Group Sequential Designs

for Clinical Trials

Christopher Jennison

Department of Mathematical Sciences,

University of Bath, UK

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

Föreningen för Medicinsk Statistik

Göteberg,

September 24, 2007

1. Motivation of interim monitoring

There are reasons of

ethics

administration (accrual, compliance, . . .)

economics

to monitor accumulating data in a clinical trial.

Subjects should not be exposed to unsafe, ineffective or inferior treatments.

An effective new treatment should be made available as rapidly as possible.

National and international guidelines call for interim analyses to be performed — and reported.

It is standard for a Data and Safety Monitoring Board to oversee a study and consider possible 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. (*JRSS, A*, 1969), Table 2.

2. Pocock's repeated significance test

Pocock (*Biometrika*, 1977).

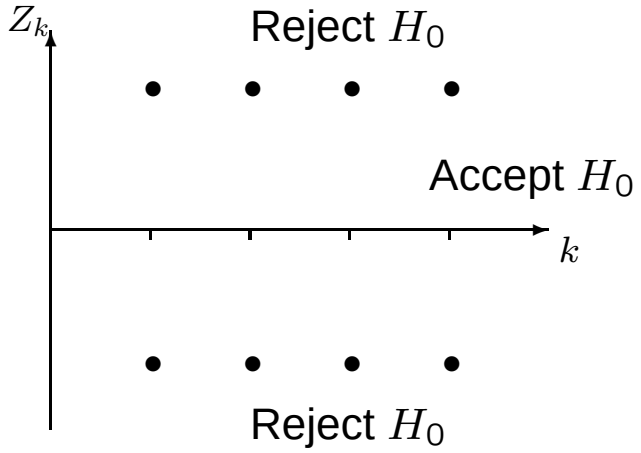
To test H_0: $\theta = 0$ vs $\theta \neq 0$ for treatment difference θ.

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 is not rejected by analysis K, stop and accept H_0.



Here, c is chosen to give the correct type I error rate.

Scope of Pocock's test

Pocock gave sample size formulae, and necessary constants, to meet type I error and power requirements.

He noted that good efficiency gains come from having just 2 or 3 groups of observations.

Pocock showed size and power are maintained if the same P-values and sample size formulae are used in tests for other response distributions:

- t-test,
- binary data,
- exponential data,
- Wilcoxon test.

He adapted the method, again via P-values, to deal with variable group sizes.

3. Regulatory requirements

The U. S. *Federal Register* (1985) published regulations for new drug applications including the requirement that the analysis of a Phase III trial

> "assess . . . the effects of any interim analyses"

This was elaborated in a Guideline (*FDA*, 1988):

> "The process of examining . . . data accumulating in a clinical trial . . . can introduce bias. Therefore all interim analyses, ***formal or informal***, by any study participant, sponsor staff member, or data monitoring group should be described in full even if treatment groups were not identified. The need for ***statistical adjustment*** because of such analyses should be addressed. . . . "

$\Rightarrow$ Need for a pre-specified stopping rule.

Adjustment, even for purely "administrative" analyses.

Updated guidelines

The FDA guidelines were updated in the *Federal Register* (1998) as

"E9 Statistical Principles for Clinical Trials".

This was prepared under the auspices of the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH).

The document lists recommendations for statistical principles and methodology applied to clinical trials in the pharmaceutical industry.

It advocates use of ***group sequential designs*** and gives detailed recommendations for trial conduct, including ***trial monitoring, interim analysis, early stopping, sample size adjustment*** and the role of an ***independent data and safety monitoring board*** (DSMB).

4. Types of hypothesis testing problems

Two-sided test:

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

One-sided test:

To show treatment A is superior to B,

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

Non-inferiority test:

To show treatment A is not clinically

inferior to treatment B,

testing H_0: $\theta \leq -\delta$ against $\theta > -\delta$.

Two-sided tests

If interested in a difference between treatments in either direction, we should test

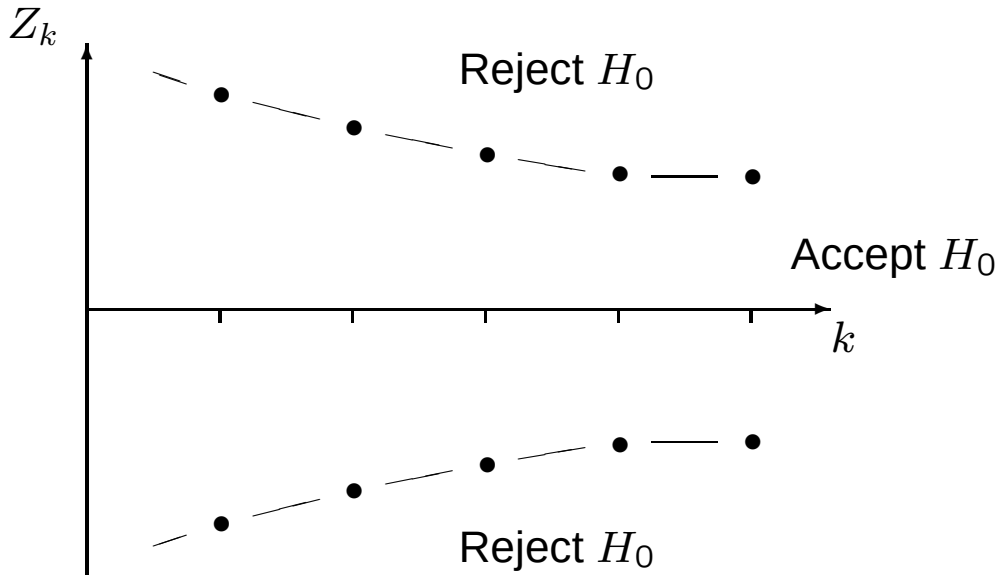
$$H_0\text{: } \theta = 0 \quad \text{against} \quad \theta \neq 0,$$

requiring

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

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

A typical boundary is:



E.g., Wang & Tsiatis (*Biometrics*, 1997).

One-sided tests

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

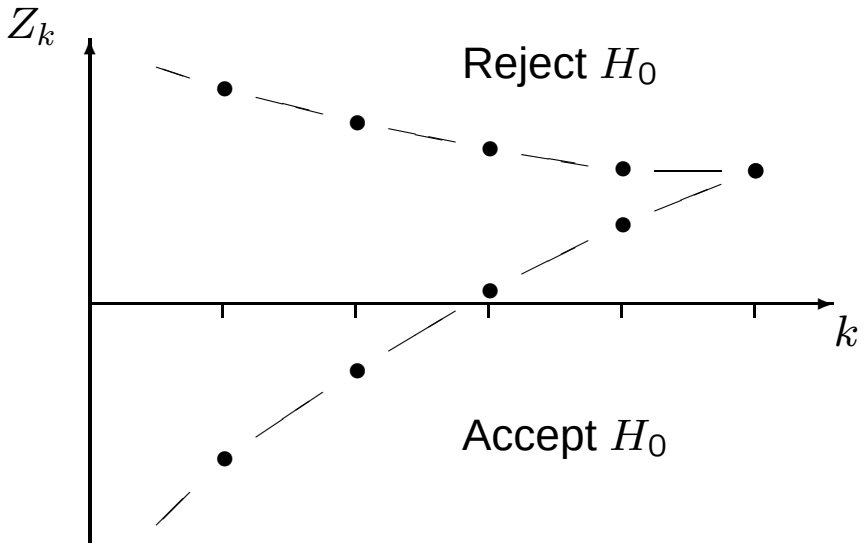
$$H_0\text{: } \theta \leq 0 \quad \text{against} \quad \theta > 0,$$

requiring

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

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

A typical boundary is:



E.g., DeMets & Ware (*Biometrika*, 1980),

Whitehead (1997) *The Design and Analysis of Clinical Trials.*

Demonstrating non-inferiority

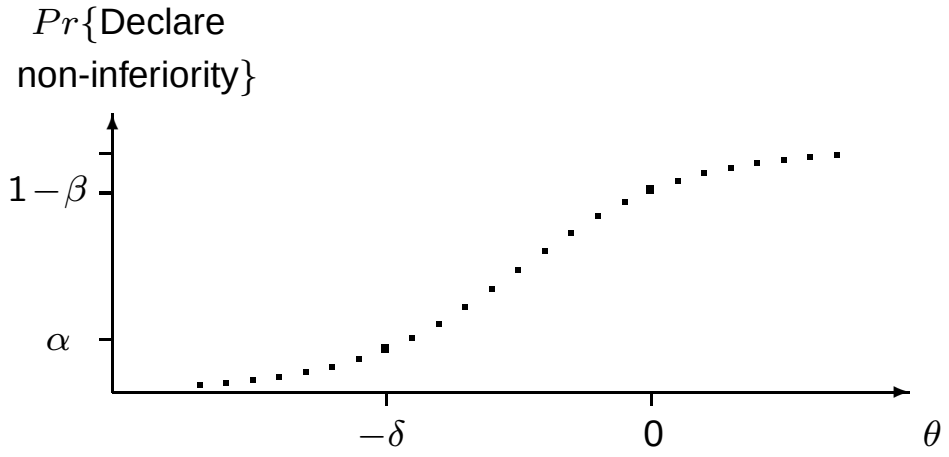
Suppose it is desired to show a new treatment is "no worse than" the current standard.

If the new treatment has mean μ_A and the standard μ_B, we wish to show $\mu_A \geq \mu_B - \delta$, where δ is a clinically acceptable fall in mean response.

Let $\theta = \mu_A - \mu_B$. We require

$$Pr\{\text{Declare non-inferiority} \,|\, \theta = -\delta\} \leq \alpha,$$

$$Pr\{\text{Declare non-inferiority} \,|\, \theta = 0\} \geq 1 - \beta.$$

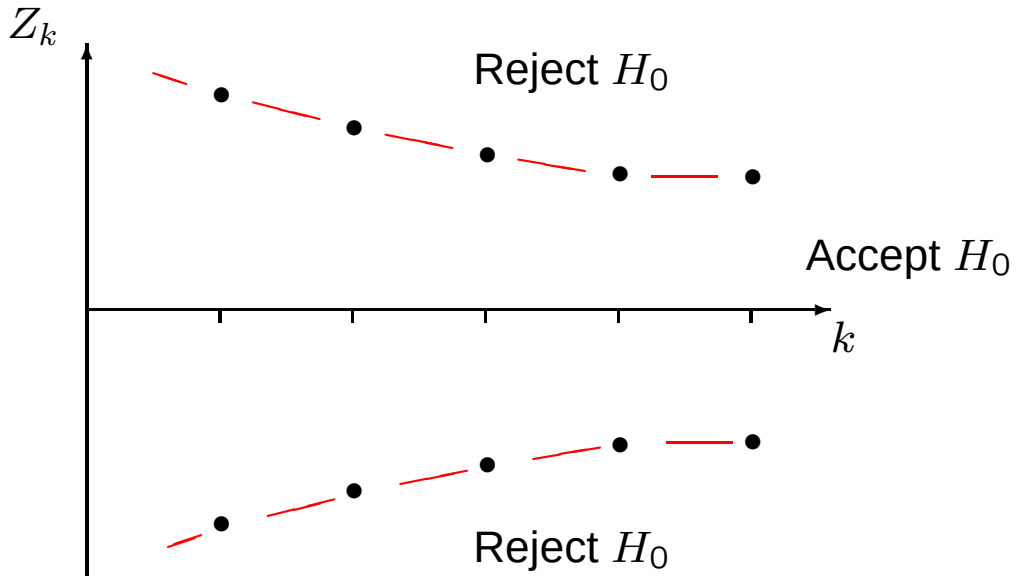


— the previous one-sided test with θ values shifted by $-\delta$.

5. Types of stopping rule

i) Stopping for a positive outcome

Two-sided test of H_0: $\theta = 0$ against $\theta \neq 0$.

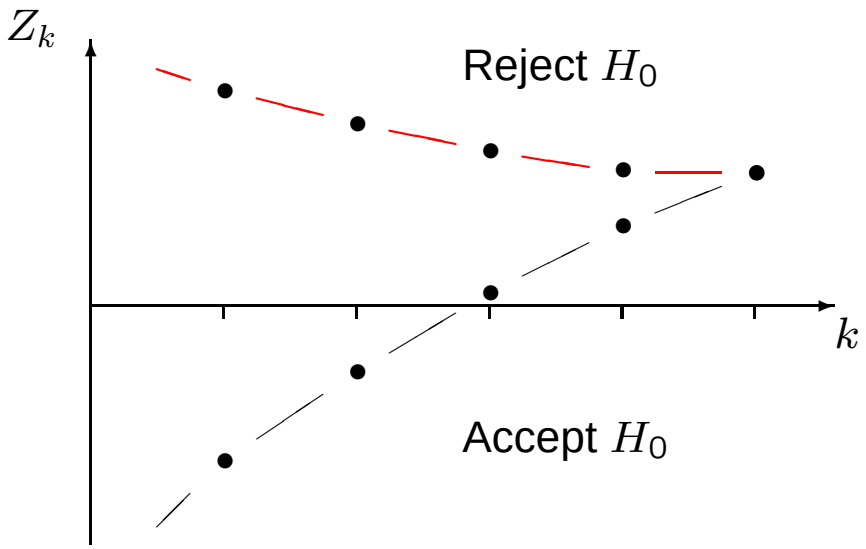


Crossing either the upper or lower boundary leads to rejection of H_0 and conclusion of a treatment difference.

Types of early stopping

i) Stopping for a positive outcome

One-sided test of H_0: $\theta \leq 0$ against $\theta > 0$.

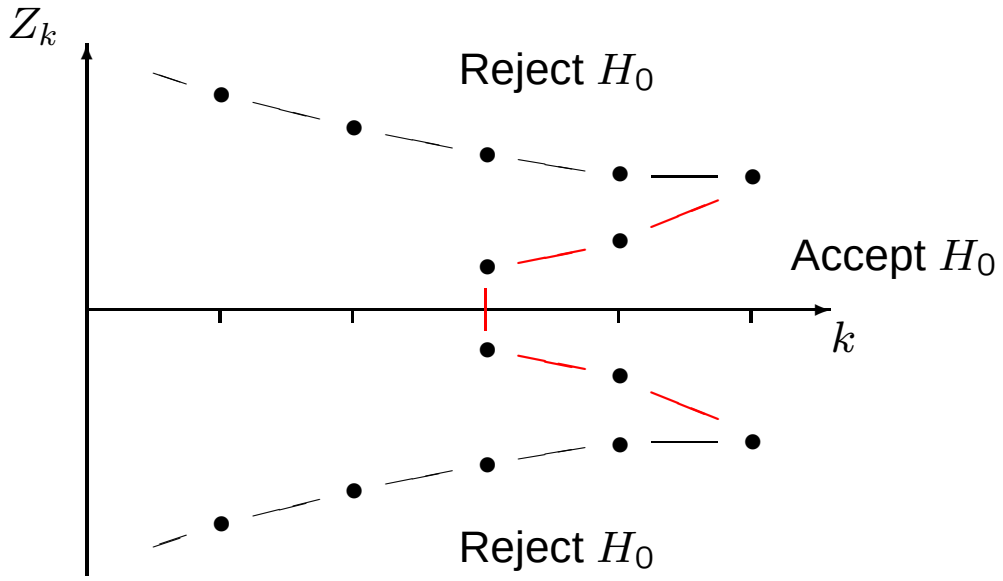


Crossing the upper boundary leads to rejection of H_0 in favour of $\theta > 0$ and conclusion that the new treatment is superior.

Types of early stopping

ii) Stopping for futility

Two-sided test of H_0: $\theta = 0$ against $\theta \neq 0$.



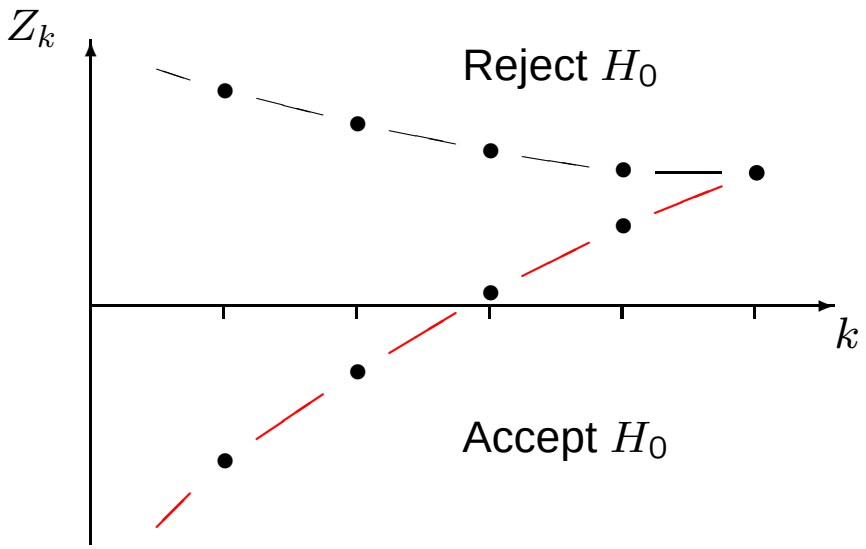
An inner boundary can be added for a two-sided test.

Crossing this inner boundary leads to early stopping with acceptance of H_0 in order to “abandon a lost cause”.

Types of early stopping

ii) Stopping for futility

One-sided test of H_0: $\theta \leq 0$ against $\theta > 0$.

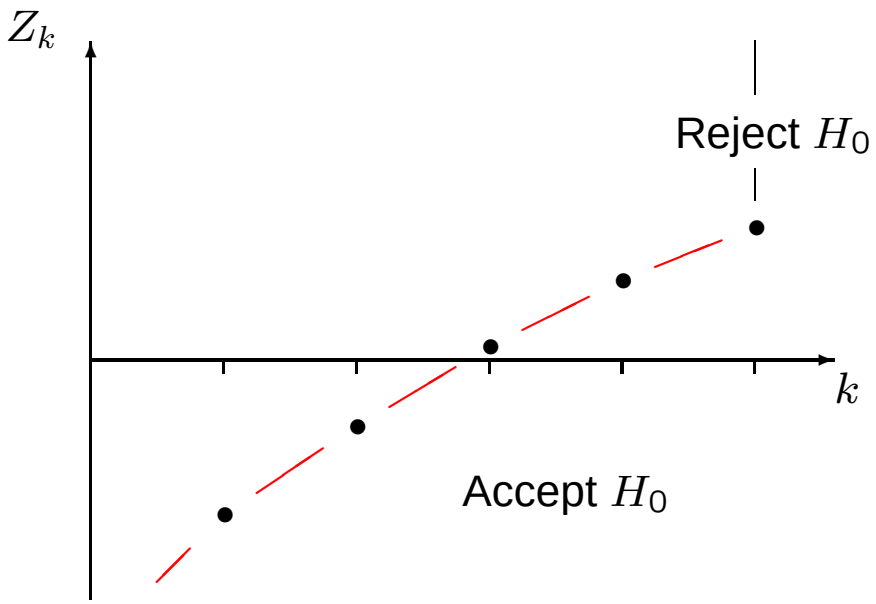


Crossing the lower boundary leads to an early decision to accept H_0.

Types of early stopping

ii) Stopping for futility

One-sided test of H_0: $\theta \leq 0$ against $\theta > 0$.



In some instances, it may be desirable to continue and check other aspects of the new treatment, even though results on the primary outcome are favourable.

The lower “futility” boundary remains for early termination when the study is unlikely to lead to a positive conclusion.

6. Underlying theory

Reference: JT, Ch. 11

Suppose our main interest is in the parameter θ and let $\widehat{\theta}_k$ be the estimate of θ from 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 $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$

Often, $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$ are approximately multivariate normal,

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

and

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

Joint distribution of z-statistics

Similar results apply to the sequence of *standardised statistics,* Z_k, for testing a null hypothesis H_0: $\theta = 0$.

At analysis k, we have

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

For the sequence of statistics Z_k,

$(Z_1, \ldots, Z_K)$ 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 $k_1 < k_2$.

Scope of this distribution theory

The preceding results for the joint distribution of $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$ apply when θ is a parameter in:

a general normal linear model,

a general model fitted by maximum likelihood.

This theory supports the analysis of *longitudinal data* and comparisons *adjusted for covariates*.

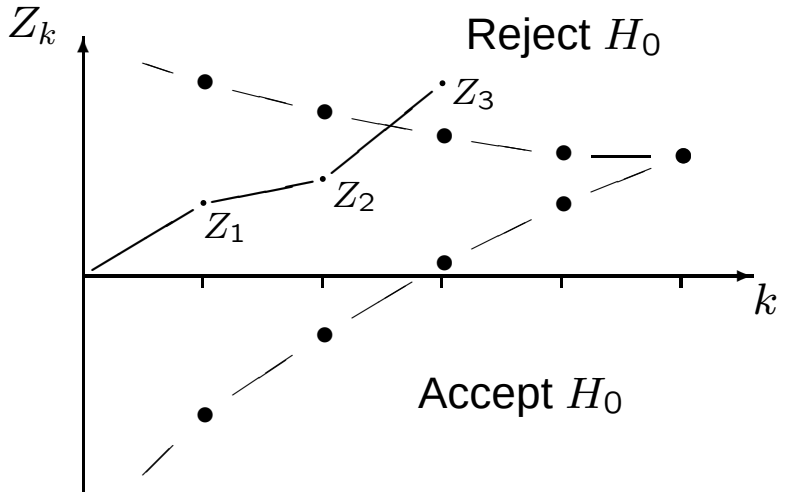
Results also apply to ***survival data,*** covering

estimates of a treatment effect parameter in Cox's proportional hazards regression model,

a sequence of log-rank statistics for comparing two survival curves.

7. Computations for group sequential tests

Reference: JT, Ch. 19



In order to find $Pr\{\text{Reject } H_0|\theta\}$, etc., we need to calculate the probabilities of basic events such as

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

Combining such probabilities gives properties of a given group sequential boundary.

Constants and group sizes can then be chosen to define a test with a specific type I error probability and power.

Numerical integration

We can write

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

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

Numerical integration replaces integrals by sums,

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

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

So, we have

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

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

$$f_2(z_2(i_2)|z_1(i_1))\; w_3(i_3)\; f_3(z_3(i_3)|z_2(i_2)).$$

Numerical integration

For an event at analysis k, we have a k-fold summation:

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

$$f_2(z_2(i_2)|z_1(i_1)) \ \ldots \ w_k(i_k)\ f_k(z_k(i_k)|z_{k-1}(i_{k-1})).$$

The structure of these 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 can give accuracy to 5 or 6 decimal places.

For further details, see JT, Ch. 19.

8. State of the art

Group sequential tests have been proposed for the testing problems and types of early stopping listed earlier.

Flexible “error spending” versions are able to handle unpredictable group sizes or information sequences.

For a thorough account, see Jennison and Turnbull’s

Group Sequential Methods with

Applications to Clinical Trials.

Software packages are available:

PEST, EAST, SEQSTAT, ADDPLAN, . . .

9. 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

First, 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:

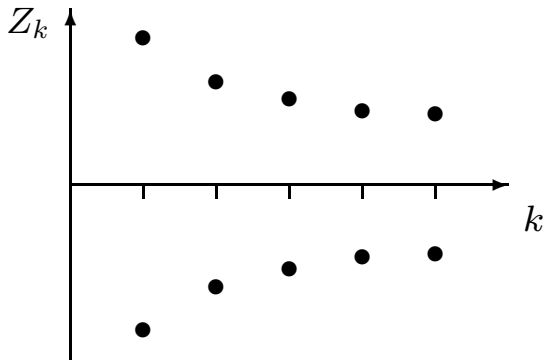
5 (fewer if we stop early)

Stopping boundary:

O'Brien & Fleming

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

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



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

$$1.026 \times 65.67 = 68.$$

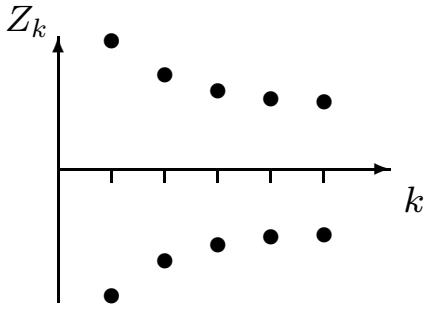
Divide this into 5 groups of 13 or 14 observations per treatment.

A choice of designs with K analyses

O'Brien & Fleming

Reject H_0 at analysis k if

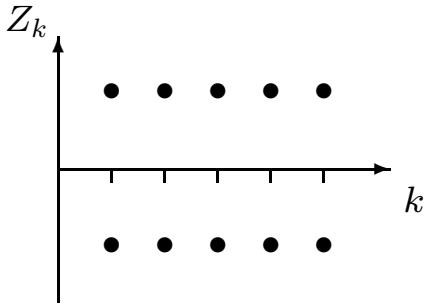
$$|Z_k| > c\sqrt{K/k}.$$



Pocock

Reject H_0 at analysis k if

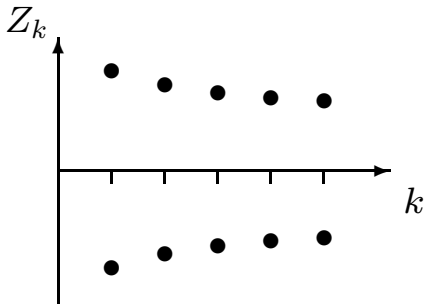
$$|Z_k| > c.$$



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*)

Properties of different designs

Sample sizes are per treatment.

Fixed sample size is 66.

K	Maximum sample size	Expected sample size $\theta = 0$	$\theta = \pm 0.2$	$\theta = \pm 0.4$
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