Adaptive population enrichment: switching to a sub-population in response to interim trial data

Christopher Jennison

Department of Mathematical Sciences,

University of Bath, UK

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

2nd Conference of the Central European Network — CEN 2011

Zurich, September 2011

Motivating example

Consider a new treatment developed to disrupt a disease's biological pathway.

Patients with high levels of a biomarker associated with this pathway should gain particular benefit, but the treatment's wider action may also help the general patient population.

In a clinical trial with *enrichment* we

- 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.

Results may support a licence for the full population or just the sub-population.

Overview of this talk

1. Adaptation to focus on a sub-population: enrichment designs

    Decision rules

    Closed testing procedures

    Power of adaptive designs

2. Estimation following an enrichment design

    Bias of MLEs

    Adjusted estimates

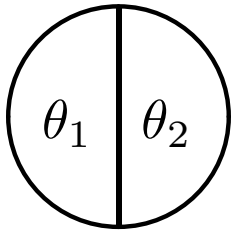
3. Confidence intervals following an enrichment design

    Problems of consistency between CIs and overall hypothesis tests

    Consistent CIs

1. Enrichment: Switching to a patient sub-population

Sub-population (proportion λ_1)



Rest of the population (proportion λ_2)

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

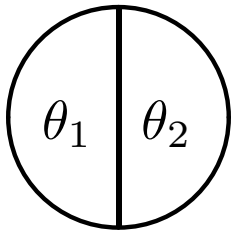
We may wish to test:

The null hypothesis for the full population, H_3: $\theta_3 \leq 0$ vs $\theta_3 > 0$,

The null hypothesis for the sub-population, H_1: $\theta_1 \leq 0$ vs $\theta_1 > 0$.

Enrichment: Example

Sub-population (proportion λ_1)



Rest of the population (proportion λ_2)

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

The design has two analyses and one-sided type I error probability 0.025.

Sample size is set to achieve power 0.9 at $\theta_3 = 20$.

Data in each stage are summarised by a Z-value:

	Stage 1	*Stage 2*	*Overall*
H_3: $\theta_3 \leq 0$	$Z_{1,3}$	$Z_{2,3}$	$Z_3 = \frac{1}{\sqrt{2}}Z_{1,3} + \frac{1}{\sqrt{2}}Z_{2,3}$

Enrichment: Example

Two stage design testing for a whole population effect, θ_3.

	Stage 1	*Stage 2*	*Overall*
H_3: $\theta_3 \leq 0$	$Z_{1,3}$	$Z_{2,3}$	$Z_3 = \frac{1}{\sqrt{2}} Z_{1,3} + \frac{1}{\sqrt{2}} Z_{2,3}$

Decision rules:

If $Z_{1,3} < 0$ Stop at Stage 1, Accept H_3

If $Z_{1,3} \geq 0$ Continue to Stage 2, then

- If $Z_3 < 1.95$ Accept H_3
- If $Z_3 \geq 1.95$ Reject H_3

Enrichment: Example

Assume the sub-population comprises half the total population, so $\lambda_1 = \lambda_2 = 0.5$.

Properties of design for the whole population effect, θ_3:

θ_1	θ_2	θ_3	*Power for* H_3: $\theta_3 \leq 0$
20	20	20	0.90
10	10	10	0.37
20	0	10	0.37

Is it feasible to identify at Stage 1 that θ_3 is low but θ_1 may be higher, so one might switch resources to test a sub-population?

Enrichment: A closed testing procedure

We wish to be able to consider two null hypotheses:

H_3: $\theta_3 \leq 0$ Treatment is not effective in the whole population,

H_1: $\theta_1 \leq 0$ Treatment is not effective in the sub-population.

Since $\theta_3 = 0.5\,\theta_1 + 0.5\,\theta_2$, either of H_1 and H_3 may be true on its own.

To apply a ***closed testing procedure*** (Marcus et al, *Biometrika*, 1976) we also need a test of the intersection hypothesis:

$$H_{13}: \quad \theta_1 \leq 0 \text{ and } \theta_3 \leq 0.$$

Then to reject H_1 overall, while protecting the family-wise type I error rate, we need to reject both H_1 and H_{13} in individual tests at significance level α.

Similarly, we can reject H_3 overall if both H_3 and H_{13} are rejected in level α tests.

Enrichment: An adaptive design

At Stage 1, if $\hat{\theta}_3 < 0$, stop to accept H_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_1 only, needing to reject H_1 and H_{13}.
Else,	Continue with full population and test H_3, needing to reject H_3 and H_{13}.

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

Enrichment: An adaptive design

Each null hypothesis, H_i say, is tested in a 2-stage group sequential test.

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

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

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

	Stage 1	*Stage 2*
H_3	$Z_{1,3}$	$Z_{2,3}$
H_{13}	$Z_{1,3}$	$Z_{2,3}$

where $Z_{i,3}$ is based on $\hat{\theta}_3$ from responses in Stage i.

So, there is no change from the original test of H_3.

Enrichment: An adaptive design

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

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

When switching to the sub-population, we use:

	Stage 1	*Stage 2*
H_1	$Z_{1,1}$	$Z_{2,1}$
H_{13}	$Z_{1,3}$	$Z_{2,1}$

where $Z_{i,j}$ is based on $\hat{\theta}_j$ from responses in Stage i.

The need to reject the intersection hypothesis H_{13} adds an extra requirement to the simple test of H_1.

Simulation results: Power of non-adaptive and adaptive designs

				Non-adaptive	***Adaptive***		
	θ_1	θ_2	θ_3	*Full pop^n*	*Sub-pop^n only*	*Full pop^n*	*Total*
1.	30	0	15	**0.68**	0.43	0.42	**0.85**
2.	20	0	10	**0.37**	0.24	0.26	**0.51**
3.	20	20	20	**0.90**	0.03	0.87	**0.90**
4.	20	10	15	**0.68**	0.11	0.60	**0.71**

Cases 1 & 2: Testing focuses (correctly) on H_1, but it is still possible to find an effect (wrongly) for the full population. Overall power is increased.

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

Case 4: Adaptation improves overall power a little.

Increasing power for finding a sub-population effect

Greater power for the sub-population can be achieved by using $Z_{1,1}$ rather than $Z_{1,3}$ as the Stage 1 statistic in the test of H_{13}.

This gives the following results:

				Non-adaptive	***Adaptive***		
	θ_1	θ_2	θ_3	*Full pop^n*	*Sub-pop^n only*	*Full pop^n*	*Total*
1.	30	0	15	**0.68**	0.47	0.40	**0.87**
2.	20	0	10	**0.37**	0.35	0.23	**0.58**
3.	20	20	20	**0.90**	0.04	0.74	**0.78**
4.	20	10	15	**0.68**	0.16	0.51	**0.56**

Benefits in Case 2 are balanced by loss of overall power in Cases 3 and 4.

Increasing power for finding a sub-population effect

As a compromise between the two previous methods, a combination* of $Z_{1,3}$ and $Z_{1,1}$ may be used as the Stage 1 statistic for the test of H_{13}.

This leads to the following results:

				Non-adaptive	***Adaptive***		
	θ_1	θ_2	θ_3	*Full pop^n*	*Sub-pop^n 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**

*Specifically, $(Z_{1,3} + Z_{1,1})/\sqrt{(2+\sqrt{2})}$, which is $N(0,1)$ under H_{13}.

Enrichment: Example

The rules for staying with the full population or switching to the sub-population can be adjusted, but we cannot eliminate the probability of making an error in these decisions.

This is to be expected. The standard error of interim estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ is 12.3 — much higher than the differences between θ_1 and θ_2 that interest us.

We conclude that

Restricting attention to a sub-population can be effective in improving power.

However, higher overall sample size is needed for accurate sub-population inference.

Increasing power for finding a sub-population effect

To match the non-adaptive test in cases 2 and 3, and obtain the benefits of adaptation elsewhere, increase the overall sample size by 30%.

With a combination* of $Z_{1,3}$ and $Z_{1,1}$ as the Stage 1 statistic for testing H_{13}, we obtain the following results:

				Non-adaptive	***Adaptive, 1.3 x sample size***		
	θ_1	θ_2	θ_3	*Full pop^n*	*Sub-pop^n only*	*Full pop^n*	*Total*
1.	30	0	15	**0.68**	0.49	0.45	**0.94**
2.	20	0	10	**0.37**	0.38	0.30	**0.69**
3.	20	20	20	**0.90**	0.03	0.92	**0.94**
4.	20	10	15	**0.68**	0.15	0.68	**0.82**

*Using $(Z_{1,3} + Z_{1,1})/\sqrt{(2+\sqrt{2})}$.

2. Estimation following an enrichment design

Consider estimating θ_1, θ_2 and θ_3, after a trial conducted according to the enrichment design we have just seen.

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.

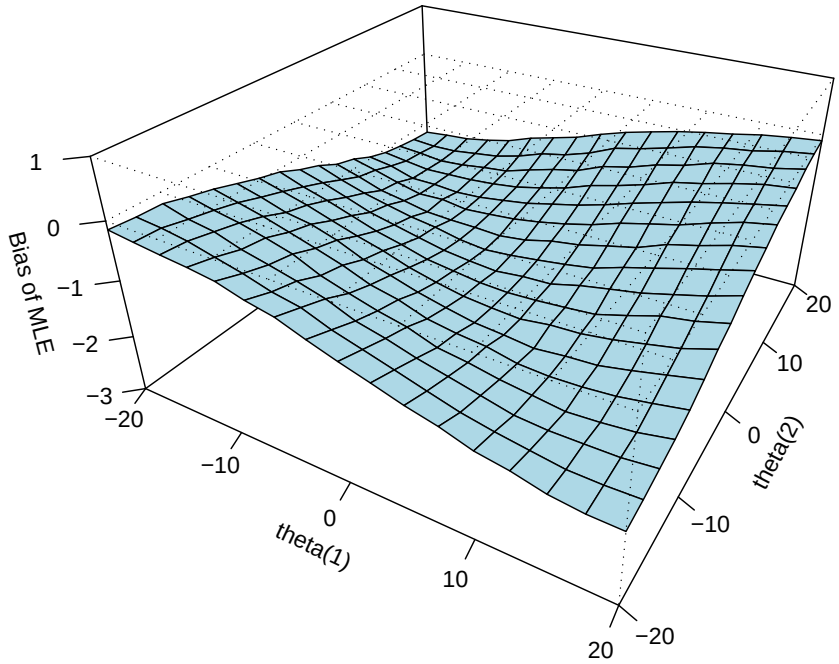
Thus, we are considering *unconditional* estimation.

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 $\hat{\theta}_{3,M}$.

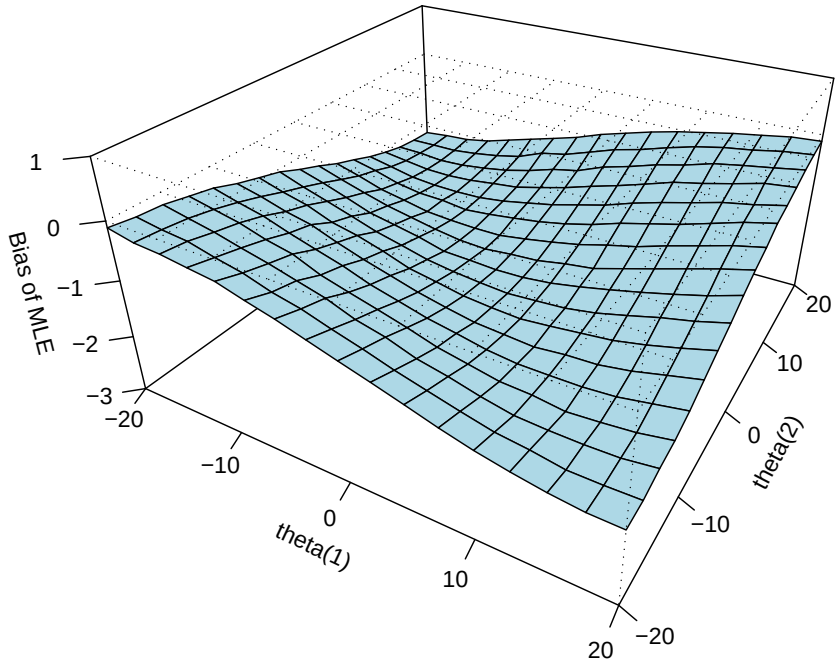
Estimation after an enrichment design



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 method (*Biometrika*, 1986) for estimating a single treatment effect after a sequential trial can be applied in more than one dimension.

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

Let $\hat{\theta}_M = (\hat{\theta}_{1,M}, \hat{\theta}_{2,M}, \hat{\theta}_{3,M})$.

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.)

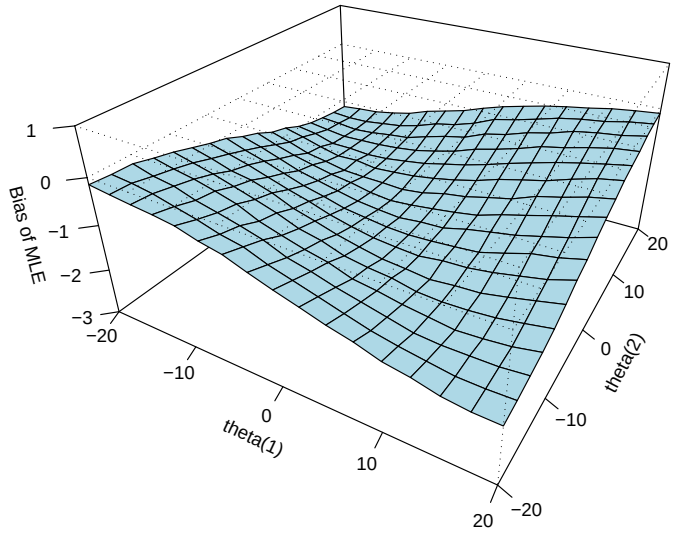
We can estimate the functions $b_1(\theta)$ and $b_3(\theta)$ by simulation.

Hence, we obtain adjusted estimators:

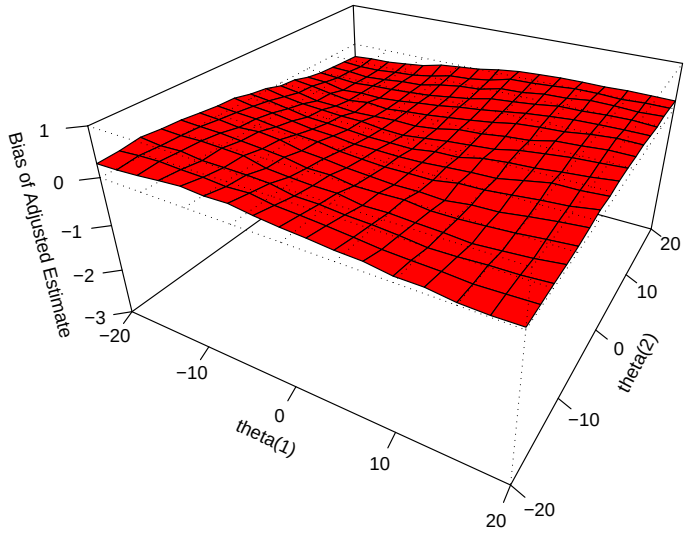
$$\hat{\theta}_{1,adj} = \hat{\theta}_{1,M} - b_1(\hat{\theta}_M) \quad \text{and} \quad \hat{\theta}_{3,adj} = \hat{\theta}_{3,M} - b_3(\hat{\theta}_M).$$

Estimation after an enrichment design

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



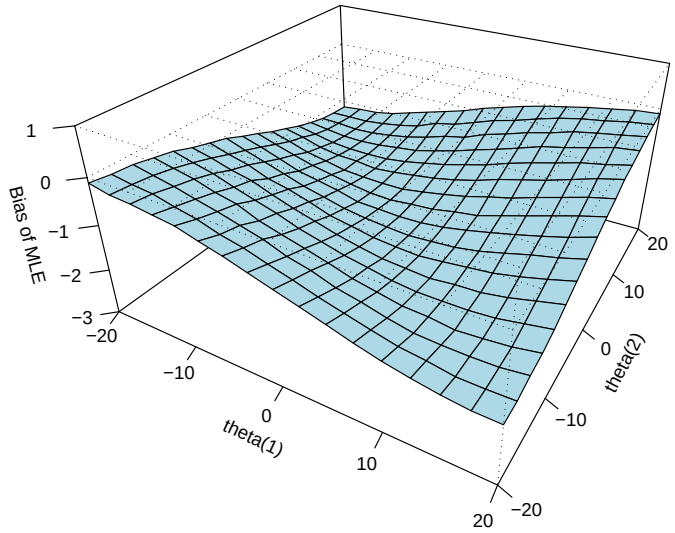
Bias of $\hat{\theta}_{1,M}$, MLE for the sub-population effect θ_1



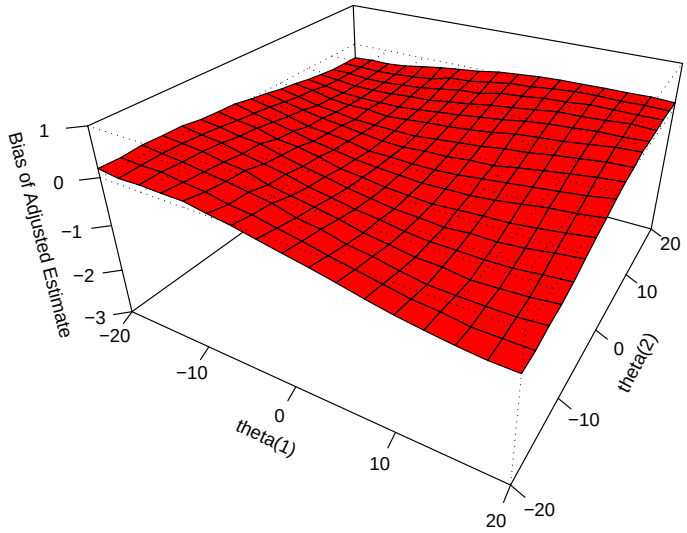
Bias of $\hat{\theta}_{1,adj}$, adjusted estimator of θ_1

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

3. Confidence intervals following an enrichment design

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

Thus, we test:

For the full population, H_3: $\theta_3 \leq 0$ vs $\theta_3 > 0$,

For the sub-population, H_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” achieves this.

Enrichment design: Confidence intervals on termination

Suppose 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_1: $\theta_1 \leq 0$ is rejected, then $\psi_1 \geq 0$

If H_3: $\theta_3 \leq 0$ is rejected, then $\psi_3 \geq 0$

— but preferably $\psi_1 > 0$ and $\psi_3 > 0$.

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.

We shall start from a proposal of Hayter & Hsu (*JASA*, 1994) for constructing a joint CI consistent with a stepwise decision procedure in a fixed sample size design.

Enrichment design: Confidence intervals on termination

Consider a closed testing procedure based on final P-values

P_1 for testing H_1: $\theta_1 \leq 0$,

P_3 for testing H_3: $\theta_3 \leq 0$,

and, with the Bonferroni adjustment,

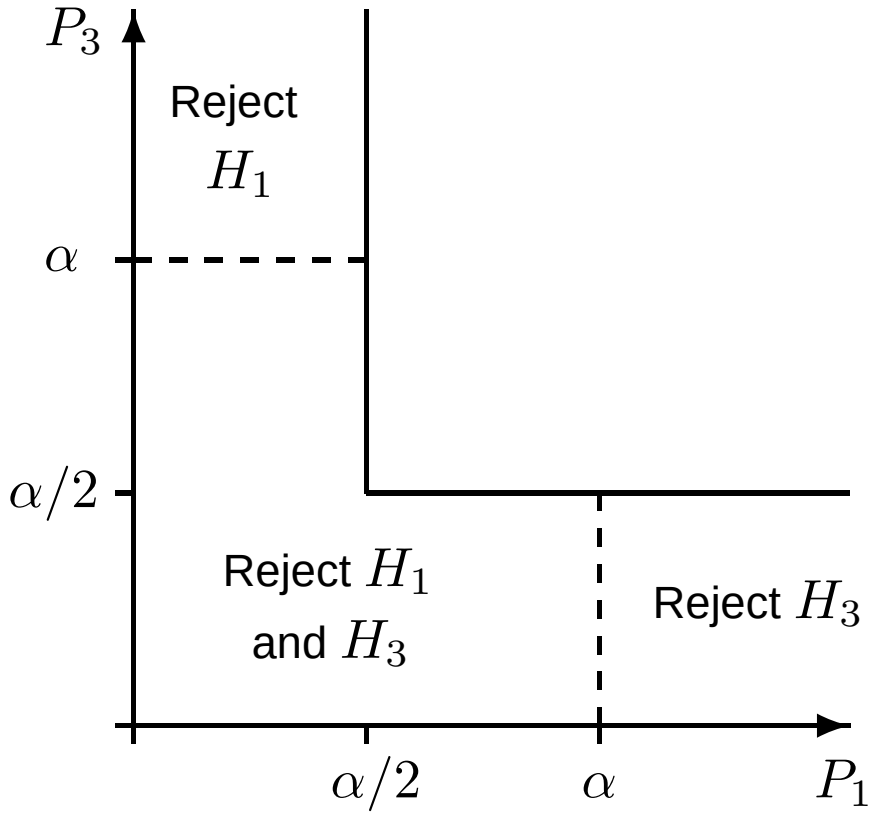
$P_{13} = 2 \min(P_1, P_3)$ for testing H_{13}: $\theta_1 \leq 0$ and $\theta_3 \leq 0$.

In an enrichment design these P-values can be formed by applying a combination test to data from the two stages, before and after adaptation.

If enrichment occurs, there is no final P_3 but we can simply set this equal to 1 in the definition of P_{13}.

Enrichment design: Confidence intervals on termination

Rejection regions for the closed testing procedure defined using P_1, P_3 and P_{13} as defined above are:



Enrichment design: Confidence intervals on termination

Writing $\theta = (\theta_1, \theta_3)$, a CI for θ can be formed from a family of tests of

$$H(\theta^*)\text{: } \theta = \theta^* \text{ vs } \theta_1 \geq \theta_1^* \text{ or } \theta_3 \geq \theta_3^*.$$

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

In our example, define

$$P_1(\theta_1^*) = \text{ the final P-value for testing } H\text{: } \theta_1 = \theta_1^* \text{ vs } \theta_1 \geq \theta_1^*,$$

$$P_3(\theta_3^*) = \text{ the final P-value for testing } H\text{: } \theta_3 = \theta_3^* \text{ vs } \theta_3 \geq \theta_3^*.$$

These P-values can be formed using a combination test, e.g., a weighted inverse normal test for data from the two stages.

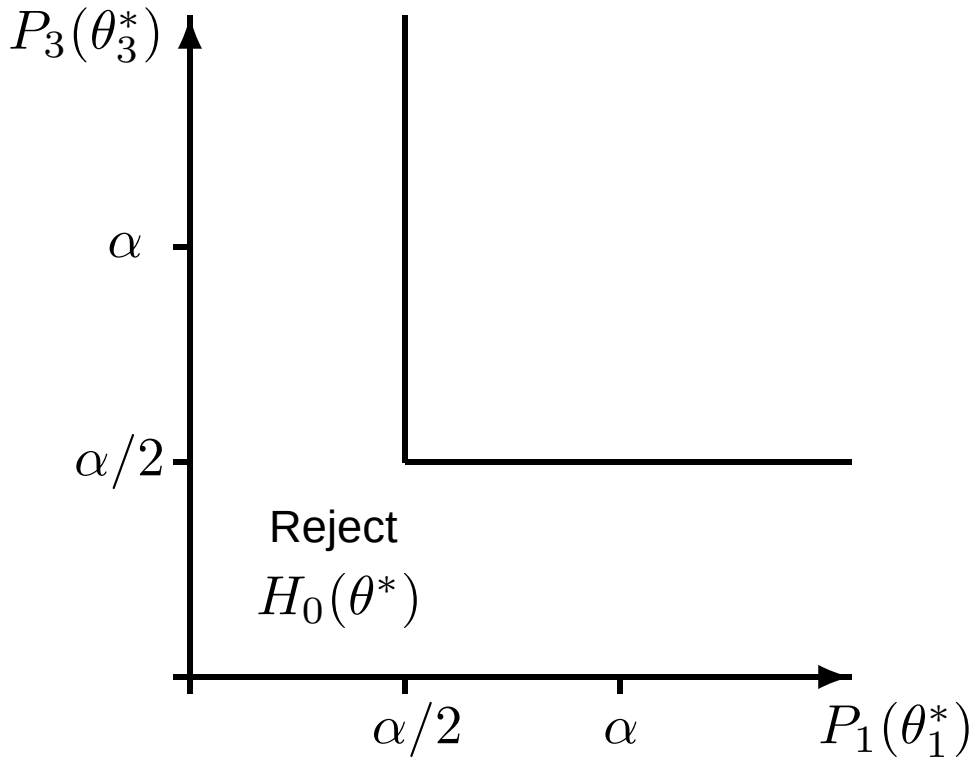
The test of $H(\theta^*)\text{: } \theta = \theta^*$ will be based on $P_1(\theta_1^*)$ and $P_3(\theta_3^*)$.

Testing $H_0(\theta^*)$: $\theta = \theta^*$

Case 1: $\theta_1^* \leq 0$ and $\theta_3^* \leq 0$

Case 2: $\theta_1^* > 0$ and $\theta_3^* > 0$

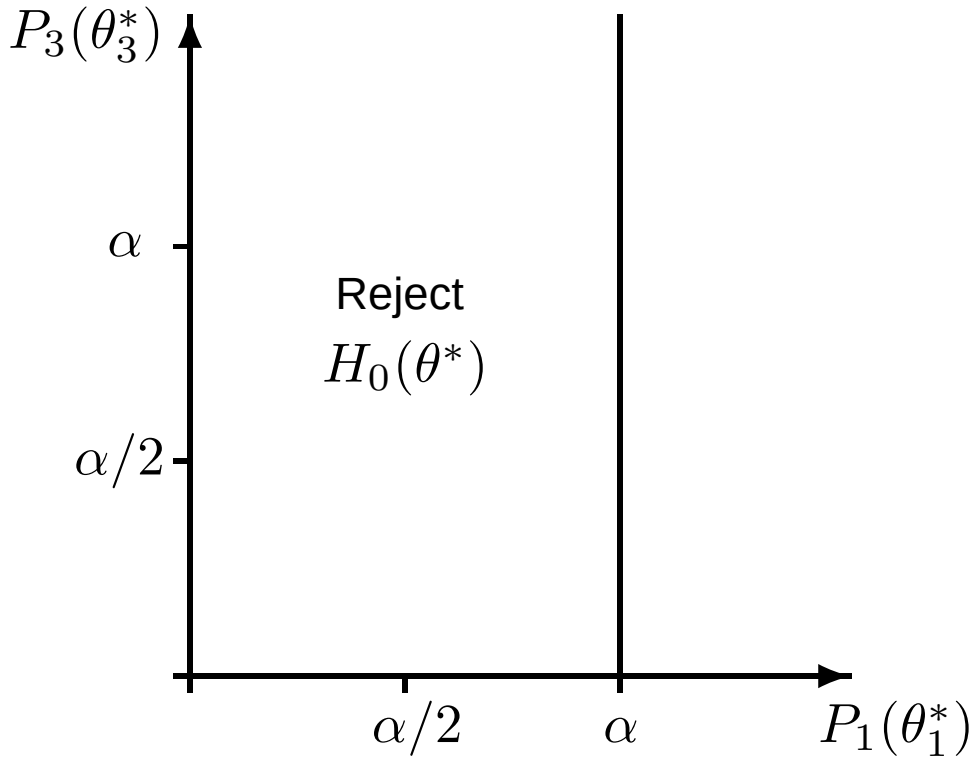
Reject $H_0(\theta^*)$ if at least one of $P_1(\theta_1^*)$ and $P_3(\theta_3^*)$ is less than $\alpha/2$.



Testing $H_0(\theta^*)$: $\theta = \theta^*$

Case 3: $\theta_1^* \leq 0$ and $\theta_3^* > 0$

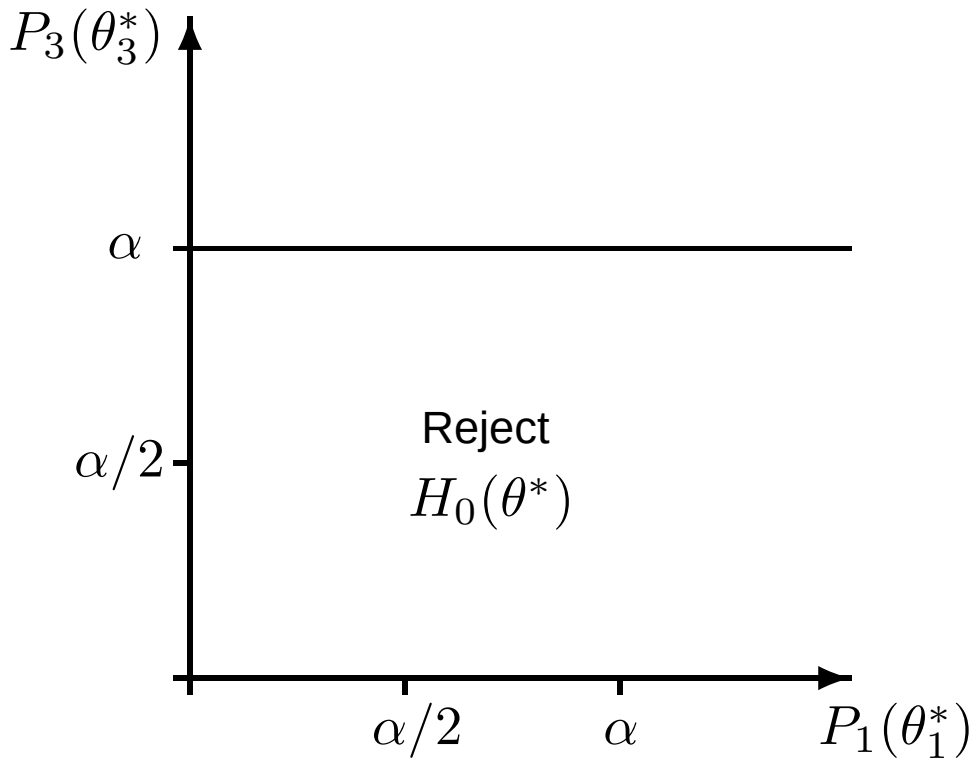
Reject $H_0(\theta^*)$ if $P_1(\theta_1^*) < \alpha$.



Testing $H_0(\theta^*)$: $\theta = \theta^*$

Case 4: $\theta_1^* > 0$ and $\theta_3^* \leq 0$

Reject $H_0(\theta^*)$ if $P_3(\theta_3^*) < \alpha$.



Enrichment design: Confidence intervals on termination

It is straightforward to check consistency of these CIs with the original closed testing procedure.

Suppose, for example, that H_1 is rejected by the overall test.

Then, $H(\theta^*)$ is rejected for all values θ^* with $\theta_1^* \leq 0$,

hence, the upper CI (ψ_1, ∞) for θ_1 has $\psi_1 \geq 0$.

Thus,

the CI for θ_1 excludes $\theta_1 = 0$ when H_1 is rejected and

the CI for θ_3 excludes $\theta_3 = 0$ when H_3 is rejected.

However, we cannot rule out the possibility that, say, H_1 is rejected but the CI for θ_1 is the open interval $(0, \infty)$.

For refinements to avoid this problem, see Strassberger & Bretz (*Statistics in Medicine*, 2008) and Guilbaud (*Biometrical Journal*, 2008).

Enrichment design: Confidence intervals on termination

Note that the closed testing procedure assumed here uses a different form of P_{13} from that introduced in our first form of enrichment design.

Thus, it is important to think ahead to questions of inference on termination when setting up an adaptive design.

One should also check the chosen form of adaptive design has satisfactory power.

Apart from the definition of P_{13}, there are no particular problems in taking the CI construction from a fixed sample problem to a sequential, adaptive trial design.

There is freedom in the definition of P_{13}

(i) to allow for the correlation between $\hat{\theta}_1$ and $\hat{\theta}_3$ and avoid conservatism in the type I error rate

(ii) to focus on higher power for rejecting H_1 or for rejecting H_3, as preferred.

Conclusions

Enrichment designs can deliver on the promise of adapting to a sub-population, when this is appropriate, based on observed data.

Since interim data are noisy, benefits arise from taking the right decision *some of the time*. For a reliable choice between full population and sub-population, sample sizes must be high.

Methods to adjust MLEs to obtain approximately unbiased point estimates extend to adaptive designs with multiple parameters.

Construction of joint confidence intervals for two or more parameters after an adaptive design is problematic but solutions are available.