Statistical inference after group sequential and adaptive clinical trials

Christopher Jennison

Department of Mathematical Sciences,

University of Bath, UK

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

SMi, Adaptive Designs in Clinical Drug Development

London, February 2011

Motivating example

A new treatment is intended for the full patient population but a patient subgroup, defined by a biomarker, is thought to be particularly likely to respond.

An clinical trial with *enrichment* is proposed:

- Start by comparing the new treatment against control in the full population.
- Examine responses at an interim stage.
- If there is no evidence of treatment effect, stop for futility.
- If the new treatment appears effective in the full population, continue as before.
- If the new treatment appears to benefit just the subgroup, recruit only from the subgroup and increase the numbers in this subgroup.

In hypothesis testing, we allow for the multiplicity of hypotheses.

What about estimation, P-values and confidence intervals on termination?

Overview of this talk

I shall discuss:

Inference following a group sequential test.

Approximately unbiased point estimation,

P-values,

Confidence intervals on termination.

Inference following an adaptive group sequential test.

Inference following an adaptive trial design with multiple populations.

1. Inference following a group sequential test

Consider a two-treatment comparison with normally distributed responses on treatments A and B

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

The treatment effect is $\theta = \mu_A - \mu_B$.

It is desired to test H_0: $\theta = 0$ against $\theta \neq 0$ with two-sided type I error rate α.

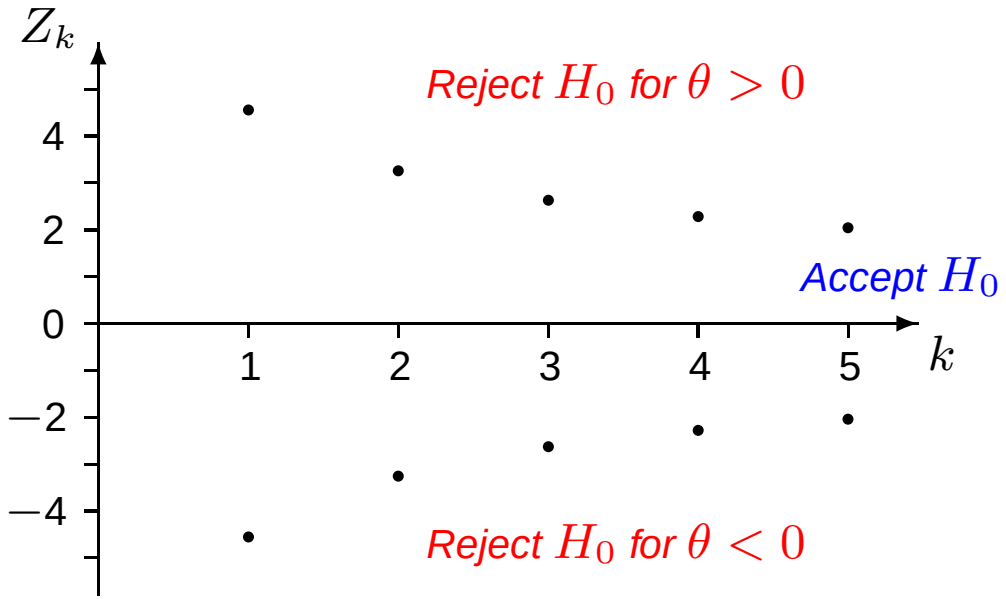
We shall use a group sequential test with $K = 5$ analyses.

At analysis k, suppose we have n_k observations on each treatment, then the standardised test statistic is

$$Z_k = \frac{\sum_{i=1}^{n_k} (X_i - Y_i)}{\sqrt{(2\, n_k\, \sigma^2)}}.$$

Inference following a group sequential test

We shall use an O'Brien & Fleming (*Biometrics*, 1979) boundary.

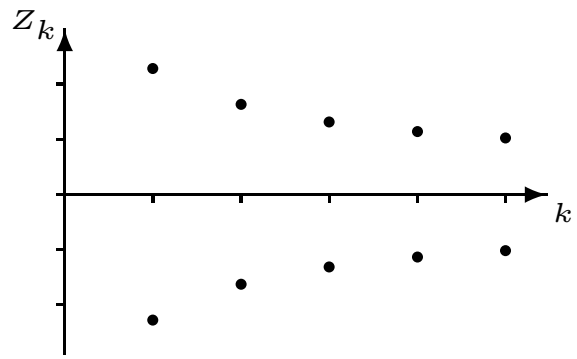


The trial stops to reject H_0 at analysis k if $Z_k \leq -b_k$ or $Z_k \geq b_k$.

If analysis 5 is reached and $-b_5 < Z_5 < b_5$, then H_0 is accepted.

Setting $b_k = \sqrt{5/k}\, 2.040$ gives two-sided type I error rate equal to 0.05.

(a) Estimating θ after a group sequential test



On termination of the test, the maximum likelihood estimate (MLE) of θ is

$$\hat{\theta}_M = \sum_{i=1}^{n_k} (X_i - Y_i)/n_k.$$

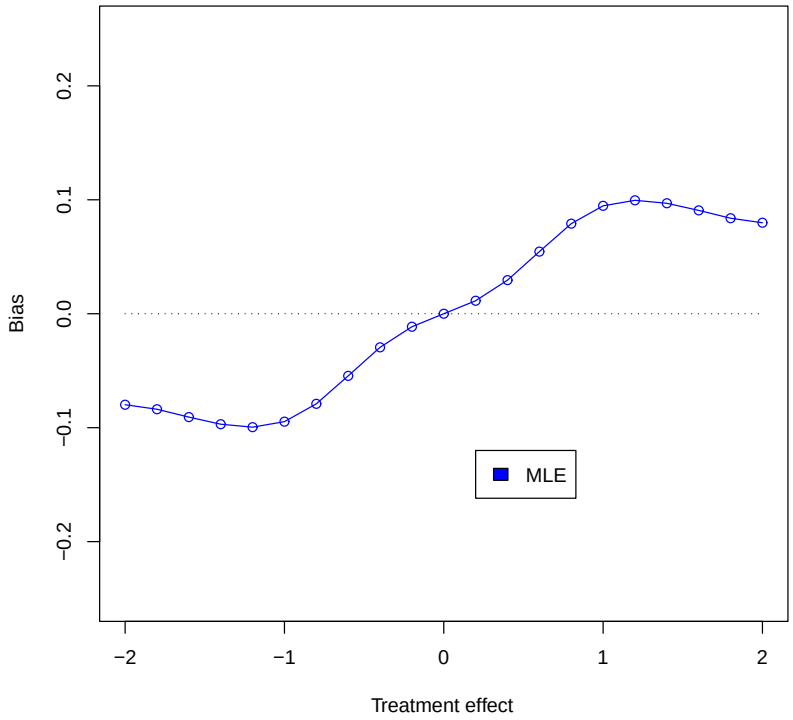
For positive values of θ, high values of $\hat{\theta}$ lead to early stopping, while lower values lead to collection of more data and the chance for $\hat{\theta}$ to increase.

This results in an upward bias of the MLE, so $E_\theta(\hat{\theta}_M) > \theta$ for $\theta > 0$.

Similarly, $E_\theta(\hat{\theta}_M) < \theta$ for negative values of θ.

Bias of the MLE of θ after a 5 group O'Brien & Fleming test

The bias of the MLE can be calculated as a function of the true effect size, θ.



With sample size chosen to give power 0.9 for detecting a treatment effect of ± 1, bias of the MLE is around 0.1 at $\theta = 1$ and -0.1 at $\theta = -1$.

Correcting the bias of the MLE

Denote the bias function of the MLE by

$$b(\theta) = E_\theta(\hat{\theta}_M) - \theta.$$

Whitehead (*Biometrika* 1986) suggested correcting the MLE by subtracting an estimate of its bias.

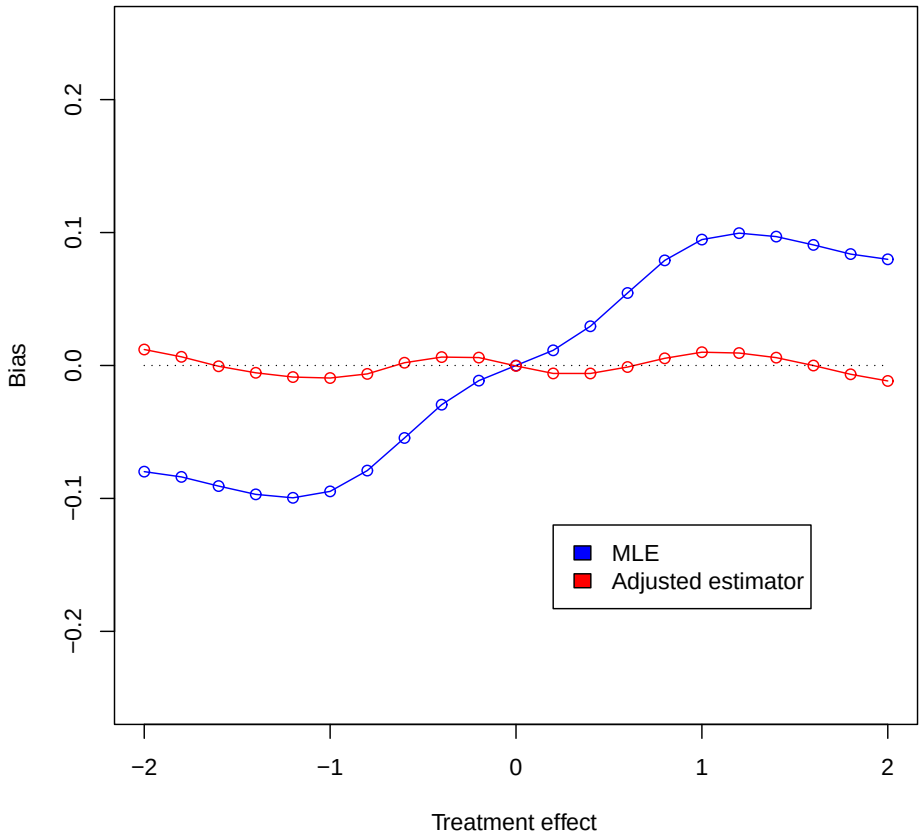
Since the true θ is unknown, the bias of the MLE is estimated by $b(\hat{\theta}_M)$.

The adjusted estimator is then

$$\hat{\theta}_{adj} = \hat{\theta}_M - b(\hat{\theta}_M).$$

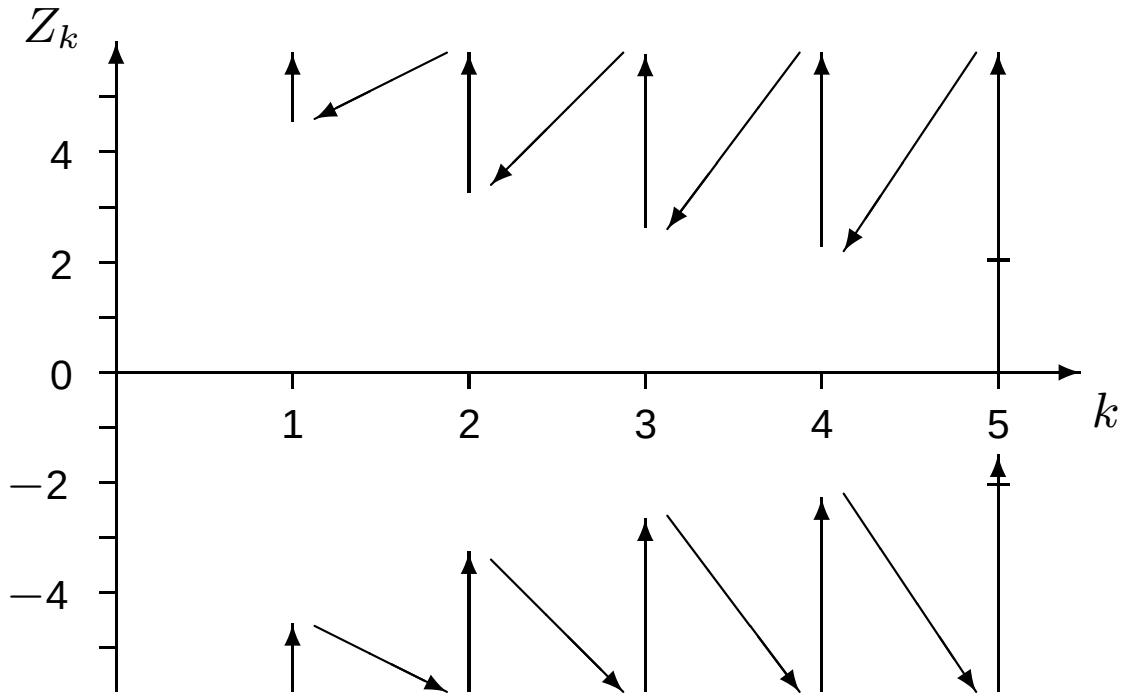
Adjusted estimator of θ after a 5 group O'Brien & Fleming test

Simulation results show that Whitehead's adjusted estimator has much smaller bias than the MLE on which it is based.



(b) P-value for H_0: $\theta = 0$ after a group sequential test

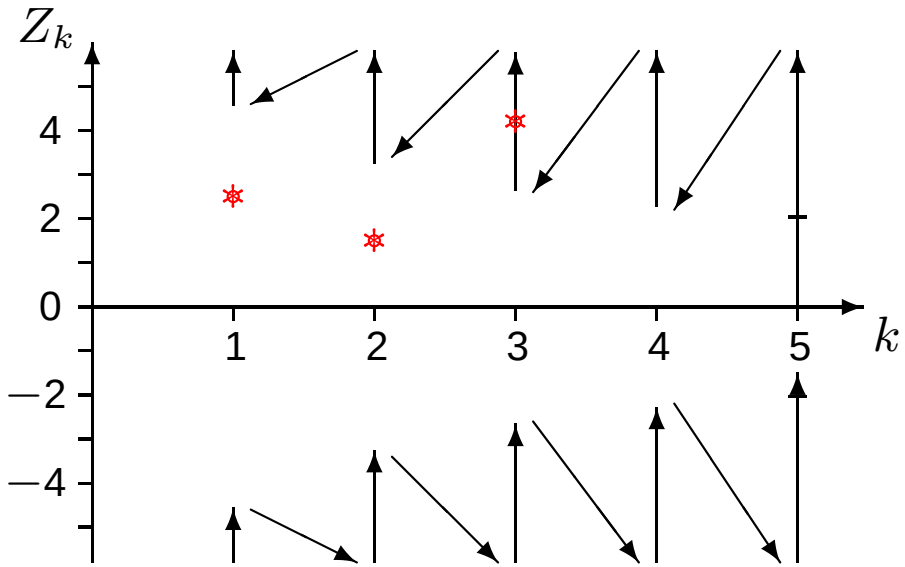
P-values and confidence intervals are based on an ordering of the sample space.



We shall use the “stage-wise” ordering depicted above.

P-value on termination of a group sequential test

The one-sided P-value for testing H_0: $\theta = 0$ against $\theta > 0$ is the probability under H_0 of observing such a high outcome in the specified ordering.



For example, if the test stops at analysis 3 with $Z_3 = 4.2$, the one-sided P-value is

$$P_{\theta=0}\{Z_1 \geq 4.56 \text{ or } Z_2 \geq 3.23 \text{ or } Z_3 \geq 4.2\} = 0.00063.$$

P-value on termination of a group sequential test

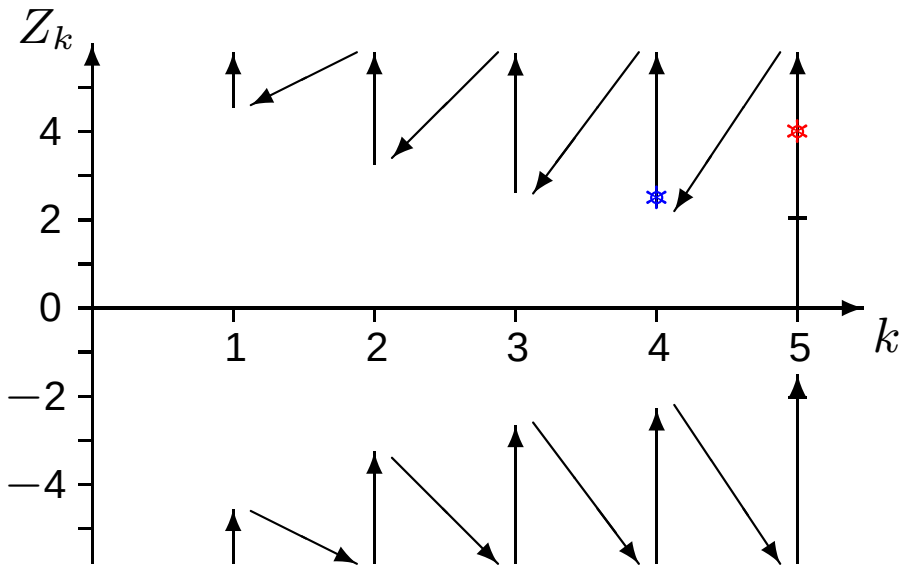
The one-sided P-value for testing H_0: $\theta = 0$ against $\theta < 0$ is the probability under H_0 of observing such a low outcome in the specified ordering.

The two-sided P-value for testing H_0: $\theta = 0$ against $\theta \neq 0$ is two times the smaller of the one-sided P-values.

Symmetry of the stopping boundary in our example implies that, on stopping at analysis 3 with $Z_3 = 4.2$, the two-sided P-value is

$$P_{\theta=0}\{|Z_1| \geq 4.56 \text{ or } |Z_2| \geq 3.23 \text{ or } |Z_3| \geq 4.2\} = 0.0013.$$

P-value on termination of a group sequential test



In the stage-wise ordering, stopping at a later analysis with a large excess over the boundary can give a higher MLE of θ, but lower rank in the sample space ordering.

For the two outcomes shown above,

$$Z_4 = 2.5 \quad \text{and} \quad \hat{\theta}_M = 0.85,$$

$$Z_5 = 4.0 \quad \text{and} \quad \hat{\theta}_M = 1.22.$$

Should these points be in the opposite order?

P-value on termination of a group sequential test

Other sample space orderings are possible — and several have been proposed.

Since the monotone likelihood ratio property does not hold for this sample space, there is no “best” ordering.

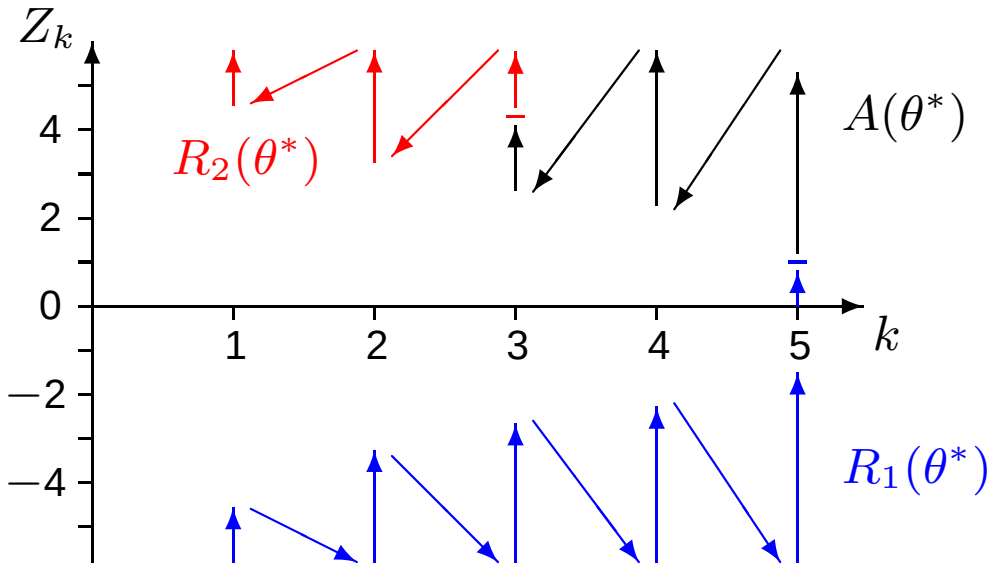
The stage-wise ordering can be applied without knowledge of what future group sizes would have been, making it a natural choice for use with error spending designs for group sequential tests.

Under the above definitions, based on a specific ordering of the sample space:

- The P-value has a $U(0,1)$ distribution under H_0.
- If the group sequential test has two-sided type I error probability α, then the two-sided P-value is $\leq \alpha$ precisely when the test stops with rejection of H_0.
- The two-sided P-value tends to be low when θ is away from zero.

(c) Confidence interval on termination of a group sequential test

A $100(1-\alpha)$% confidence interval (CI) for θ is obtained by inverting a family of level α, two-sided hypothesis tests.



For each θ^*, partition the sample space into regions $R_1(\theta^*)$, $A(\theta^*)$ and $R_2(\theta^*)$, containing outcomes of increasing rank in the sample space ordering, with

$$P_{\theta^*}\{R_1(\theta^*)\} = \alpha/2, \quad P_{\theta^*}\{A(\theta^*)\} = 1-\alpha, \quad P_{\theta^*}\{R_2(\theta^*)\} = \alpha/2.$$

Confidence interval on termination of a group sequential test

The null hypothesis $H_0(\theta^*)$: $\theta = \theta^*$ is accepted for outcomes in $A(\theta^*)$.

The CI for θ is the set of all values θ^* for which $H_0(\theta^*)$ is accepted:

$$\{\theta^*\text{: Observed outcome } \in A(\theta^*)\}$$

Direct definition:

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

The CI contains those values of θ for which (k^*, Z^*) is in the middle $(1-\alpha)$ of the probability distribution of outcomes under θ.

This can be seen to be the interval (θ_1, θ_2) where

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

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

Confidence interval on termination of a group sequential test

Example:

If the test stops at analysis 3 with $Z_3 = 4.2$, the 95% confidence interval for θ is

$$(0.60,\ 2.32)$$

using the stage-wise ordering.

In contrast:

The "naive" 95% fixed sample CI is

$$(0.88,\ 2.42).$$

But, it is ***not appropriate*** to use this fixed sample interval.

Its derivation does not take account of the special nature of the sample space.

Thus, coverage probability of the interval calculated in this manner is *not* 95%.

Consistency of hypothesis testing and CI on termination

Suppose a group sequential study is conducted to test H_0: $\theta = 0$ vs $\theta \neq 0$ with type I error probability α.

Then, a $(1 - \alpha)$ confidence interval on termination should contain $\theta = 0$ if and only if H_0 is accepted.

This happens automatically if outcomes for which we reject H_0 are at the top and bottom ends of the sample space ordering — and any sensible ordering does this.

Note that a naive $(1 - \alpha)$ level CI on termination fails to include $\theta = 0$ if an *unadjusted* α level significance test rejects H_0. Due to the "multiple looks" effect, this can occur with probability considerably higher than α.

Should inference be “conditional”?

Strickland & Casella (*Biometrical Journal*, 2003) and Fan & DeMets (*J. Biopharm. Statistics*, 2006) consider conditional inference following a group sequential test.

R.A. Fisher (*Statistical Methods and Scientific Inference*, 1959) discussed “conditioning on recognizable subsets”. The stopping time of a group sequential test defines such a subset of outcomes, and one may consider properties of inferences conditional on the stage at which the study stops.

However, the stopping time itself is informative about θ and, for certain types of boundary, can even imply the decision to accept or reject H_0.

What if you only report a confidence interval for θ when you have rejected H_0?

Would you consider inference conditional on rejecting H_0 in a fixed sample study?

I shall focus on *unconditional* inference, aiming for consistency with the conclusions of the basic hypothesis test. This will be a significant choice for trials with multiple parameters and null hypotheses.

2. Adaptive designs: The combination test

Bauer & Köhne (*Biometrics*, 1994) introduced the combination test as a key tool for constructing adaptive trial designs. We describe this method using Z-statistics:

Define the null hypothesis H_0 (with a one-sided alternative).

Design Stage 1, fixing sample size and test statistic for this stage.

Stage 1

Observe the Z-value Z_1 for testing H_0.

Design Stage 2 in the light of Stage 1 data.

Stage 2

Observe Z_2 for testing H_0, based on Stage 2 data only.

Clearly, Z_1 has the usual $N(0,\,1)$ distribution under H_0.

Under H_0, $Z_2 \sim N(0,\,1)$ conditionally — and hence unconditionally — on the Stage 2 design, and Z_2 is independent of Z_1.

The inverse normal combination test

Weights w_1 and w_2, with $w_1^2 + w_2^2 = 1$, are stipulated before the start of the study.

Since Z_1 and Z_2 are independent $N(0,\, 1)$ variates under H_0, it follows that

$$w_1\, Z_1 + w_2\, Z_2 \;\sim\; N(0,\, 1).$$

Hence, the Z-values can be combined in an overall test, rejecting H_0 if

$$w_1\, Z_1 + w_2\, Z_2 \;>\; 1 - \Phi(\alpha).$$

This test has type I error rate α under H_0, despite allowing mid-study modifications which do not necessarily follow pre-planned rules.

Bauer & Köhne refer to this as the *inverse normal test* (in contrast to their *inverse* χ^2 *test* based on $P_1\, P_2 = \{1 - \Phi(Z_1)\}\,\{1 - \Phi(Z_2)\}$).

Confidence interval following a combination test

Consider the two-treatment comparison with normal responses introduced earlier.

Suppose an adaptive trial design with an inverse normal combination test is used to test H_0: $\theta \leq 0$ against $\theta > 0$.

We shall construct a $(1-\alpha)$ level upper CI for θ.

With n_1 subjects per treatment in Stage 1, $\hat{\theta}_1 = \sum_{i=1}^{n_1} (X_i - Y_i)/n_1$ and

$$Z_1 \;=\; \frac{\hat{\theta}_1}{\sqrt{(2\,\sigma^2/n_1)}} \;\sim\; N(0,\,1) \quad \text{under} \quad \theta = 0.$$

To test $H_0(\theta^*)$: $\theta = \theta^*$ against $\theta > \theta^*$, define

$$Z_1(\theta^*) \;=\; \frac{\hat{\theta}_1 - \theta^*}{\sqrt{(2\,\sigma^2/n_1)}} \;=\; Z_1 - \frac{\theta^*}{\sqrt{(2\,\sigma^2/n_1)}}.$$

Then, $Z_1(\theta^*) \;\sim\; N(0,\,1)$ under $H_0(\theta^*)$.

Confidence interval following a combination test

The second stage of the trial is designed in the light of Stage 1 data.

With n_2 subjects per treatment in Stage 2 and $\hat{\theta}_2$ based on these subjects only,

$$Z_2(\theta^*) \;=\; \frac{\hat{\theta}_2 - \theta^*}{\sqrt{(2\,\sigma^2/n_2)}} \;=\; Z_2 - \frac{\theta^*}{\sqrt{(2\,\sigma^2/n_2)}}.$$

Under $H_0(\theta^*)$: $\theta = \theta^*$, $Z_2(\theta^*) \sim N(0,\,1)$ and is independent of $Z_1(\theta^*)$.

Applying the combination test, $H_0(\theta^*)$ is rejected if

$$w_1\,Z_1(\theta^*) + w_2\,Z_2(\theta^*) \;>\; 1 - \Phi(\alpha).$$

A $(1-\alpha)$ level, upper CI for θ is the set of values θ^* for which $H_0(\theta^*)$ is accepted, i.e.,

$$\{\theta^*\colon\; w_1\,Z_1(\theta^*) + w_2\,Z_2(\theta^*) \;\leq\; 1 - \Phi(\alpha)\}.$$

Confidence interval following a combination test

This methodology can be extended to accommodate additional features of adaptive designs based on combination tests.

Early stopping

The two-stage design may include a stopping rule:

If $Z_1 \geq b_1$,	stop and reject H_0,
If $a_1 < Z_1 < b_1$,	continue to Stage 2,
If $Z_1 \leq a_1$	stop and accept H_0.

This rule can be expressed in terms of $Z_1(\theta^*)$ since

$$Z_1 \geq b_1 \;\Leftrightarrow\; Z_1(\theta^*) \geq b_1(\theta^*) = b_1 - \theta^*/\sqrt{(2\,\sigma^2/n_1)}$$

and

$$Z_1 \leq a_1 \;\Leftrightarrow\; Z_1(\theta^*) \leq a_1(\theta^*) = a_1 - \theta^*/\sqrt{(2\,\sigma^2/n_1)}.$$

Confidence interval following a combination test

Early stopping, continued

Now order the sample space according to

1. $Z_1(\theta^*) \leq a_1(\theta^*)$ by increasing values of $Z_1(\theta^*)$,
2. $a_1(\theta^*) < Z_1(\theta^*) < b_1(\theta^*)$ by increasing $w_1 Z_1(\theta^*) + w_2 Z_2(\theta^*)$,
3. $Z_1(\theta^*) \geq b_1(\theta^*)$ by increasing $Z_1(\theta^*)$.

This has similarities to the stage-wise ordering of the sample space of a group sequential test.

Now define the rejection region for $H_0(\theta^*)$: $\theta = \theta^*$ as the set of outcomes at the top of this ordering with probability α under $\theta = \theta^*$.

As before, the $(1 - \alpha)$ level, upper CI for θ is the set of values θ^* for which $H_0(\theta^*)$ is accepted.

Confidence interval following a combination test

Combination tests with more than two stages

Lehmacher & Wassmer (*Biometrics*, 1999) extend Bauer & Köhne's construction to create adaptive group sequential designs combining Z-statistics from several stages — with the opportunity of re-design at each stage.

Brannath, Posch & Bauer (*JASA*, 2002) discuss "recursive" combination tests. That is, two-stage tests in which the second stage can be sub-divided into two stages, and iterating this process gives multiple stages.

Brannath et al. show how to construct an upper confidence interval after a recursive combination test. Since a Lehmacher & Wassmer test can be expressed as a recursive combination test, this approach will provide a CI on termination.

Brannath, Mehta & Posch (*Biometrics*, 2009) consider other adaptive designs based on preserving the conditional type I error probability. They propose methods CIs on termination based on the stage-wise ordering which are (slightly) conservative.

3. Inference following an enrichment design

A trial protocol defines a population that may benefit from the new treatment.

It is believed the treatment could be particularly effective in a sub-population defined by a physiological or genetic biomarker.

Enrichment: Restricting recruitment to a sub-population

At an interim analysis, the options are:

Stop the trial for futility,

Continue as originally planned,

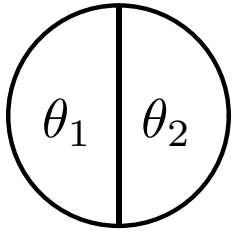
Restrict the remainder of the study to the defined sub-population.

Restricting recruitment to the sub-population will affect the licence that a positive outcome can support.

The possibility of testing more than one null hypothesis means a multiple testing procedure must be used.

Enrichment: Example

Sub-population 1 (proportion λ_1)



Rest of the population (proportion λ_2)

Overall treatment effect is $\theta_3 = \lambda_1\theta_1 + \lambda_2\theta_2$.

We may wish to test:

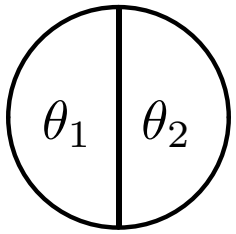
The null hypothesis for the full population, $H_{0,3}$: $\theta_3 \leq 0$ vs $\theta_3 > 0$,

The null hypothesis for sub-population 1, $H_{0,1}$: $\theta_1 \leq 0$ vs $\theta_1 > 0$.

Enrichment: Example

Sub-population 1

(proportion λ_1)



Rest of the population

(proportion λ_2)

First, consider a design testing for a whole population effect, $\theta_3 = \lambda_1\theta_1 + \lambda_2\theta_2$, in the case $\lambda_1 = \lambda_2 = 0.5$.

The test of $H_{0,3}$: $\theta_3 \leq 0$ has one-sided type I error probability 0.025 and sample size is set to achieve power 0.9 at $\theta_3 = 20$.

An interim analysis is conducted after half the planned sample size.

If $\hat{\theta}_3 < 0$ at the interim analysis, stop for futility with acceptance of $H_{0,3}$.

Enrichment: Example

Properties of design for the whole population effect.

θ_1	θ_2	θ_3	*Power for* $H_{0,3}$*:* $\theta_3 \leq 0$
20	20	20	0.90
10	10	10	0.37
20	0	10	0.37

The third row represents a case where only the sub-population benefits from the new treatment, so θ_1 is high and θ_2 is low.

Our aim is to identify such cases and switch resources to the sub-population in order to improve power.

We need to specify a sampling rule which states when to continue in the full population and when to “enrich” the sub-population.

Enrichment: An adaptive sampling rule

At Stage 1, if $\hat{\theta}_3 < 0$ stop to accept $H_{0,3}$: $\theta_3 \leq 0$.

If $\hat{\theta}_3 > 0$ and the trial continues:

If $\hat{\theta}_2 < 0$ and $\hat{\theta}_1 > \hat{\theta}_2 + 8$ Restrict to sub-population 1 and test $H_{0,1}$ only.

Else, Continue with full population, test $H_{0,1}$ and $H_{0,3}$.

The same *total* sample size for Stage 2 is retained in both cases, increasing the numbers for the sub-population when enrichment occurs.

This sampling rule defines the sample space, so we are already in a position to consider estimation of θ_1 and θ_3 on termination.

We shall complete the definition of the testing procedure when we move on to the related topic of CIs on termination.

Estimation after an enrichment design

Consider estimating θ_1, θ_2 and θ_3, the treatment effects in the sub-population, the complement of the sub-population, and the full population.

Maximum likelihood estimates are obtained as follows:

If the trial stops at Stage 1

Base $\hat{\theta}_{1,M}$, $\hat{\theta}_{2,M}$ and $\hat{\theta}_{3,M}$ on Stage 1 data.

If the trial continues to Stage 2 with the full population

Base $\hat{\theta}_{1,M}$, $\hat{\theta}_{2,M}$ and $\hat{\theta}_{3,M}$ on combined Stage 1 and Stage 2 data.

If the trial continues to Stage 2 with only the sub-population

Base $\hat{\theta}_{1,M}$ on combined Stage 1 and Stage 2 data,

Base $\hat{\theta}_{2,M}$ on Stage 1 data

Set $\hat{\theta}_{3,M} = \lambda_1 \hat{\theta}_{1,M} + \lambda_2 \hat{\theta}_{2,M}$.

Estimation after an enrichment design

We obtain $\hat{\theta}_{1,M}$, $\hat{\theta}_{2,M}$ and $\hat{\theta}_{3,M}$ for all trials, irrespective of the interim decision to stop, continue, or focus on the sub-population.

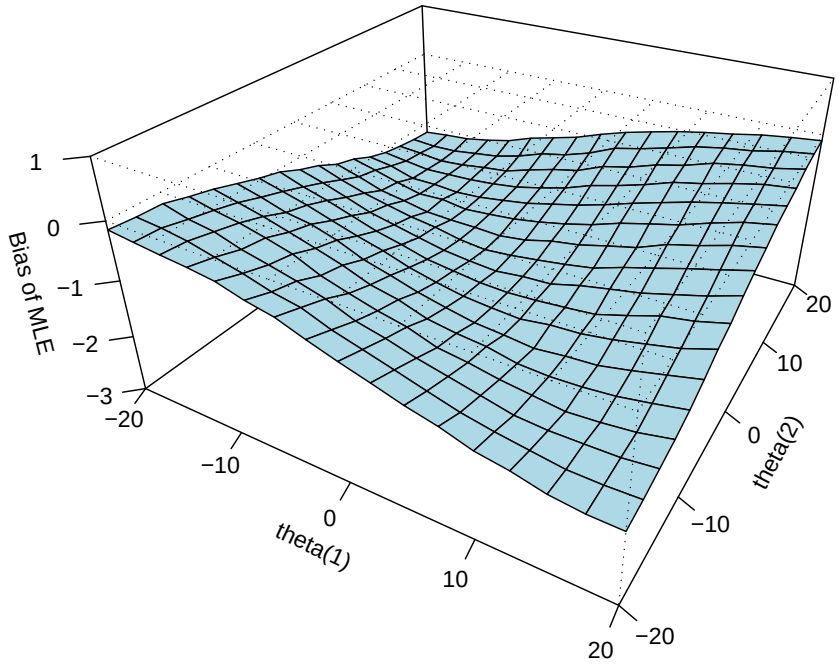
Thus, we are considering *unconditional* estimation.

We shall look, in particular, at estimates of θ_1 and θ_3.

We should expect bias in the MLEs:

> If $\hat{\theta}_1$ from Stage 1 is high, there is a greater chance of focusing on the sub-population and increasing its sample size.
>
> If $\hat{\theta}_1$ from Stage 1 is low, there is less chance of focusing on the sub-population, so this $\hat{\theta}_1$ keeps a high weight in the MLE.
>
> This will produce a negative bias in $\hat{\theta}_{1,M}$.
>
> Similar reasoning indicates negative bias in $\hat{\theta}_{2,M}$ and, hence, in $\hat{\theta}_{3,M}$.

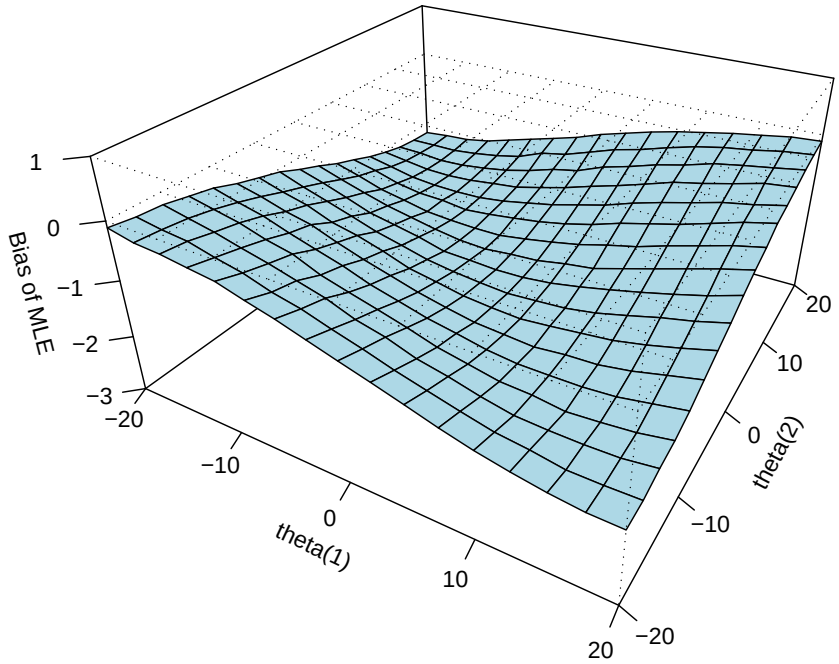
Estimation after an enrichment design



Bias of $\hat{\theta}_{1,M}$, MLE for the sub-population treatment effect

Biases of -2 represent 10% of the effect size under investigation.

Estimation after an enrichment design



Bias of $\hat{\theta}_{3,M}$, MLE for the full population treatment effect

Biases of -2 represent 10% of the effect size under investigation.

Correcting the bias of the MLE

Whitehead's (1986) method can be applied in more than one dimension.

Write $\theta = (\theta_1, \theta_2, \theta_3)$, where $\theta_3 = \lambda_1 \theta_1 + \lambda_2 \theta_2$.

Denote the bias functions of the MLEs of θ_1 and θ_3 by

$$b_1(\theta) = E_\theta(\hat{\theta}_{1,M}) - \theta_1 \quad \text{and} \quad b_3(\theta) = E_\theta(\hat{\theta}_{3,M}) - \theta_3.$$

(Note that the bias depends on both θ_1 and θ_2.)

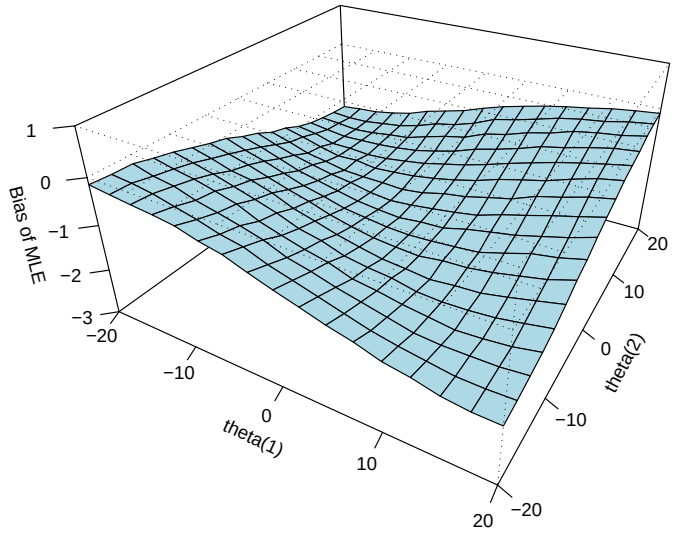
We can estimate the functions $b_1(\theta)$ and $b_3(\theta)$ by simulation.

Hence, we obtain adjusted estimators:

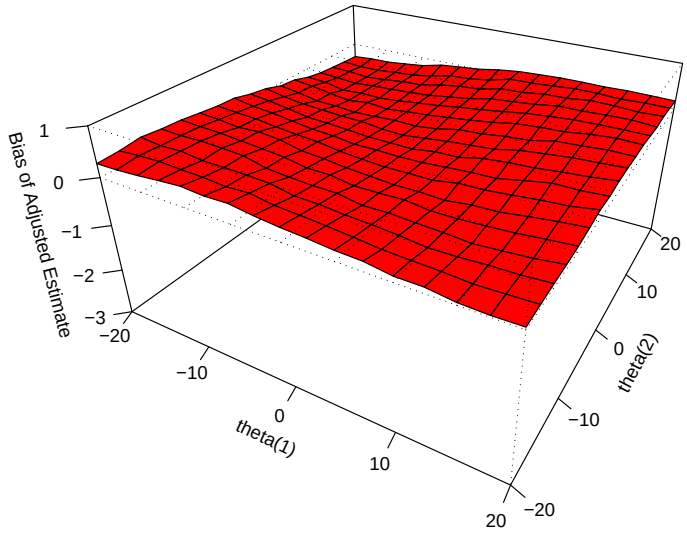
$$\hat{\theta}_{1,adj} = \hat{\theta}_{1,M} - b_1(\hat{\theta}_M) \quad \text{and} \quad \hat{\theta}_{3,adj} = \hat{\theta}_{3,M} - b_3(\hat{\theta}_M).$$

Estimation after an enrichment design

The adjusted estimator $\hat{\theta}_{1,adj}$ has much smaller bias than the MLE, $\hat{\theta}_{1,M}$.



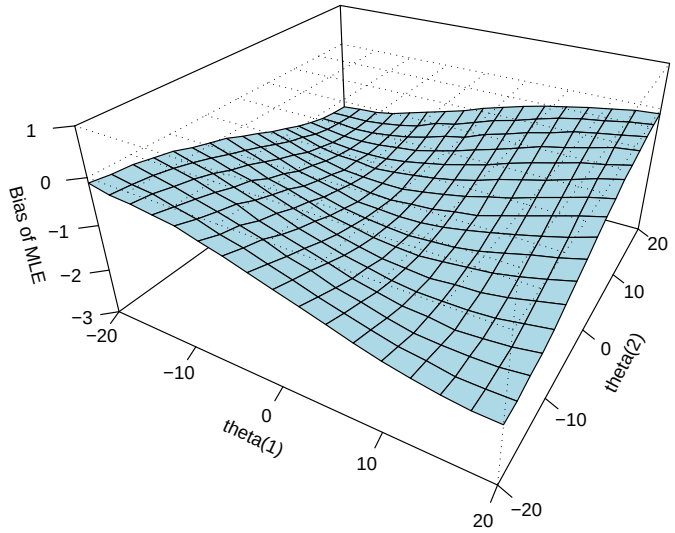
Bias of $\hat{\theta}_{1,M}$, MLE for the sub-population effect θ_1



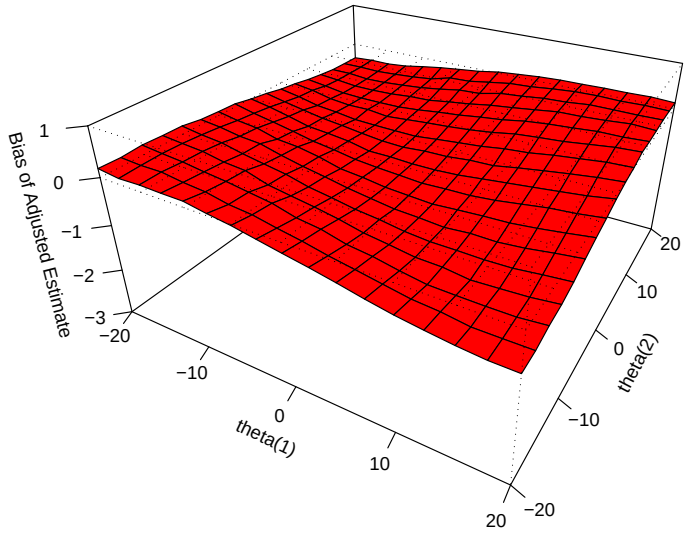
Bias of $\hat{\theta}_{1,adj}$, adjusted estimator of θ_1

Estimation after an enrichment design

The adjusted estimator $\hat{\theta}_{3,adj}$ has much smaller bias than the MLE, $\hat{\theta}_{3,M}$.



Bias of $\hat{\theta}_{3,M}$, MLE for the full population effect θ_3



Bias of $\hat{\theta}_{3,adj}$, adjusted estimator of θ_3

Enrichment: Multiple hypothesis testing

Recall that interest lies in showing a treatment effect in the full population or, failing that, in the sub-population.

Thus, we may wish to test:

For the full population, $H_{0,3}$: $\theta_3 \leq 0$ vs $\theta_3 > 0$,

For the sub-population, $H_{0,1}$: $\theta_1 \leq 0$ vs $\theta_1 > 0$.

The responses observed in Stage 1 will determine which hypotheses are of interest at the end of the trial and the sample sizes available for testing these.

An appropriate multiple testing procedure is needed to provide proper control of the false positive rate. A “closed testing procedure” can achieve this.

Closed testing procedures (Marcus et al, *Biometrika*, 1976)

With hypotheses H_i: $\theta_i \leq 0,\ i = 1, \ldots, k,$ define the intersection hypothesis $H_I = \cap_{i \in I} H_i$ for each subset I of $\{1, \ldots, k\}$.

Construct a level α test of each intersection hypothesis H_I: this test rejects H_I with probability at most α whenever all hypotheses specified in H_I are true.

Closed testing procedure

The hypothesis H_i: $\theta_i \leq 0$ is rejected overall if, and only if, H_I is rejected for every set I containing index i.

This procedure controls the family-wise error rate strongly at level α, i.e.,

$$Pr\{\text{Reject any true } H_i\} \leq \alpha \quad \text{for all } (\theta_1, \ldots, \theta_k).$$

With such strong control, the probability of choosing to focus on the parameter θ_{i^*} and then falsely claiming significance for null hypothesis H_{i^*} is at most α.

Enrichment: Multiple hypothesis testing

A closed testing procedure will require tests for 3 hypotheses:

$$\begin{aligned} H_{0,3}\text{:} &\quad \theta_3 \leq 0 \\ H_{0,1}\text{:} &\quad \theta_1 \leq 0 \\ H_{0,13}\text{:} &\quad \theta_1 \leq 0 \text{ and } \theta_3 \leq 0. \end{aligned}$$

Given the definition $\theta_3 = \lambda_1 \theta_1 + \lambda_2 \theta_2$, any subset of these three hypotheses may be true.

We test all three hypotheses in two-stage tests, stopping to accept a hypothesis at Stage 1 if the Z-statistic is negative.

With Z-statistics Z_1 and Z_2 from Stages 1 and 2, H_i is rejected if

$$Z_1 \geq 0 \quad \text{and} \quad \tfrac{1}{\sqrt{2}} Z_1 + \tfrac{1}{\sqrt{2}} Z_2 \geq 1.95.$$

Enrichment: Multiple hypothesis testing

For Stage $i = 1$ and 2, let:

$\hat{\theta}_{i,1}$ and $\hat{\theta}_{i,3}$ be estimates of θ_1 and θ_3 obtained from responses in Stage i,

$Z_{i,1}$ and $Z_{i,3}$ denote Z-statistics for $H_{0,1}$ and $H_{0,3}$ based on $\hat{\theta}_{i,1}$ and $\hat{\theta}_{i,3}$.

When continuing with the full population, we use Z-statistics:

	Stage 1	*Stage 2*
$H_{0,3}$	$Z_{1,3}$	$Z_{2,3}$
$H_{0,13}$	$Z_{1,13}$	$Z_{2,3}$

Here, we define $Z_{1,13}$ as a weighted combination of $Z_{1,1}$ and $Z_{1,3}$,

$$Z_{1,13} = (Z_{1,1} + Z_{1,3})/\sqrt{(2+\sqrt{2})},$$

and this has a $N(0,1)$ distribution under $H_{0,13}$.

Enrichment: Multiple hypothesis testing

When switching to sub-population 1, we use:

	Stage 1	*Stage 2*
$H_{0,1}$	$Z_{1,1}$	$Z_{2,1}$
$H_{0,13}$	$Z_{1,13}$	$Z_{2,1}$

Even though the goal of demonstrating a treatment effect in the full population has been abandoned, it is still necessary to reject the intersection hypothesis $H_{0,13}$ in order for the overall procedure to reject $H_{0,1}$.

This procedure does improve power when the new treatment is effective in the sub-population only.

Enrichment: Power of non-adaptive and adaptive designs

				Non-adaptive	***Adaptive***		
	θ_1	θ_2	θ_3	*Full pop^n*	*Sub-pop 1 only*	*Full pop^n*	*Total*
1.	30	0	15	**0.68**	0.47	0.41	**0.88**
2.	20	0	10	**0.37**	0.33	0.25	**0.58**
3.	20	20	20	**0.90**	0.04	0.83	**0.87**
4.	20	10	15	**0.68**	0.15	0.57	**0.72**

Cases 1 & 2: Overall power is increased. Testing focuses (correctly) on $H_{0,1}$, but it is still possible to find an effect (wrongly) for the full population.

Case 3: Restricting to the sub-population slightly reduces power for finding an effect in the full population.

Case 4: Adaptation improves overall power a little.

Enrichment design: Confidence intervals on termination

We desire a $(1-\alpha)$ level joint upper confidence interval for θ_1 and θ_3 on conclusion of the enrichment design.

A rectangular interval has the form

$$\theta_1 \in (\psi_1, \infty), \quad \theta_3 \in (\psi_3, \infty).$$

For consistency with the outcomes of hypothesis tests, we require:

If $H_{0,1}$: $\theta_1 \leq 0$ is rejected, then $\psi_1 > 0$,

If $H_{0,3}$: $\theta_3 \leq 0$ is rejected, then $\psi_3 > 0$.

As in the univariate case, a CI can be formed from a family of hypothesis tests by taking the set of parameter values accepted by their hypothesis tests.

Enrichment design: Confidence intervals on termination

Posch et al. (*Statistics in Medicine*, 2005) discuss the construction of a joint CI for multiple parameters after an adaptively designed trial.

They note that it is difficult to achieve the desired consistency property between joint CIs and hypothesis test outcomes.

A route to a solution:

One reason that creating satisfactory CIs can be problematic is that elements of the closed testing procedure are defined without thinking ahead to construction of CIs.

An alternative approach is to start with a good method for producing a joint CI on termination, then define a closed testing procedure to fit with this.

Hayter & Hsu (*JASA*, 1994) present such a method for fixed sample size designs, constructing joint CIs which are consistent with stepwise decision procedures.

Enrichment design: Confidence intervals on termination

A route to a solution:

A key feature of Hayter & Hsu's method is that, in tests of $H_0(\theta^*)$: $\theta = \theta^*$, they use different forms of test in different regions of the parameter space.

For our example, I propose:

To test $H_0(\theta^*)$: $\theta = \theta^*$ *with* $(\theta_1^* \leq 0,\, \theta_2^* \leq 0)$ *or* $(\theta_1^* > 0,\, \theta_2^* > 0)$

Test H_0: $\theta_1 = \theta_1^*$ and H_0: $\theta_2 = \theta_2^*$ separately, then apply Bonferroni's rule and reject $H_0(\theta^*)$ if at least one significance level is less than $\alpha/2$.

To test $H_0(\theta^*)$: $\theta = \theta^*$ *with* $(\theta_1^* \leq 0,\, \theta_2^* > 0)$

Test H_0: $\theta_1 = \theta_1^*$ and reject $H_0(\theta^*)$ if the significance level is less than α.

Enrichment design: Confidence intervals on termination

A route to a solution:

and finally,

To test $H_0(\theta^*)$*:* $\theta = \theta^*$ *with* $(\theta_1^* > 0,\ \theta_2^* \leq 0)$

Test H_0: $\theta_2 = \theta_2^*$ and reject $H_0(\theta^*)$ if the significance level is less than α.

These types of test can be incorporated in the definitions of tests of $H_{0,1},\ H_{0,3}$ and $H_{0,13}$ in our example of an enrichment design.

It will then follow that the joint CI for θ_1 and θ_3 on termination has the desired properties of excluding $\theta_1 = 0$ when $H_{0,1}$ is rejected and excluding $\theta_3 = 0$ when $H_{0,3}$ is rejected.

Conclusions

There are methods for obtaining approximately unbiased point estimates, P-values and confidence intervals after a group sequential test.

P-values and confidence intervals for a single treatment effect can also be obtained after an adaptive group sequential design.

Methods for point estimation extend simply to multiple parameters.

Construction of joint confidence intervals for two or more parameters after an adaptive design poses problems but these can be overcome.