Estimation and P-values for two-stage Adaptive Designs

Werner Brannath

VO ”Sequential and Adaptive Designs”,
University Bremen

Adaptive two stage combination tests

Notation: p and q the p-values from stage 1 and 2 for

$$H_0 : \theta \leq 0 \quad \text{versus} \quad H_1 : \theta > 0$$

p and q are independent under H_0.

Two stage combination test: Prefix a monotone combination function $C(p, q)$ and rejection bounds c and α_1.

$$\begin{array}{lll} \text{We reject } H_0 \text{ if} & \text{either} & p \leq \alpha_1 \quad \text{(stage 1)} \\ & \text{or} & C(p, q) \leq c \quad \text{(stage 2)} \end{array}$$

Level condition: We must prefix α_1, $C(p, q)$ and c such that

$$P_0(\{p \leq \alpha_1\} \cup \{C(p, q) \leq c\}) = \alpha$$

Examples for combination functions

- Fisher's product test:

$$C(p, q) = p \cdot q$$

- Inverse normal method:

$$C(p, q) = \Phi\left(w_1\, \Phi^{-1}(p) + w_2\, \Phi^{-1}(q)\right), \quad w_1^2 + w_2^2 = 1$$

Corresponds to two stage GSD with information times $t_1 \leq 1$ if $w_1 = \sqrt{t_1}$ and no adaptations are done.

Problem

- The usual confidence intervals do not provide the correct coverage probability. The non-coverage probability may be substantially larger than α (like the type I error rate of the naive test).
- The usual p-values may be anti-conservative.
- The maximum likelihood estimates may be severely biased.

Repeated Confidence Intervals

(WASSMER & LEHMACHER, 1997; LEHMACHER & WASSMER, 1999; BRANNATH ET AL. 2002, LAWRENCE & HUNG, 2003; PROSCHAN ET AL., 2003)

Repeated confidence intervals

Duality between hypothesis tests and confidence sets:

p_Δ stage 1 and q_Δ stage 2 p-values for $H_{0,\Delta} : \theta \leq \Delta$.

p_Δ and q_Δ p-clud under $H_{0,\Delta}$ and increasing in Δ.

Apply two stage combination test to all $H_{0,\Delta}$:

We reject $H_{0,\Delta}$ if either $p_\Delta \leq \alpha_1$ (stage 1)

or $C(p_\Delta, q_\Delta) \leq c$ (stage 2)

Remark: The rule "$p_\Delta \leq \alpha_1$" should *not* be understood as a *stopping rule*, but as *rejection rule* which we apply at stage 1.

Lower repeated confidence bounds

Stage 1: Solve the equation $p_\Delta = \alpha_1 \quad \rightarrow \quad \delta_1$ such that

$$p_\Delta \leq \alpha_1 \quad \Longleftrightarrow \quad \Delta \leq \delta_1$$

$\rightarrow$ (δ_1, ∞) one-sided confidence interval at first stage.

Stage 2: Solve $C(p_\Delta, q_\Delta) = c \quad \rightarrow \quad \delta_2$ such that

$$C(p_\Delta, q_\Delta) \leq c \quad \Longleftrightarrow \quad \Delta \leq \delta_2$$

$\rightarrow$ (δ_2, ∞) one-sided confidence interval at second stage.

Lower repeated confidence bounds

Denote $L \in \{1, 2\}$ the random stage at which recruitment is stopped. (The symbol L stands for *last* stage).

Let $\delta_L = \delta_1$ if $L = 1$ and $\delta_L = \delta_2$ if $L = 2$.

Theorem: (δ_L, ∞) has coverage probability $1 - \alpha$ independently from the stopping rule L and the adaptations.

Remark: For the one-sided CI we can even use the larger bound $\delta_2' = \max(\delta_1, \delta_2)$ instead of δ_2 at the second stage.

Lower repeated confidence bounds

Proof of the Theorem:

$$\begin{aligned} \mathbf{P}_\Delta(\delta_L > \Delta) &\leq P_\Delta(\{\delta_1 > \Delta\} \cup \{\delta_2 > \Delta\}) \\ &= \mathbf{P}_\Delta(\{p_\Delta \leq \alpha_1\} \cup \{C(p_\Delta, q_\Delta) \leq c\}) = \alpha \end{aligned}$$

where

- in the second probability statement we assume that the trial is always continued until the second stage,
- the first equality follows from the dual combination test,
- the last equality follows from the level condition of the combination test.

Example I

Primary efficacy endpoint: Infarct size measured by the cumulative release of α-HDBH within 72 hours after administration of the drug (area under the curve, AUC).

θ the mean α-HDBH AUC difference between control c and treatment t, $H_0 : \theta \leq 0$ vs. $H_1 : \theta > 0$

Inverse normal combination test:

$$C(p, q) = \Phi\left(\sqrt{0.5} \cdot \Phi^{-1}(p) + \sqrt{0.5} \cdot \Phi^{-1}(q)\right)$$

O'Brien & Fleming at one sided level $\alpha = 0.025$

$\rightarrow \alpha_1 = 0.0026$, $c = 0.024$.

Example I (cont.)

Stage 1: sample sizes: $n_{1c} = 88$, $n_{1t} = 91$,
standard deviation: $\hat{\sigma}_{1c} = 26.0$, $\hat{\sigma}_{1t} = 22.5$
treatment difference: $\hat{\theta}_1 = 4.0$, $\sigma_{\hat{\theta}_1} = 3.64$

p_Δ according to t-test for $H_0 : \theta = \Delta$.

Solving $p_\Delta = 0.0026$ $\longrightarrow$ classical CI at level 0.0026

$$\delta_1 = \hat{\theta}_1 - t_{\nu,0.9974} \cdot \sigma_{\hat{\theta}_1} = -6.3$$

First stage confidence interval is $(-6.3, \infty)$

Example I (cont.)

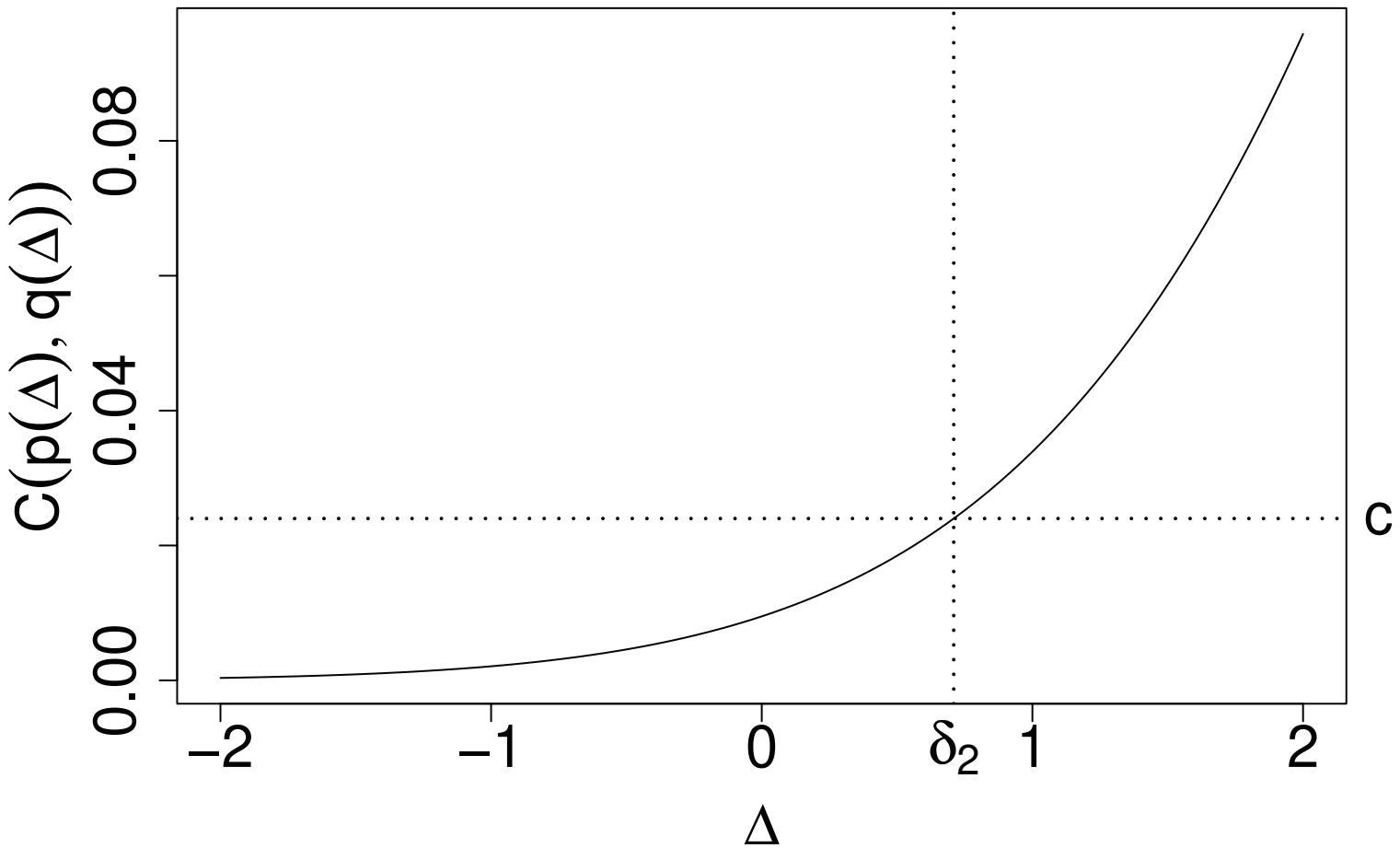
Stage 2: sample sizes $n_{2c} = 322$, $n_{2t} = 321$,
standard deviations: $\hat{\sigma}_{2c} = 26.1$, $\hat{\sigma}_{2t} = 28.5$,
treatment difference: $\hat{\theta}_2 = 4.8$, $\sigma_{\hat{\theta}_2} = 2.16$

q_Δ according to t-test for $H_0 : \theta = \Delta$ from second stage data.

Solving $C(p_\Delta, q_\Delta) = 0.024$ numerically $\longrightarrow$ $\delta_2 = 0.71$

Second stage confidence interval is $(0.71, \infty)$

Example I (cont.): Determination of δ_2



Properties of repeated confidence bounds

- One need not pre-specify the adaptation and stopping rule to keep the nominal coverage probability.
- Price for the flexibility with regard to stopping rule is strict conservatism: we must control the level for the worst case rule, also when actually not following this rule.
- H_0 is rejected with the combination test iff $\delta_L > 0$.
- The first stage bound δ_1 is the classical confidence bound at level α_1.

Normal approximations and inverse normal method (Lehmacher and Wassmer, 1999)

If the stage wise estimates $\hat{\theta}_i$ $(i = 1, 2)$ for the treatment effect are (approximately) independent and normal with mean treatment effect Δ and variance $\sigma^2_{\hat{\theta}_i} = I_1^{-1}$, then

$$p(\Delta) = 1 - \Phi(\sqrt{I_1} \cdot (\hat{\theta}_1 - \Delta)) \quad \text{and} \quad q(\Delta) = 1 - \Phi(\sqrt{I_2} \cdot (\hat{\theta}_2 - \Delta))$$

are (approximately) independent p-values for $H_{0,\Delta} : \theta \leq \Delta$.

With inverse normal combination function:

$$\delta_2 = \widehat{\theta}_w - \frac{\Phi^{-1}(1-c)}{w_1 \cdot \sqrt{I_1} + w_2 \cdot \sqrt{I_2}}, \quad \widehat{\theta}_w = \frac{w_1 \cdot \sqrt{I_1} \cdot \hat{\theta}_1 + w_2 \cdot \sqrt{I_2} \cdot \hat{\theta}_2}{w_1 \cdot \sqrt{I_1} + w_2 \cdot \sqrt{I_2}}$$

Example I with normal approximation

Stage 1: $\hat{\theta}_1 = 4.0, \quad I_1 = 0.076$

$$\delta_1 = 4.0 - \Phi^{-1}(0.9974) \cdot \sqrt{I_1} = -6.2 \quad \text{(before } -6.3\text{)}$$

Stage 2: $\hat{\theta}_2 = 4.8, \quad I_2 = 0.215, \quad w_1 = w_2 = \sqrt{0.5}$

$$\widehat{\theta}_w = \frac{\sqrt{I_1} \cdot \hat{\theta}_1 + \sqrt{I_2} \cdot \hat{\theta}_2}{\sqrt{I_1} + \sqrt{I_2}} = 4.5$$

$$\delta_2 = 4.5 - \frac{\Phi^{-1}(1 - 0.024)}{(\sqrt{I_1} + \sqrt{I_2})\sqrt{0.5}} = 0.70 \quad \text{(before 0.71)}$$

Extensions

- Repeated confidence intervals can be extended to multistage adaptive designs, and can be computed even after adapting the number of interim looks
  (LEHMACHER AND WASSMER, '99; BRANNATH ET AL., '02; MEHTA ET AL., '07)
- One can incorporate a futility boundary into the dual combination tests. However, one must carefully account for the futility bound in the determination of δ_2:

  One must accept all Δ for which stage 1 p-value p_Δ falls into stage 1 acceptance region even if the second stage data suggest rejection of $H_{0,\Delta}$.
- We can use different α_1 and c for different Δ, however, to get nested dual rejection regions one must be careful in the choice of $\alpha_1(\Delta)$ and $c(\Delta)$.

Two-sided tests and confidence intervals

- One should not perform combination tests with two-sided p-values for $H_{0,\Delta} : \theta = \Delta$:

  Interpretation problem if the first and the second stage estimates point in conflictive directions.

- Better use the two one-sided combination tests at level $\alpha/2$.

- We can use the intersection of the corresponding repeated confidence intervals (as lower and upper confidence bound)

  $\rightarrow$ $(1-\alpha)100\%$ two-sided confidence interval

Two-sided confidence intervals

- At the first stage we get the classical two-sided confidence interval for all combination tests.
- With the normal approximation and normal inverse method we get at the second stage the interval

$$\left(\widehat{\theta}_w - \frac{\Phi^{-1}(1-c)}{w_1 \cdot \sqrt{I_1} + w_2 \cdot \sqrt{I_2}} \, , \, \widehat{\theta}_w + \frac{\Phi^{-1}(1-c)}{w_1 \cdot \sqrt{I_1} + w_2 \cdot \sqrt{I_2}} \right)$$

with

$$\widehat{\theta}_w = \frac{w_1 \cdot \sqrt{I_1} \cdot \hat{\theta}_1 + w_2 \cdot \sqrt{I_2} \cdot \hat{\theta}_2}{w_1 \cdot \sqrt{I_1} + w_2 \cdot \sqrt{I_2}}$$

Confidence intervals for conditional error functions

Conditional error function approach: Prefix a decreasing conditional error function $A(x)$ and first stage rejection level α_1.

Reject H_0 if $p \leq \alpha_1$ (stage 1) or $q \leq A(p)$ (stage 2).

Equivalent combination test (POSCH & BAUER 1999, WASSMER 1999):

e.g. : α_1, $C(p, q) = q - A(p)$, and $c = 0$

$\rightarrow$ One can use the same estimation methods as for combination tests

Overall p-Values

Overall p-Values

- The p-Value is the smallest significance level at which H_0 can be rejected with the given data.
- Calculation of an overall p-value requires the definition and use of an adaptive test for all levels $0 \leq \nu \leq 1$.
- The overall p-value is then the smallest significance level for which H_0 can be rejected with the given data.

- The rejection region of these adaptive tests need to be increasing (nested) in ν,
- For $\nu = \alpha$ we need to get the original adaptive test.

- With more than a single stage (GSD or adaptive design) the definition of the p-value is not unique.

Repeated p-Values

- Repeated p-values have been suggested for group sequential designs.
- The idea has be extended to adaptive designs.
- The main idea is to use the (chosen) family of group sequential boundaries (e.g. O'Brien and Fleming boundaries) not only for α, but for all significance levels $0 \leq \nu \leq 1$.
- Since the resulting local levels $\alpha_{1,\nu}$ and c_ν are increasing in ν, we obtain nested rejection regions at every stage.
- The repeated p-value P_k at stage $k = 1, 2$ is the smallest significance level ν for which we can reject H_0 at the stage k.
- P_2 is calculated only, if we proceed to the second stage ($L = 2$) (by whatever stopping criteria we use).

Repeated p-Values for the Inverse Normal Test (I)

- We consider (as an example) the inverse normal combination function

$$C(p,q) = 1 - \Phi\left(\sqrt{0.5}\,\underbrace{\Phi^{-1}(1-p)}_{z_1} + \sqrt{0.5}\,\underbrace{\Phi^{-1}(1-q)}_{z_2}\right)$$

with $\alpha_0 = 1$ and

$$\alpha_1 = 1 - \Phi\big(c_{WT}(\alpha,\Delta)\big) \quad \text{and} \quad c = 1 - \Phi\big(c_{WT}(\alpha,\Delta)\cdot 2^{\Delta-0.5}\big),$$

according to Wang & Tsiatis with some fixed Δ.

Repeated p-Values for the Inverse Normal Test (II)

- For the repeated p-value we use the rejection boundaries

$$\alpha_{1,\nu} = 1 - \Phi\big(c_{WT}(\nu, \Delta)\big) \quad \text{and} \quad c_\nu = 1 - \Phi\big(c_{WT}(\nu, \Delta) \cdot 2^{\Delta - 0.5}\big),$$

according to the Wang & Tsiatis.

- This means that the repeated p-values are calculated such that they satisfy

$$p \overset{!}{=} \alpha_{1,P_1} = 1 - \Phi\big(c_{WT}(P_1, \Delta)\big)$$

and

$$C(p, q) \overset{!}{=} c_{P_2} = 1 - \Phi\big(c_{WT}(P_2, \Delta) \cdot 2^{\Delta - 0.5}\big)$$

Example 1 (once more I)

- We used O'Brien & Fleming boundaries ($\Delta = 0$) at one-sided level $\alpha = 0.025$.
- The stage-one p-value is $p = 0.136$
- At stage 1 we solve the equation

$$0.495 \stackrel{!}{=} 1 - \Phi\big(c_{OBF}(\nu, \Delta)\big)$$

in ν, which gives: $P_1 = 0.263$

Example 1 (once more II)

- The stage-two p-value is $q = 0.013$ and

$$C(p, q) = 1 - \Phi\left(\sqrt{0.5}\Phi^{-1}(1-p) + \sqrt{0.5}\Phi^{-1}(1-q)\right) = 0.009$$

- At stage 2 we solve the equation

$$C(p, q) = 0.009 \stackrel{!}{=} 1 - \Phi\big(c_{OBF}(\nu, \Delta)/2\big)$$

in ν, which gives: $P_2 = 0.01$

Properties of repeated p-values

- L the stage at which the trial stops, then for all $u \in (0,1)$:

$$\mathbf{P}_{H_0}\left(P_L \leq u\right) \leq P_\Delta\Big(\{P_1 \leq u\} \cup \{P_2 \leq u\}\Big) =$$

$$= \mathbf{P}_{H_0}\Big(\text{adaptive test at level } u \text{ rejects } H_0\Big) = u$$

- The adaptive test rejects if and only if $P_L \leq \alpha$.
- One need not pre-specify the adaptation and stopping rule.
- Price for this flexibility is a strict conservatism, i.e. for most u (or even all, depending on the stopping rule) the above inequality is strict.

P-Values and Sample Space Orderings

- Define a strict ordering on the sample space (i.e. on the space of all possible trial outcomes), and . . .
- . . . calculate under H_0 the probability to observe an outcome that is larger then the one observed in the trial.
- The ordering specifies which outcomes are provide more evidence against H_0 (are more extreme) than others.
- In two- and multi-stage GSD and adaptive designs, the ordering is neither clear nor unique (like in single stage designs, namely by the single test statistics).
- For this reasons there a multiple ways for defining an overall p-value for GSD and adaptive designs.

Stage-wise ordering for combination tests

We order the sample space as follows:

- order according to p_1 at the first stage;
- order according to $C(p_1, p_2)$ at the second stage;
- $p_1 \leq \alpha_1$ is more extreme than any 2^{nd} stage outcome;
- $p_1 > \alpha_0$ is less extreme than any 2^{nd} stage outcome.

Corresponding overall p-value:

$$Q(p_1, p_2) = \begin{cases} p_1 & , p_1 \leq \alpha_1 \text{ or } p_1 > \alpha_0 \\ \alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_0^1 \mathbf{1}_{\{C(x,y) \leq C(p_1,p_2)\}} \, dx \, dy & , \alpha_1 < p_1 \leq \alpha_0 \end{cases}$$

Exact lower confidence bound for combination tests

- One can use the stage-wise ordering also to define an exact overall lower confidence bound.
- To this end we define for each $H_{0,\delta} : \theta \leq \delta$, $\delta \in \mathbb{R}$, a p-value based on (a slightly modified) stage-wise ordering.

  (We use for the stopping rules the p-values p_j and in the combination function the p-values $p_{j,\delta}$ of $H_{0,\delta}$.)
- We collect all δ that are accepted by their p-value.
- This leads to a one-sided interval with a finite lower bound ...
- ... that has coverage probability equal to $1 - \alpha$ (when the stage wise p-values are ind. and uniformly distributed).

Point Estimation

Maximum likelihood estimate (MLE)

Assuming normal data and balanced treatment groups the MLE can be written as

$$\hat{\theta}_{mle} = \frac{I_1}{I_1 + I_2} \cdot \hat{\theta}_1 + \frac{I_2}{I_1 + I_2} \cdot \hat{\theta}_2$$

(for small effect sizes approximatively also in other cases)

Mean Bias: $E_\Delta(\hat{\theta}_{mle} - \Delta) = Cov_\Delta(\frac{I_1}{I_1+I_2}, \hat{\theta}_1)$ (Liu et al. 2002)

One can show that always: $|E_\Delta(\hat{\theta}_{mle} - \Delta)| \leq 0.4 \cdot \sigma/\sqrt{n_1}$

Variance also depends on (unknown) adaptation/selection rule

Maximum likelihood estimate (MLE)

Mean bias of MLE for typical examples (qualitatively):

- *Stopping with early rejection:* the larger the effect size the smaller the sample size $\rightarrow$ positive mean bias.
- *Stopping for futility:* the smaller the effect size the smaller the sample size $\rightarrow$ negative mean bias.
- *Conditional or predictive power control:* the smaller the effect size the larger the sample size $\rightarrow$ positive mean bias.
- *Selecting promising treatments:* the larger the effect size the larger the sample size $\rightarrow$ negative mean bias.

Weighted maximum likelihood estimate

(LAWRENCE & HUNG, 2003; PROSCHAN ET AL., 2003; BRANNATH ET AL., 2002)

Center of a two sided repeated confidence interval:

$$\widehat{\theta}_w = \frac{w_1 \cdot \sqrt{I_1} \cdot \hat{\theta}_1 + w_2 \cdot \sqrt{I_2} \cdot \hat{\theta}_2}{w_1 \cdot \sqrt{I_1} + w_2 \cdot \sqrt{I_2}}$$

where $w_1, w_2 \geq 0$, $w_1^2 + w_2^2 = 1$ are the pre-specified weights.

Properties:

- If recruitment is stopped at stage 1 then $\widehat{\theta}_w = \hat{\theta}_1$.
- If recruitment is never stopped at the interim analysis, then $\hat{\theta}_w$ is *median unbiased*, i.e., $\hat{\theta}_w$ has median Δ.

Cases for which the estimates are similar

The two estimates are equal or differ only slightly if

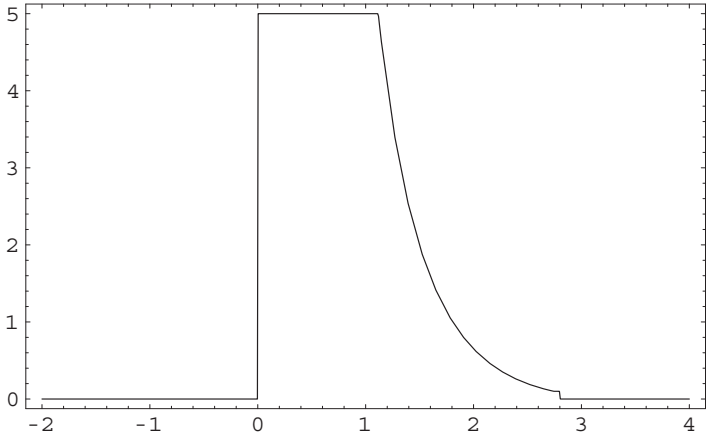
- recruitment is stopped at the interim analysis;
- recruitment is **not** stopped at the interim analysis, and
  - the first and second stage estimates are similar, $\hat{\theta}_1 \approx \hat{\theta}_2$;

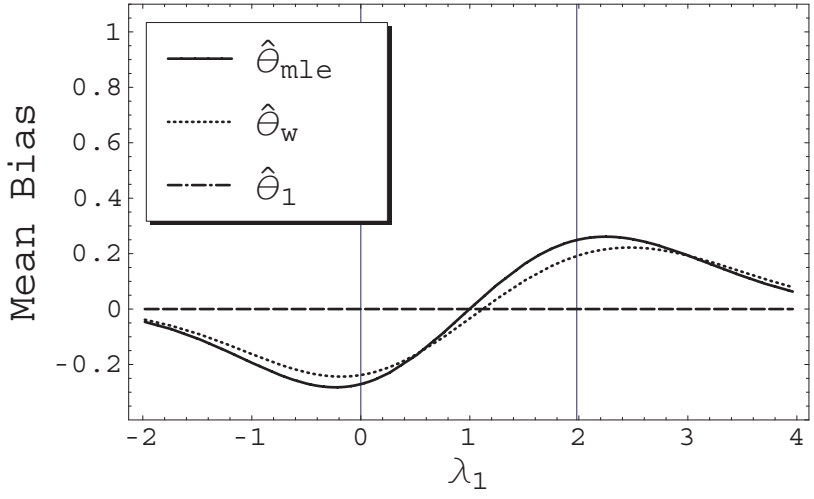
  or
  - the sample sizes are (almost) as pre-planned: $\sqrt{I_1/I_2} \approx w_1/w_2 = \sqrt{t_1/(t_2 - t_1)}$

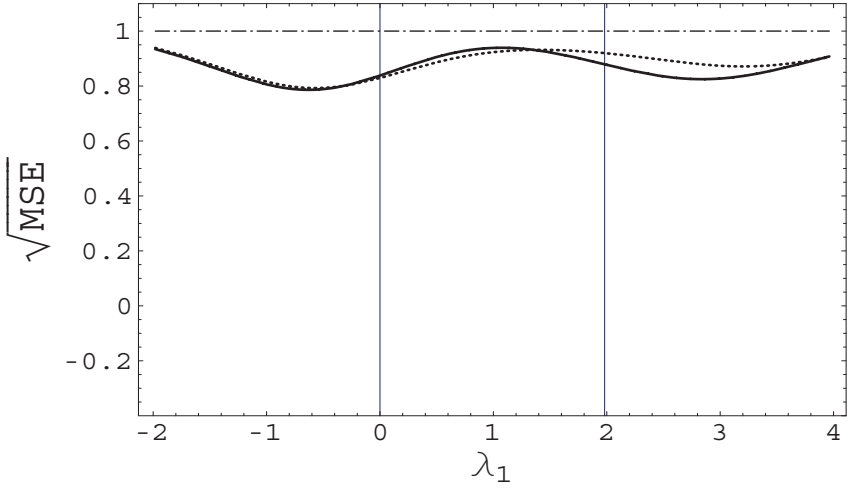
Numerical example

80% - Predictive power rule,
truncated $0.1 \cdot I_1 \leq I_2 \leq 5 \cdot I_1$.

$\widehat{\theta}_w$ with $w_1^2 = 0.5$
$\hat{\theta}_1$ first stage mean diffenrence
horizontal axis: $\lambda_1 = \sqrt{I_1} \cdot \theta$







Summary

- Univariate confidence intervals and p-values are, in general, available for adaptive adaptive designs.
- Repeated confidence intervals and p-values provide flexibility with regard to the stopping rule but are conservative.
- Using the normal approximation of stage wise estimates and the inverse normal combination function, we get explicit (and intuitive) formula for the confidence bounds.
- Maximum likelihood estimate is biased, however, seems to perform well in terms of the mean square error.
- The *weighted maximum likelihood estimate* is, in general, less biased (and median unbiased in the case of an administrative interim look).
- With a stopping rule a median unbiased estimate can be obtained via the stage wise ordering.