Group Sequential Methods

for Clinical Trials

Christopher Jennison,

Dept of Mathematical Sciences,

University of Bath, UK

http://www.bath.ac.uk/∼mascj

PSI, London,

22 October 2003

Plan of talk

1. Why sequential monitoring?
2. 1929, Dodge & Romig: 2-stage sampling
3. 1940s: methods for manufacturing
4. 1950s and 60s: methods for medical studies
5. 1970s: group sequential tests
6. Types of test, including equivalence, and types of stopping rule
7. Sequential theory, including survival data
8. A unified approach for group sequential design, monitoring and analysis
9. Nuisance parameters: updating a design
10. Survival data example
11. Error spending

1. Motivation of interim monitoring

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

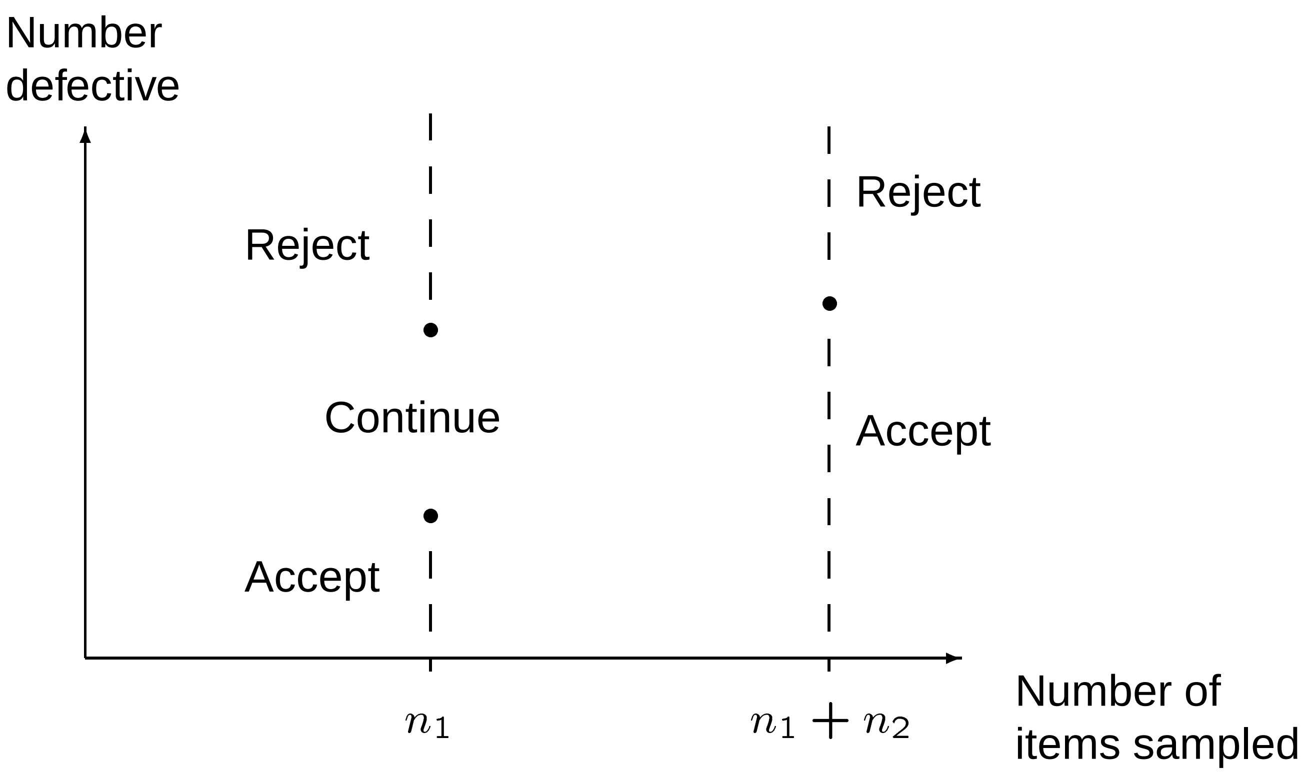
Number of tests	Error rate
1	0.05
2	0.08
3	0.11
5	0.14
10	0.19
20	0.25
100	0.37
∞	1.00

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

2. Acceptance sampling

Dodge & Romig (1929), *Bell Systems Technical Journal*.

Components are classified as effective or defective. A batch is only accepted if the proportion of defectives in a sample is sufficiently low.

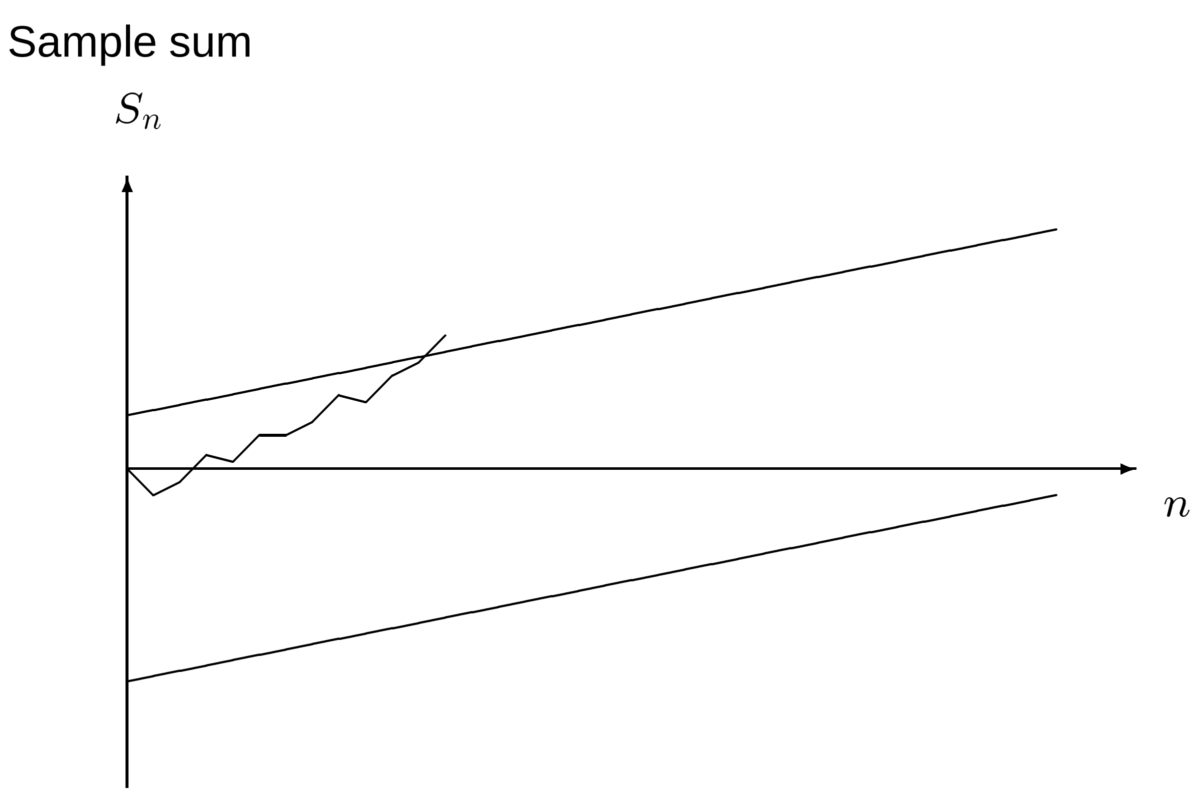


3. Manufacturing production

Barnard and Wald developed methods for industrial production and development.

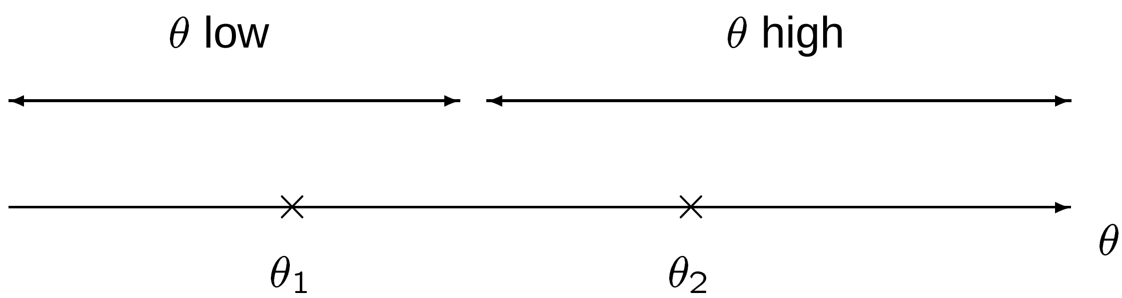
Wald (1947) published his Sequential Probability Ratio Test (SPRT) for testing between two simple hypotheses.

Stopping boundaries and continuation region



The SPRT

Ostensibly, the SPRT tests between H_1: $\theta = \theta_1$ and H_2: $\theta = \theta_2$.



In reality, it is usually used to choose between two sets of θ values.

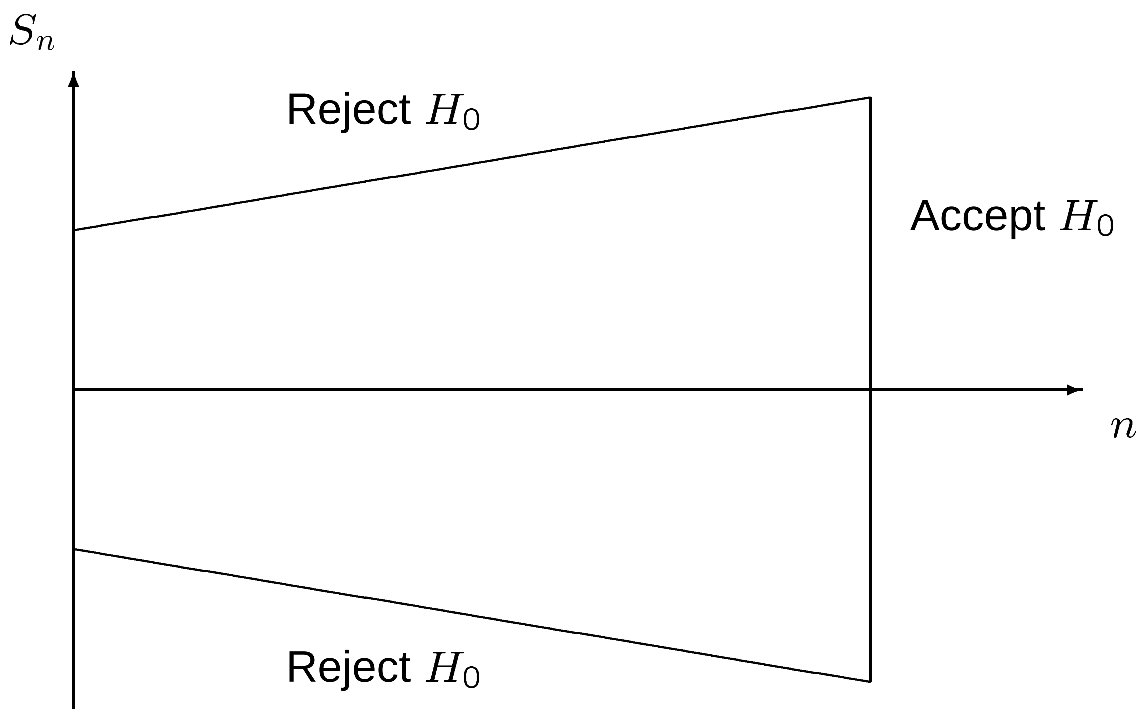
The SPRT has an "optimality" property if only θ_1 and θ_2 need be considered.

However, it assumes *continuous monitoring* of the data and has *no upper bound* on the possible sample size.

4. Sequential monitoring of clinical trials

In the 1950s, Armitage and Bross took sequential testing from industrial applications to comparative clinical trials. Their plans were fully sequential but with a bounded maximum sample size.

The "restricted" test, Armitage (1957),

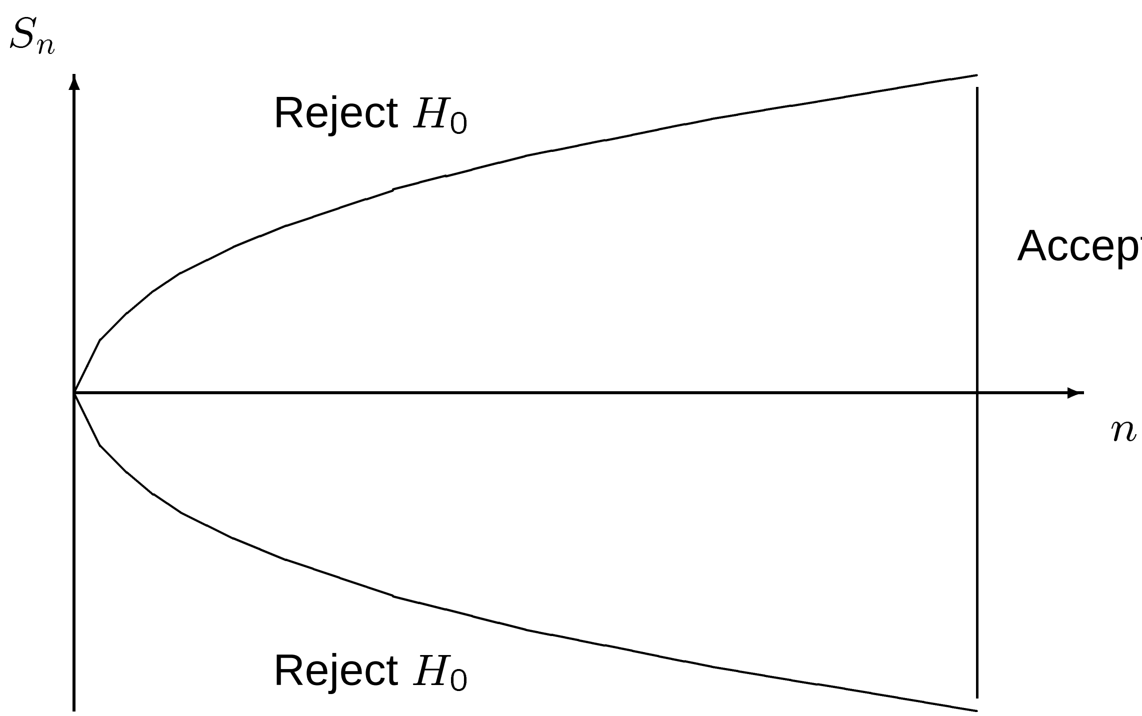


testing H_0: $\theta = 0$ against $\theta \neq 0$, where θ is the treatment difference.

Armitage's repeated significance test

Armitage, McPherson & Rowe (1969) applied a significance test of H_0: $\theta = 0$ after each new pair of observations.

Numerical calculations gave the "nominal" significance level α' to use in each of N repeated significance tests for an overall type I error probability α.



5. Group sequential tests

In practice, one can only analyse a clinical trial on a small number of occasions.

Shaw (1966): talked of a “block sequential” analysis.

Elfring & Schultz (1973): gave “group sequential” designs to compare two binary responses.

McPherson (1974): use of repeated significance tests at a small number of analyses.

Pocock (1977): provided clear guidelines for group sequential tests with given type I error and power.

O’Brien & Fleming (1979): an alternative to Pocock’s repeated significance tests.

Pocock's repeated significance test

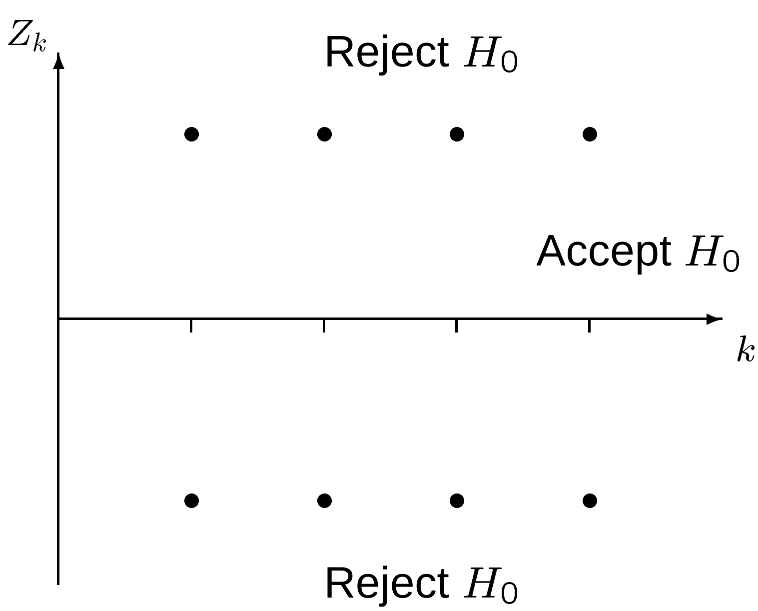
To test H_0: $\theta = 0$ against $\theta \neq 0$.

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.



6. 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:

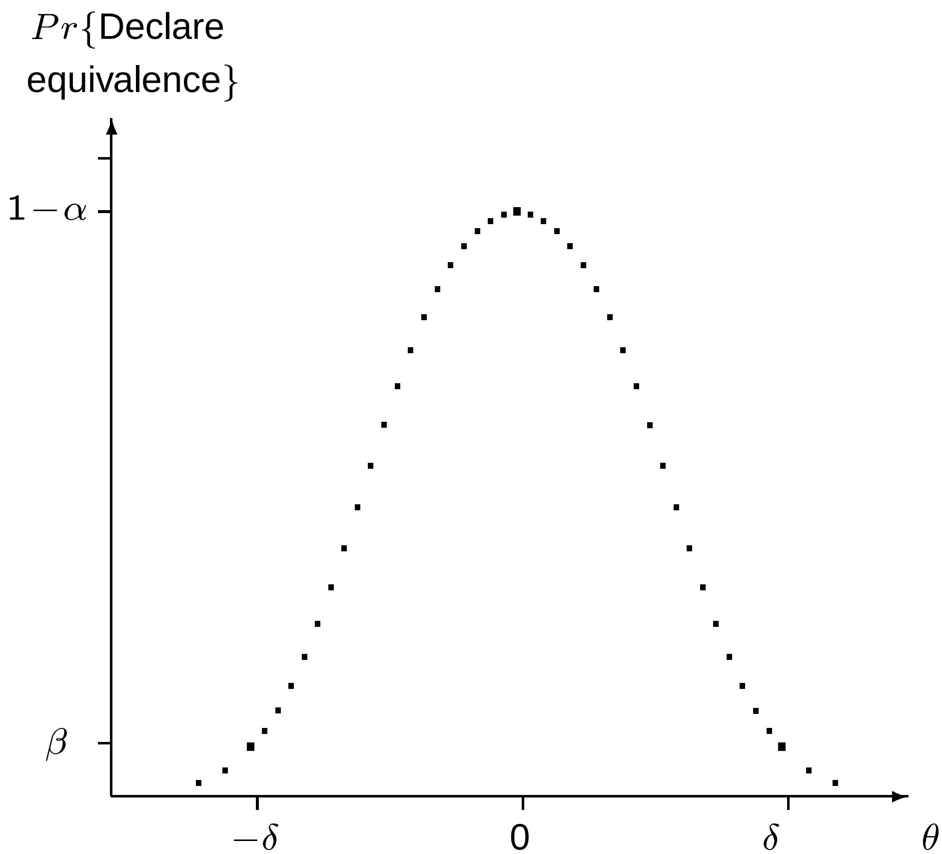
one-sided — to show treatment A is as good as treatment B, within a margin δ.

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

A two-sided equivalence test

Conduct a test of H_0: $\theta = 0$ *vs* $\theta \neq 0$ with type I error rate α and power $1 - \beta$ at $\theta = \pm\delta$.

Declare equivalence if H_0 is accepted.



Here, β represents the "consumer's risk."

In design and implementation, give priority to attaining power $1 - \beta$ at $\theta = \pm\delta$.

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, say, 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

If we are only interested in showing that a new treatment is superior to a control, we should test

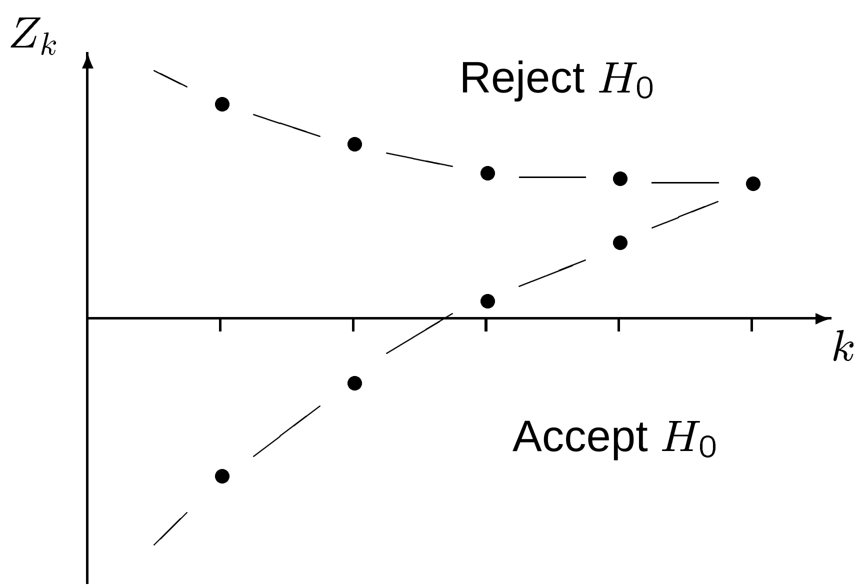
$$H_0\text{: } \theta \le 0 \quad \text{against} \quad \theta > 0,$$

requiring

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

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

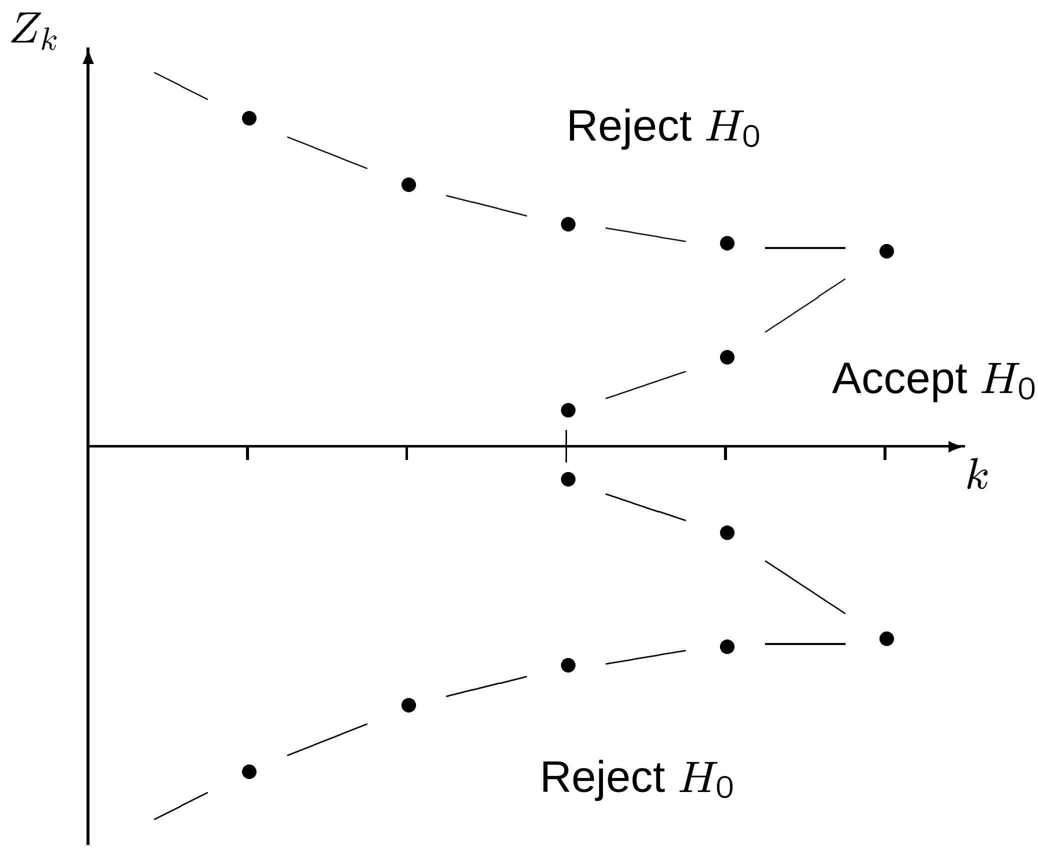
A typical boundary is:



E.g., Whitehead (1997), Pampallona & Tsiatis (1994).

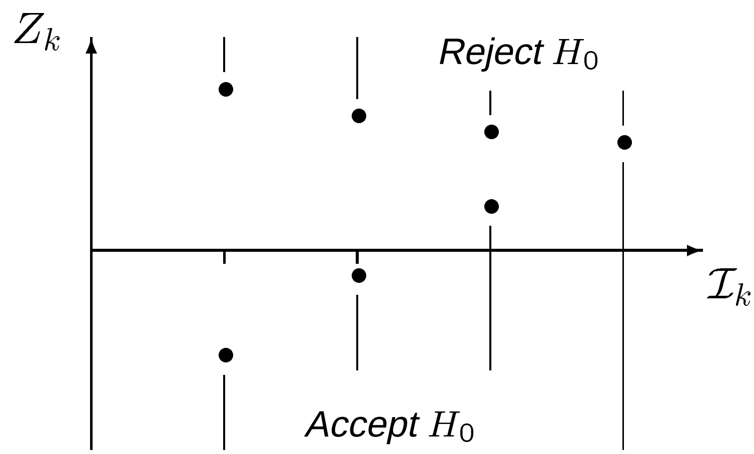
Two-sided tests with early stopping for H_0

Early stopping in favour of H_0 may be included in a two-sided test to "abandon a lost cause".

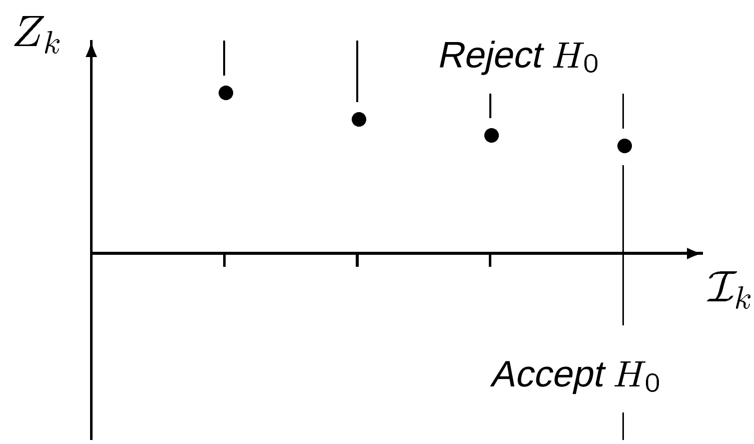


One-sided tests of H_0: $\theta = 0$ vs $\theta > 0$

Early stopping to reject H_0 or accept H_0

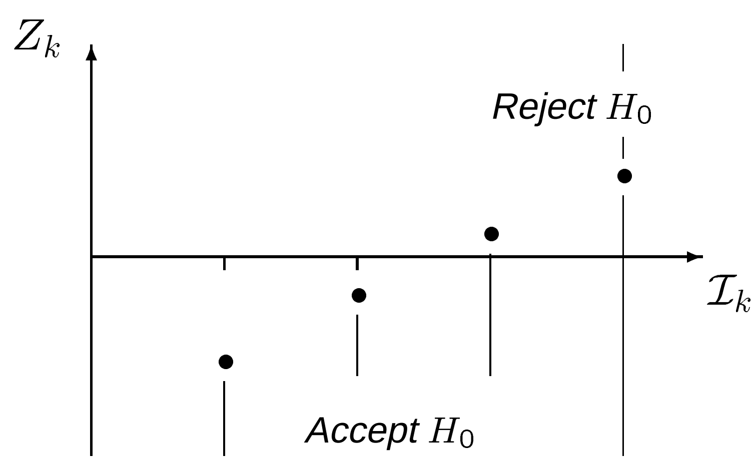


Early stopping only to reject H_0



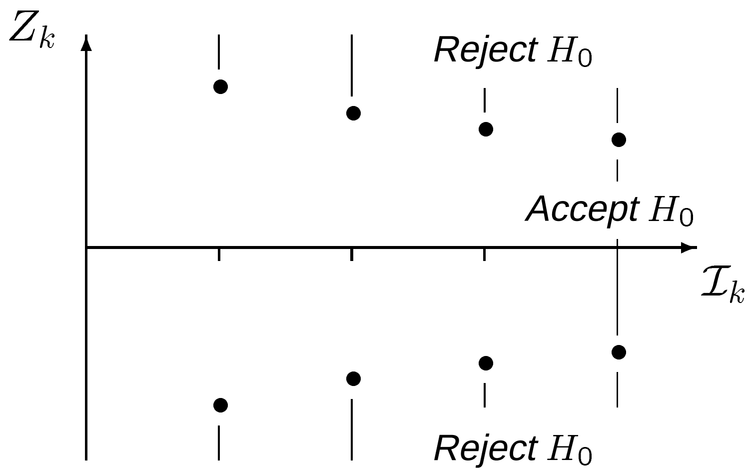
Abandoning a lost cause:

Early stopping only to accept H_0



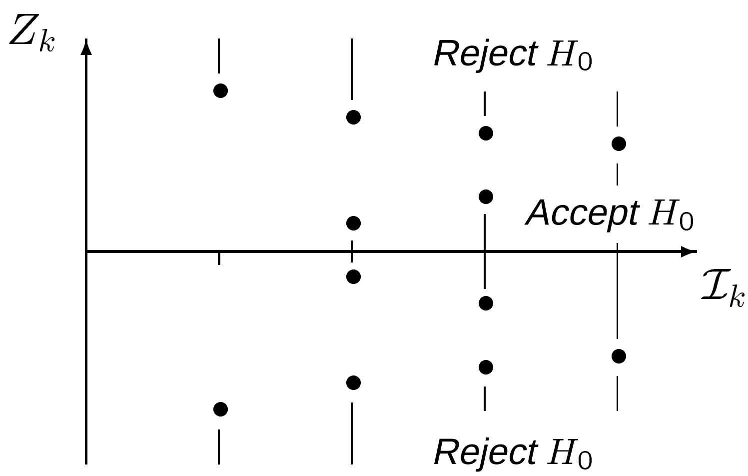
Two-sided tests of H_0: $\theta = 0$ vs $\theta \neq 0$

Early stopping to reject H_0



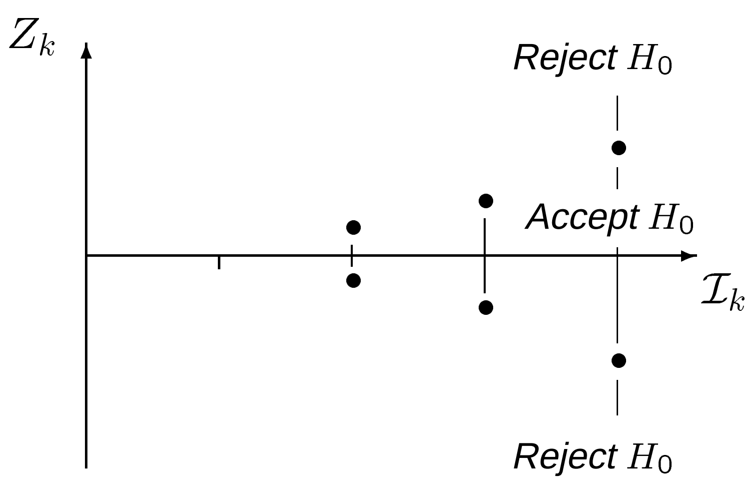
An inner wedge:

Early stopping to reject H_0 or accept H_0



Abandoning a lost cause:

Only an inner wedge



7. Joint distribution of parameter estimates

Reference: Jennison & Turnbull, Ch. 11

Suppose our main interest is in the parameter θ and let $\widehat{\theta}_k$ be the estimate of θ based on data 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 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

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

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.

Canonical joint distribution of z-statistics

In testing 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 $\mathcal{I}_1, \ldots, \mathcal{I}_K$.

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*.

8. Group sequential design, monitoring and analysis

To have the usual features of a fixed sample study,

- Randomisation, stratification, etc.,
- Adjustment for baseline covariates,
- Appropriate testing formulation,
- Inference on termination,

plus the opportunity for early stopping.

Response distributions:

- Normal, unknown variance
- Binomial
- Cox model or log-rank test for survival data
- Normal linear models
- Generalized linear models

General approach

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 in the case of an error spending test).

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

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*.

DESIGN — MONITORING — ANALYSIS

DESIGN

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, *Controlled Clinical Trials*, 1992)

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:

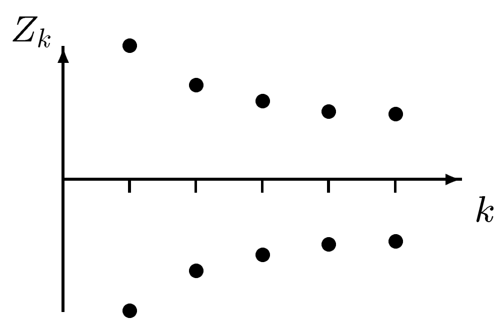
5 (fewer if we stop early)

Stopping boundary:

O'Brien & Fleming.

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

$$1.026 \times 65.67 = 68.$$

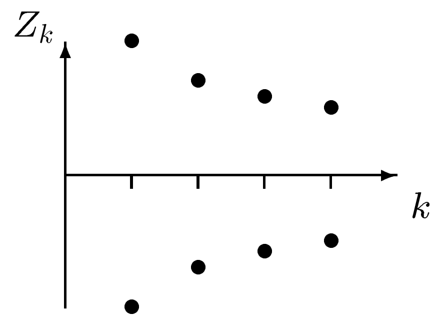
Divide this into 5 groups of 13 or 14 observations per treatment.

Some designs with K analyses

O'Brien & Fleming

Reject H_0 at analysis k if

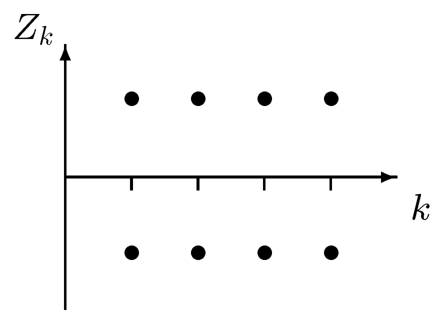
$$|Z_k| > c\sqrt{K/k}.$$



Pocock

Reject H_0 at analysis k if

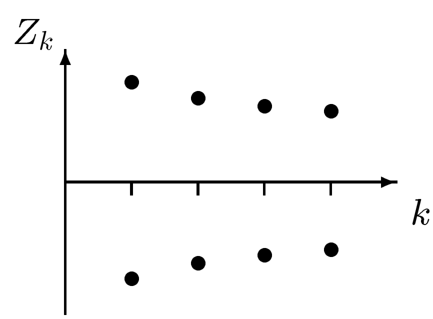
$$|Z_k| > c.$$



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 $|Z_5| < 2.040$.

The stopping rule gives

> type I error rate $\alpha = 0.050$ and
>
> power 0.902 at $\theta = 0.4$

if group sizes are equal to their design values.

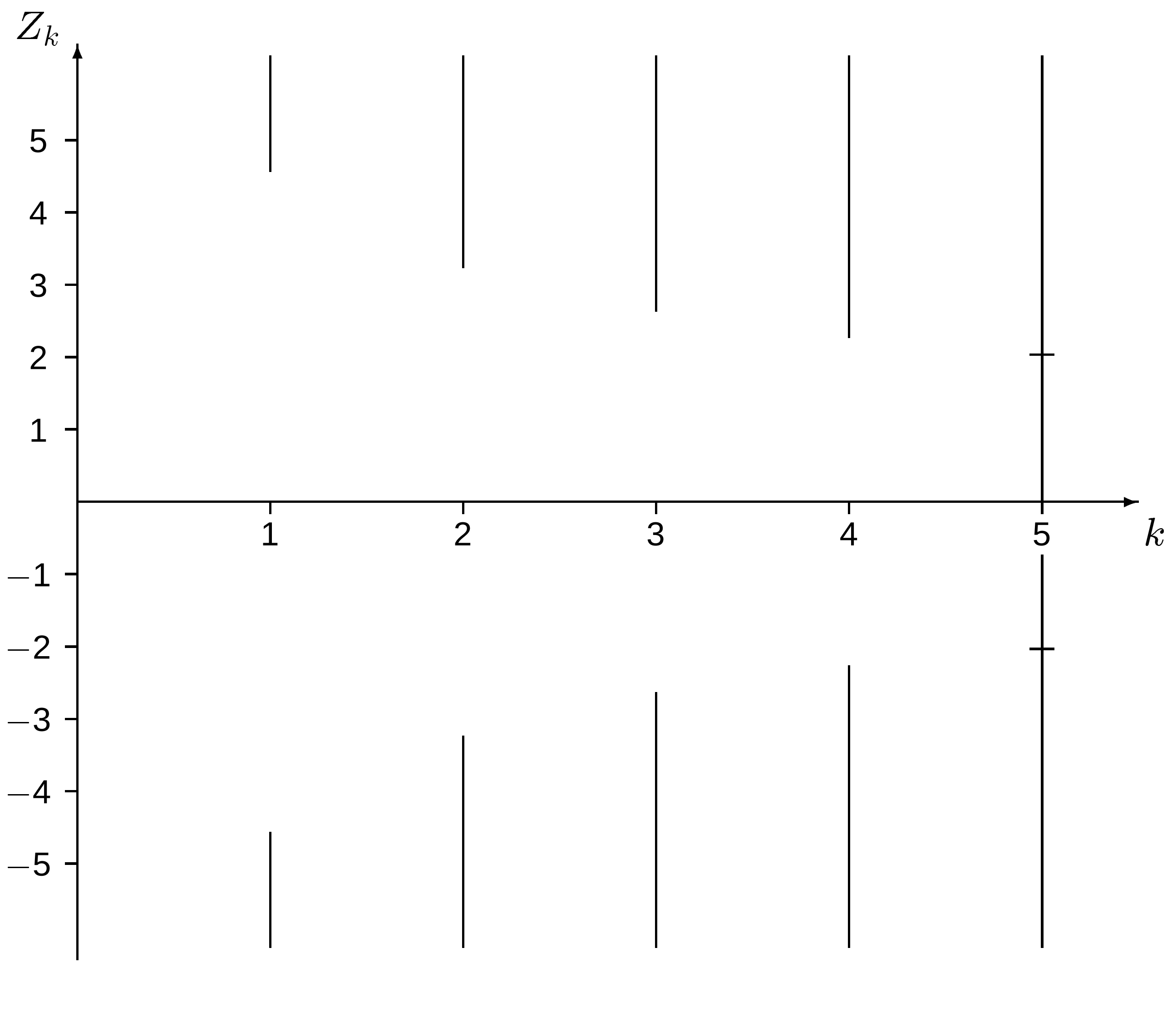
Note the minor effects of discrete group sizes.

Perturbations in error rates also arise from small variations in the *actual* group sizes.

For major departures from planned group sizes, we should really follow the “error spending” approach — see later.

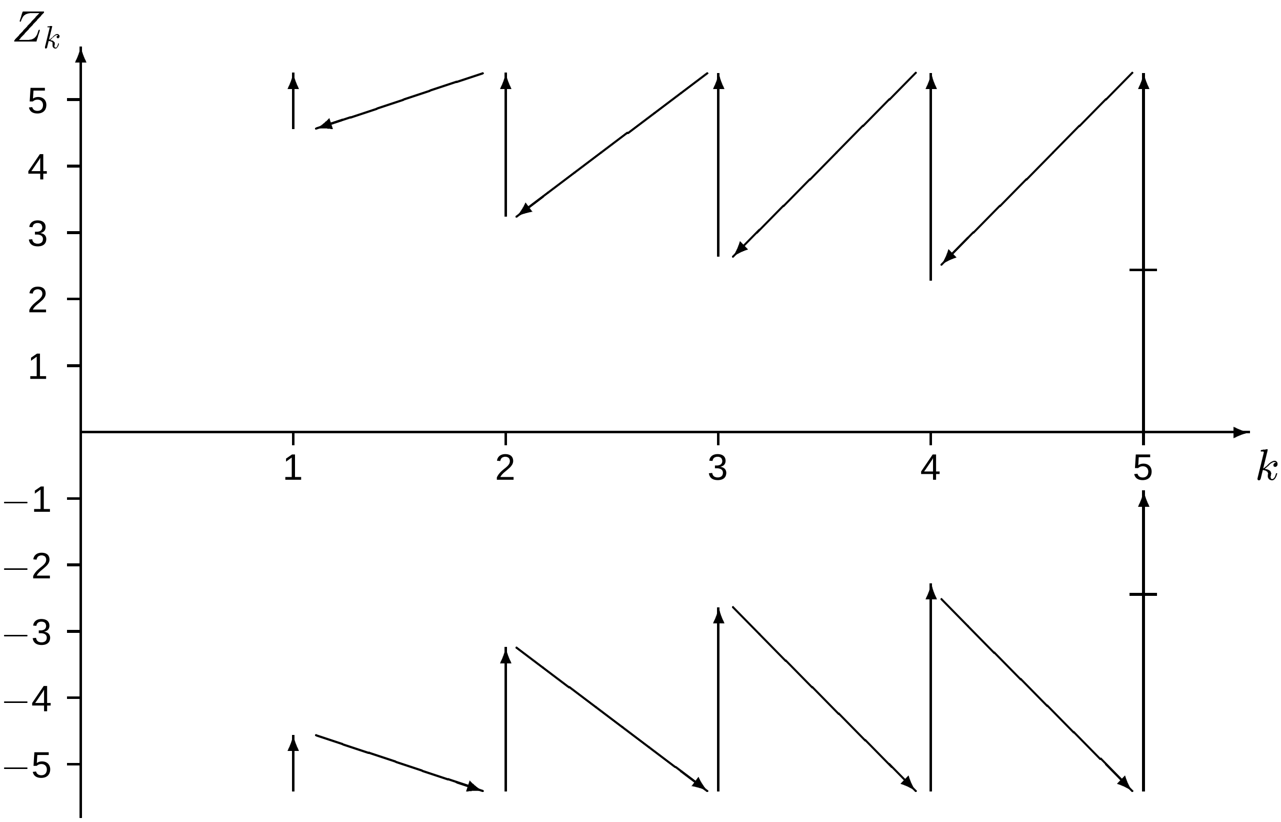
Analysis on termination

The sample space consists of all possible pairs (k, Z_k) on termination:



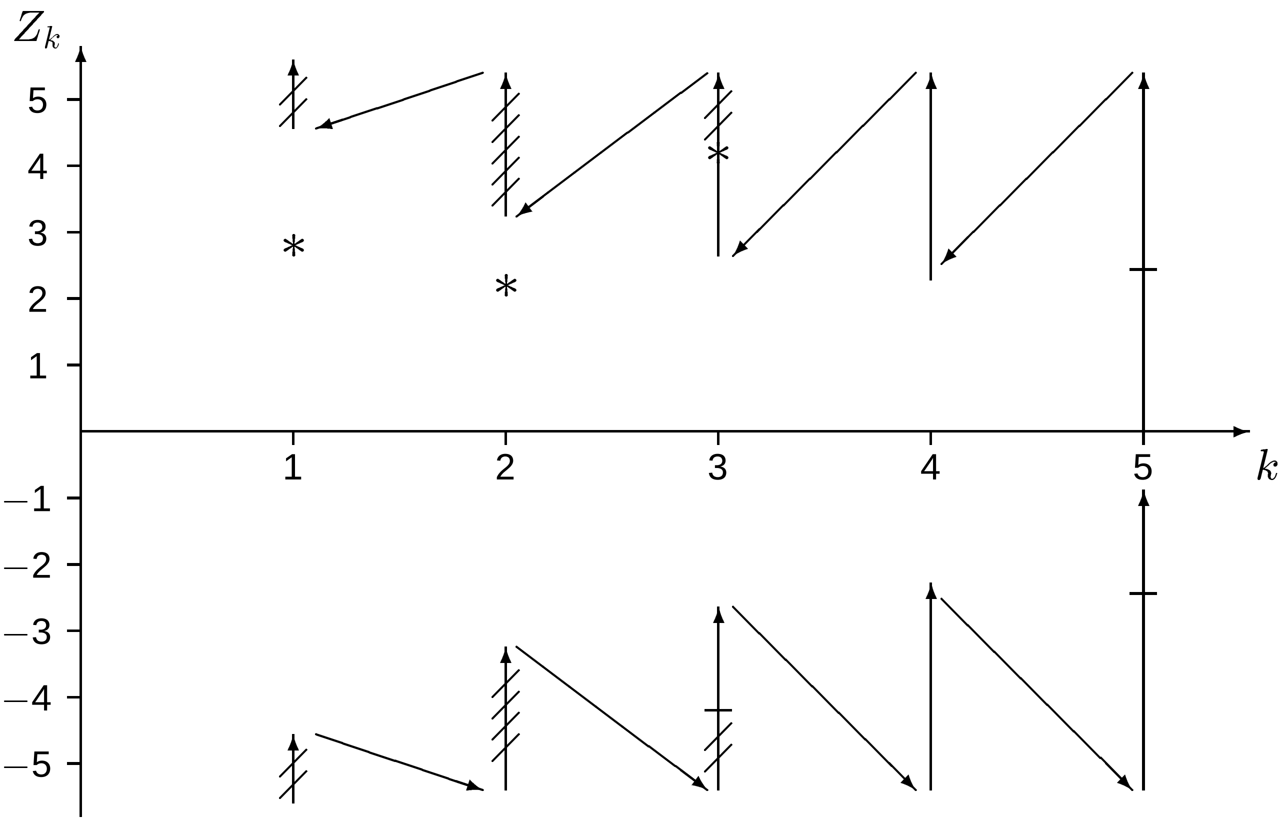
Analysis on termination

First, order the sample space.



We define P-values and confidence intervals with respect to this ordering.

The P-value for H_0: $\mu_A = \mu_B$ is the probability under H_0 of observing such an extreme outcome.



E.g., if the test stops at analysis 3 with $Z_3 = 4.2$, the two-sided P-value is

$$Pr_{\theta=0}\{|Z_1| \geq 4.56 \text{ or } |Z_2| \geq 3.23 \text{ or } |Z_3| \geq 4.2\}$$

$$= 0.0013.$$

A confidence interval on termination

Suppose the test terminates at analysis k^* with $Z_{k^*} = Z^*$.

A $100(1-\alpha)$% confidence interval for $\theta = \mu_A - \mu_B$ is the interval $(\theta_1,\, \theta_2)$ where

$$Pr_{\theta=\theta_1}\{\text{An outcome above } (k^*, Z^*)\} = \alpha/2$$

and

$$Pr_{\theta=\theta_2}\{\text{An outcome below } (k^*, Z^*)\} = \alpha/2.$$

E.g., if the test stops at analysis 3 with $Z_3 = 4.2$, the 95% confidence interval for θ is

$$(0.24,\, 0.91),$$

using our specified ordering.

Compare: fixed sample CI would be $(0.35,\, 0.95)$.

9. Updating a design as a nuisance parameter is estimated

The case of unknown variance

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 $\hat{\sigma}^2 = 0.5$,

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.

Remarks on “re-estimating” sample size

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

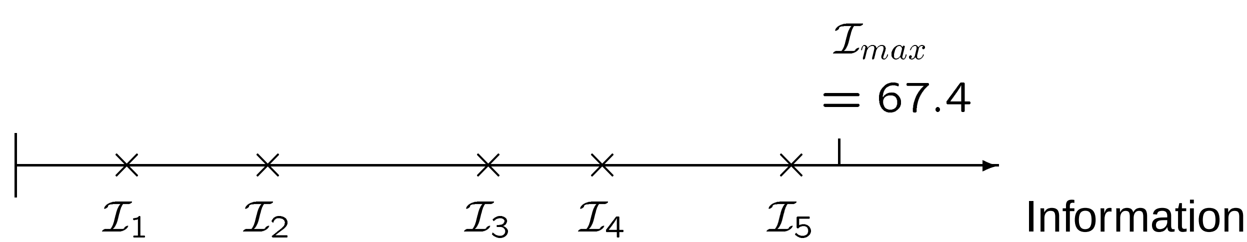
$$
\begin{aligned}
\mathcal{I}_{max} &= IF \times \{\Phi^{-1}(1-\alpha/2) + \Phi^{-1}(1-\beta)\}^2/\delta^2 \\
&= 1.026 \times (1.960 + 1.282)^2/0.4^2 = 67.4.
\end{aligned}
$$

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 $n = \ldots$

Recapitulation

Designing a group sequential test:

- Formulate the testing problem
- Create a fixed sample study design
- Choose number of analyses and boundary shape parameter
- Set maximum sample size equal to fixed sample size times the inflation factor

Monitoring:

- Find observed information at each analysis
- Compare z-statistics with critical values

Analysis:

- P-value and confidence interval on termination

This method can be applied to many response distributions and statistical models.

10. 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

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

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., $\lambda = 1.8$.

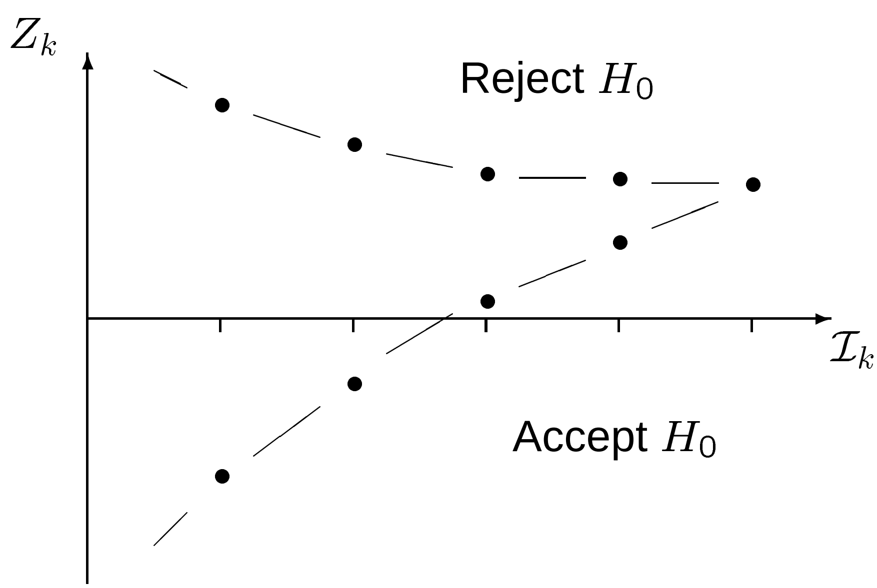
Information needed for a fixed sample study is

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

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.

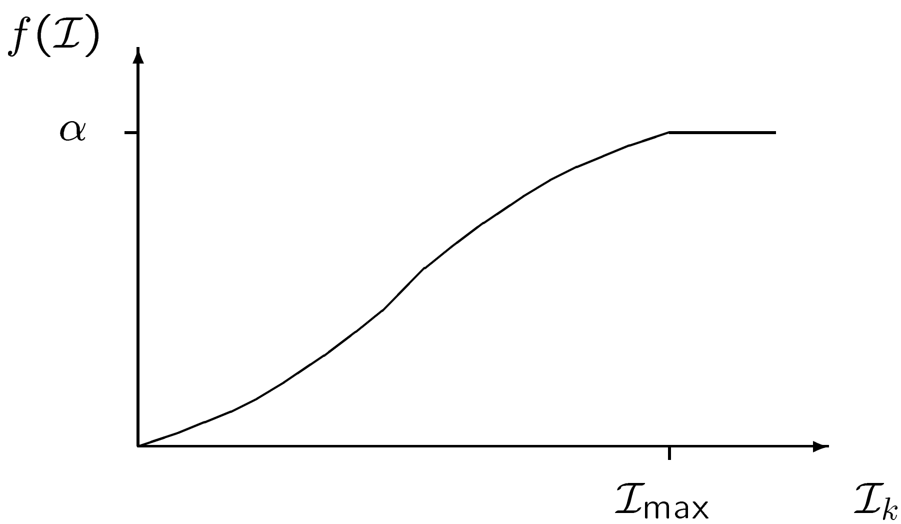
This leads to consideration of an “error spending” design.

11. Error spending tests

Lan & DeMets (1983) presented two-sided tests which "spend" type I error as a function of observed information.

Maximum information design:

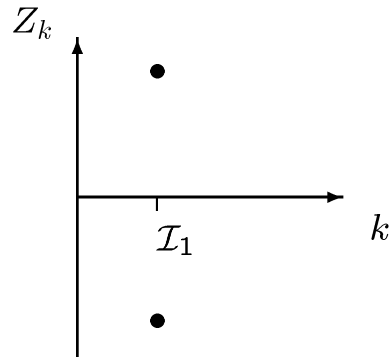
Error spending function $f(\mathcal{I})$



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.

Analysis 1:

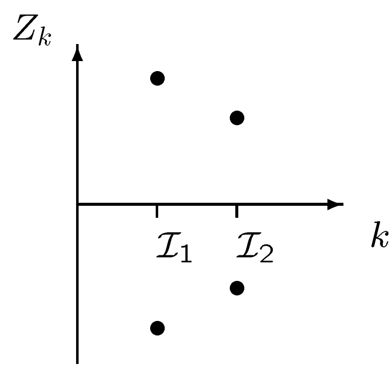


Observed information $\mathcal{I}_1$.

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

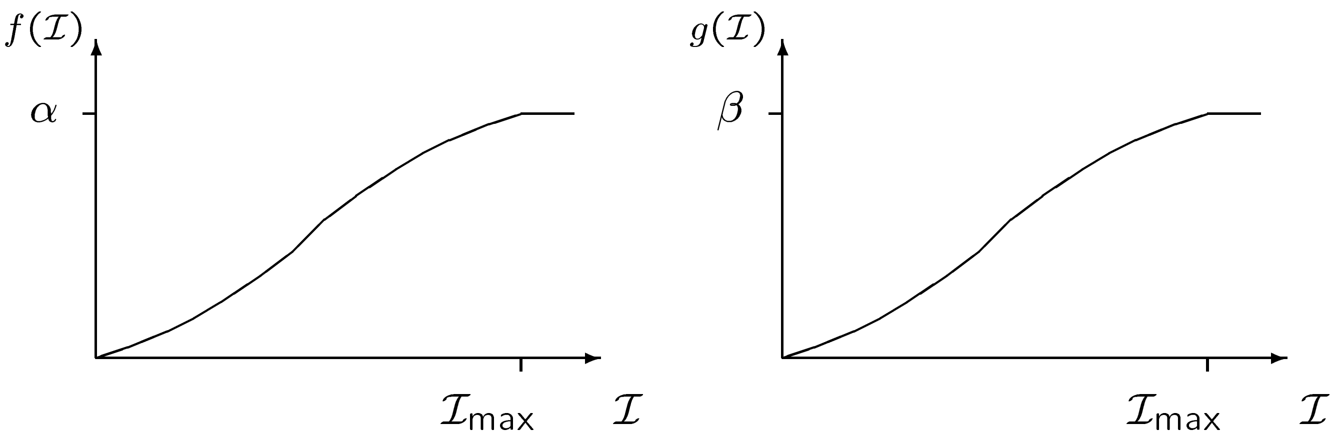
$$Pr_{\theta=0}\{|Z_1| < c_1,\, |Z_2| > c_2\}$$

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

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:

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

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

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

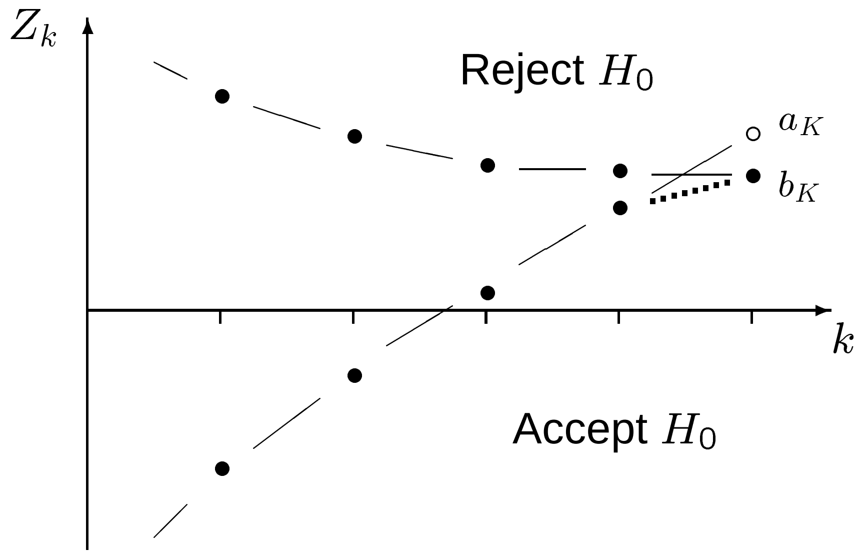
3. In a “maximum information design”, the study continues until the boundary is crossed or an analysis is reached with $\mathcal{I}_k \geq \mathcal{I}_{\mathrm{max}}$.

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

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

Over-running

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



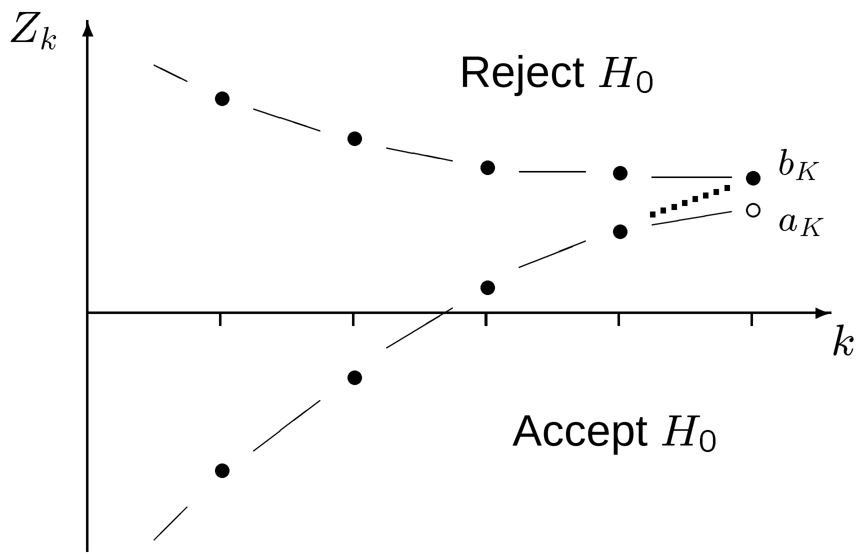
Keeping b_K as calculated guarantees type I error probability of exactly α.

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

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

Under-running

If a final information level $\mathcal{I}_K < \mathcal{I}_{\text{max}}$ 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 $1 - \beta$.

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., error spent is proportional to $(\mathcal{I}/\mathcal{I}_{\max})^2$.

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

Summary data and critical values for the Oropharynx trial

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

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

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

Covariate adjustment in the Oropharynx trial

Covariate information was recorded for subjects:

gender, initial condition, T-staging, N-staging.

These can be included in a proportional hazards regression model along with treatment effect β_1. The goal is then to test H_0: $\beta_1 = 0$ against the one-sided alternative $\beta_1 > 0$.

At stage k we have the estimate $\widehat{\beta}_1^{(k)}$,

$$v_k = \widehat{\text{Var}}(\widehat{\beta}_1^{(k)}),\ \mathcal{I}_k = v_k^{-1} \text{ and } Z_k = \widehat{\beta}_1^{(k)}/\sqrt{v_k}.$$

All these are available from standard Cox regression software.

The standardised statistics $Z_1, \ldots, Z_5$ have, approximately, the canonical joint distribution.

Covariate-adjusted group sequential analysis of the Oropharynx trial

Constructing the error spending test gives boundary values $(a_1, b_1), \ldots, (a_5, b_5)$ for $Z_1, \ldots, Z_5$.

k	$\mathcal{I}_k$	a_k	b_k	$\widehat{\beta}_1^{(k)}$	Z_k
1	4.11	-1.95	3.17	-0.79	-1.60
2	10.89	-0.61	2.59	-0.14	-0.45
3	19.23	0.43	2.20	-0.08	-0.33
4	28.10	1.28	1.90	0.04	0.20
5	30.96	1.86	1.86	0.01	0.04

Under this model and stopping rule, the study would have terminated at the 3rd analysis.

Further topics

Chapters of Jennison & Turnbull,

Group Sequential Methods with Applications to Clinical Trials:

- Ch 9. Repeated confidence intervals
- Ch 10. Stochastic curtailment
- Ch 12. Special methods for binary data
- Ch 15. Multiple endpoints
- Ch 16. Multi-armed trials
- Ch 17. Adaptive treatment assignment
- Ch 18. Bayesian approaches
- Ch 19. Numerical computations for group sequential tests