



Group sequential tests for a delayed response

We have created Bayes problems and applied Dynamic Programming to find optimal delayed-response designs of this type for a variety of criteria.

Brief summary of results

Some of the benefits of group sequential tests in reducing expected sample size are lost when response is subject to delay.

The impact depends on the ratio r of the number of responses "in the pipeline" to the total fixed sample size.

Suppose a fixed sample test would need $N = 100$ observations.

A group sequential design with 5 analyses could have $E(N) \approx 60$.

With $r = 0.2$, the optimal design has $E(N) \approx 80$ — halving the benefits of sequential testing.

Using a second, rapidly observed endpoint

Suppose, however, that a second endpoint can be observed more rapidly and this has a high correlation with the primary endpoint.

We can use this information and fit a model for both endpoints which incorporates this correlation.

Then, the estimate of the primary endpoint gains in accuracy and information increases.

In the previous example, with a correlation between endpoints of 0.7, we find $E(N) \approx 70$ — so the benefits of group sequential testing are largely restored.

7. Conclusions

- The monitoring of clinical trials poses a range of problems of statistical inference and optimal design.
- A general distribution theory gives a basis for generic methodology.
- Using Dynamic Programming to solve specially constructed Bayes decision problems provides a route to deriving optimal designs.
- This methodology can be developed to solve a variety of additional problems of practical significance.