Ordering the sample space for p-value and confidence interval computation in group sequential trials with delayed outcome

Brice Ozenne [1,2], Paul Blanche[1], and Corine Baayen [3,4],

[1] Section of Biostatistics, Department of Public Health, University of Copenhagen.

[2] Neurobiology Research Unit, University Hospital of Copenhagen, Rigshospitalet.

[3] Biometric Division, H. Lundbeck A/S.

[4] Global Biometrics, Ferring Pharmaceuticals.

25th April 2024 - ADMTP Workshop

Group sequential design (GSD)

Statistical framework to perform repeated significance testing:

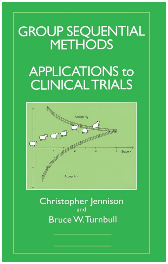
- early stopping for efficacy or futility

Uses sets of boundaries to control type I and type II error

- boundaries evaluated once the data is collected

Key result: test statistics are asymptotically multivariate normal

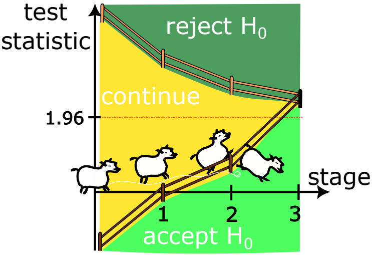
- integration of the multivariate Gaussian density



GROUP SEQUENTIAL METHODS

APPLICATIONS to CLINICAL TRIALS

Christopher Jennison and Bruce W. Turnbull



An example of trial

Endpoint: clinical score measured at week 0, 6, 12 $Y(t)$

Exposure: drug vs. placebo E

Estimand: $\mathbb{E}\left[Y(12)|E=1\right] - \mathbb{E}\left[Y(12)|E=0\right]$ θ

Design: 2 interim analyses and a final analysis $k \in \{1,2,3\}$

- maximum planned information $\mathcal{I}_{\max}$
  sample size at each stage n_k
- . . .

Model: mixed model each stage

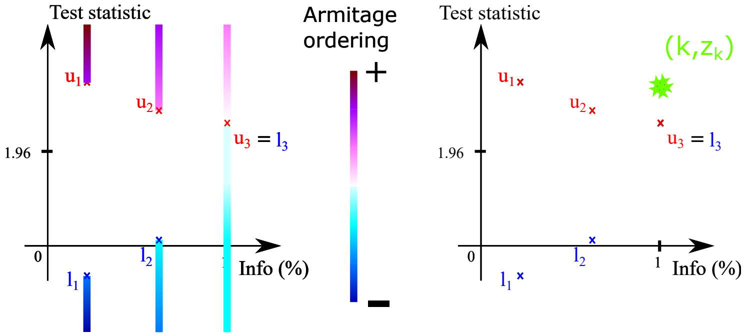
- estimate and test statistic $\widehat{\theta}_k,\ Z_k$
- current information $\sigma^{-2}_{\widehat{\theta}_k}$ $\mathcal{I}_k$
- current information fraction $\mathcal{I}_k/\mathcal{I}_{\max}$ Info (%)

Estimating p-values

Suppose we stop at stage k with a test statistic z_k.

P-value: Probability to obtain a result that is at least as extreme as the observed result under the null:

$$p = \mathbb{P}_0\left[(\kappa, Z_\kappa) \succeq (k, z_k)\right] \text{ where } \kappa = \min(k : Z_k \notin [l_k, u_k])$$

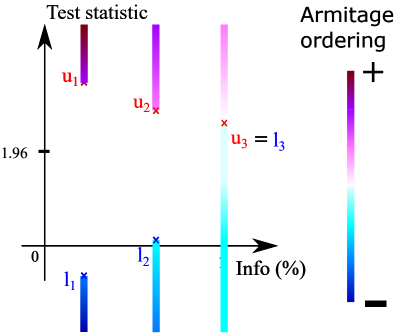


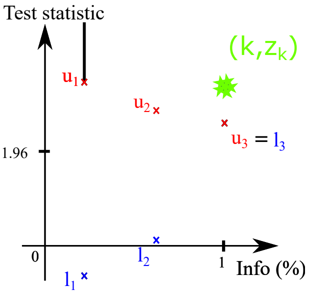
Estimating p-values

Suppose we stop at stage k with a test statistic z_k.

P-value: Probability to obtain a result that is at least as extreme as the observed result under the null:

$$p = \mathbb{P}_0\left[(\kappa, Z_\kappa) \succeq (k, z_k)\right] \text{ where } \kappa = \min(k : Z_k \notin [l_k, u_k])$$





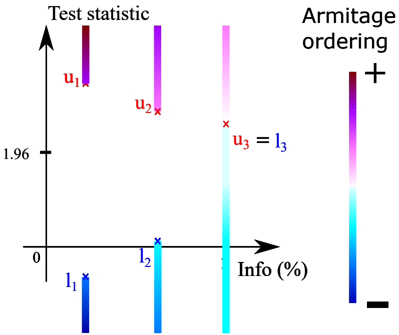
P-value calculation

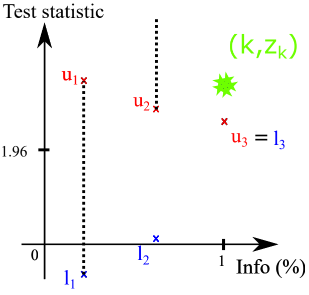
Estimating p-values

Suppose we stop at stage k with a test statistic z_k.

P-value: Probability to obtain a result that is at least as extreme as the observed result under the null:

$$p = \mathbb{P}_0\left[(\kappa, Z_\kappa) \succeq (k, z_k)\right] \text{ where } \kappa = \min(k : Z_k \notin [l_k, u_k])$$





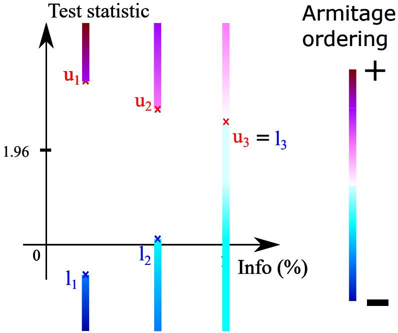
P-value calculation

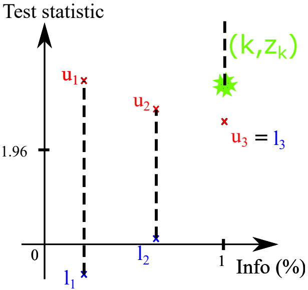
Estimating p-values

Suppose we stop at stage k with a test statistic z_k.

P-value: Probability to obtain a result that is at least as extreme as the observed result under the null:

$$p = \mathbb{P}_0\left[(\kappa, Z_\kappa) \succeq (k, z_k)\right] \text{ where } \kappa = \min(k : Z_k \notin [l_k, u_k])$$





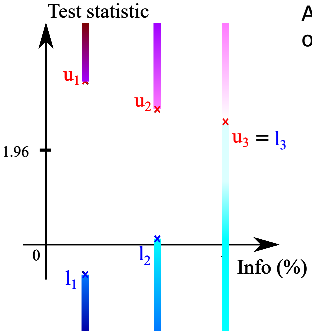
P-value calculation

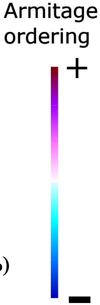
Estimating p-values

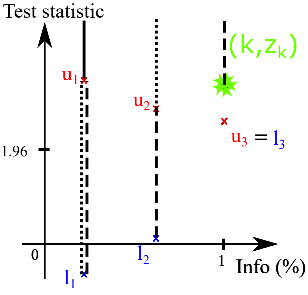
Suppose we stop at stage k with a test statistic z_k.

P-value: Probability to obtain a result that is at least as extreme as the observed result under the null:

$$p = \mathbb{P}_0\left[(\kappa, Z_\kappa) \succeq (k, z_k)\right] \text{ where } \kappa = \min(k : Z_k \notin [l_k, u_k])$$







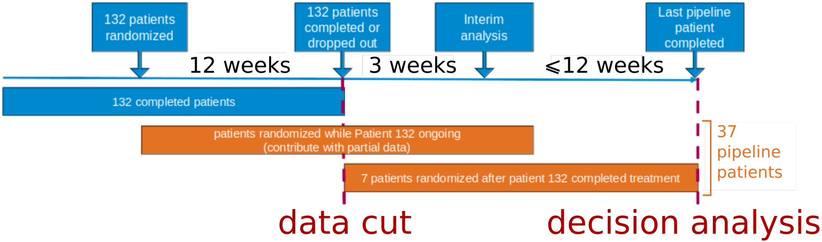
P-value calculation

Need for extending GSD methodology

GSD is a well established methodology
. . . for outcomes measured immediately after treatment.

In practice, delayed outcome are common

- leading to incomplete data at interim



132 patients randomized

132 patients completed or dropped out

Interim analysis

Last pipeline patient completed

12 weeks

3 weeks

⩽12 weeks

132 completed patients

patients randomized while Patient 132 ongoing (contribute with partial data)

7 patients randomized after patient 132 completed treatment

37 pipeline patients

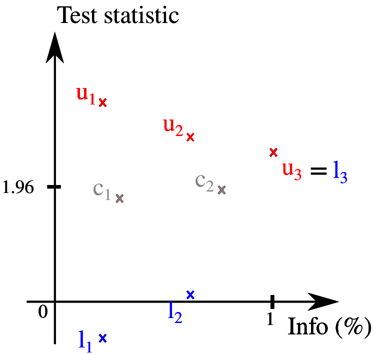
data cut

decision analysis

Adding a decision analysis

At the k-th interim:
u_k efficacy boundary
l_k futility boundary
$\mathcal{I}_k$ information
Z_k test statistic

At the k-th decision:
c_k decision boundary
$\widetilde{\mathcal{I}}_k$ information
$\widetilde{Z}_k$ test statistic



Hampson and Jennison (2013); Jennison (2022) define the boundaries to control the type 1 and type 2 error.

- Method 1-3: Corine Baayen's talk (Friday morning)

Adding a decision analysis

At the k-th interim:

u_k efficacy boundary

l_k futility boundary

$\mathcal{I}_k$ information

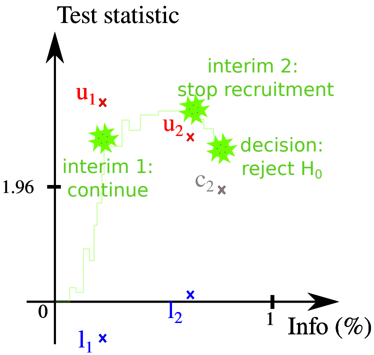
Z_k test statistic

At the k-th decision:

c_k decision boundary

$\widetilde{\mathcal{I}}_k$ information

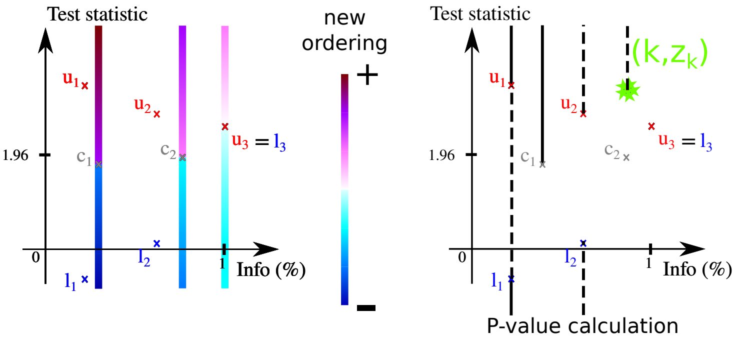
$\widetilde{Z}_k$ test statistic



Hampson and Jennison (2013); Jennison (2022) define the boundaries to control the type 1 and type 2 error.

- Method 1-3: Corine Baayen's talk (Friday morning)

What is more extreme? (Hampson and Jennison, 2013)



$$p = \mathbb{P}_0\left[\left(\kappa, \widetilde{Z}_\kappa\right) \succeq \left(k, \widetilde{z}_k\right)\right]$$

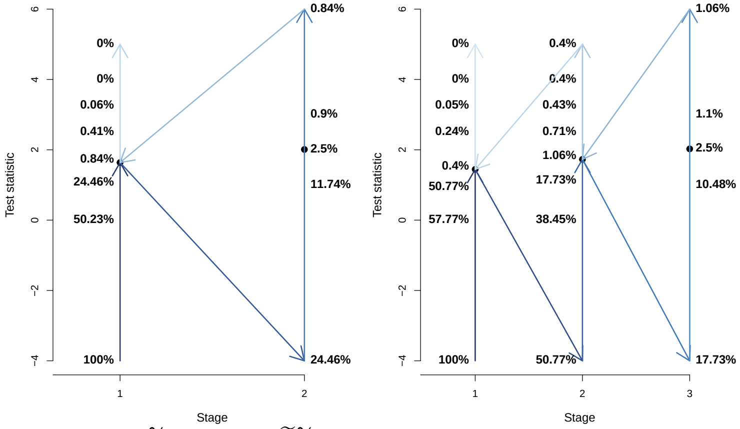
✔ p-values agree with boundaries and spent type 1 error

✔ continuous p-values taking all values between 0 and 1

✔ only depends on current and past information

- does not depend on test statistic(s) at interim

P-value space (Method 1, binding)



- 2 stage: $\mathcal{I}_1^{\%} = 58\%$, $\widetilde{\mathcal{I}}_1^{\%} = 68\%$
- 3 stages: $\mathcal{I}_1^{\%} = 40\%$, $\widetilde{\mathcal{I}}_1^{\%} = 50\%$, $\mathcal{I}_2^{\%} = 65\%$, $\widetilde{\mathcal{I}}_2^{\%} = 75\%$

Simulation results (n ≈ 500, 10000 datasets)

Binding futility	Method 1	
	2 stages	3 stages
Type 1 error	2.42%	2.50%
Power	81.00%	80.87%
CI=[NA;NA]	0	0.01%
Coverage	94.85%	95.30%

Confidence interval: $[\widehat{\theta}_L; \widehat{\theta}_U]$

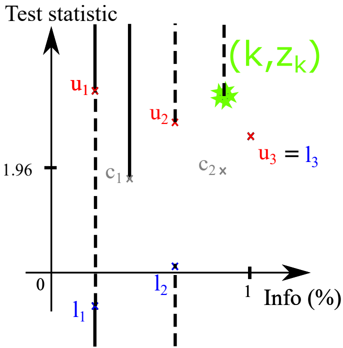
Suppose we stop at stage k with test statistics z_k and $\widetilde{z}_k$

$$\text{Find } \widehat{\theta}_L \text{ solving } \mathbb{P}_{\widehat{\theta}_L}\left[\left(\kappa, \widetilde{Z}_\kappa\right) \succeq (k, \widetilde{z}_k)\right] = \alpha/2$$

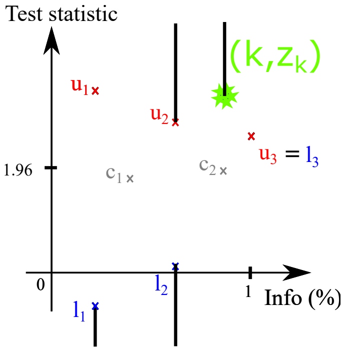
$$\text{Find } \widehat{\theta}_U \text{ solving } \mathbb{P}_{\widehat{\theta}_U}\left[\left(\kappa, \widetilde{Z}_\kappa\right) \succeq (k, \widetilde{z}_k)\right] = 1 - \alpha/2$$

Modification for non-binding futility

- continuation is possible even when $z_k < l_k$



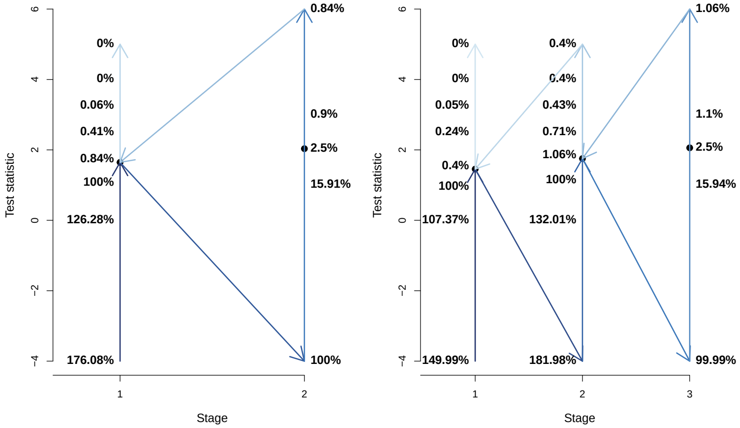
+



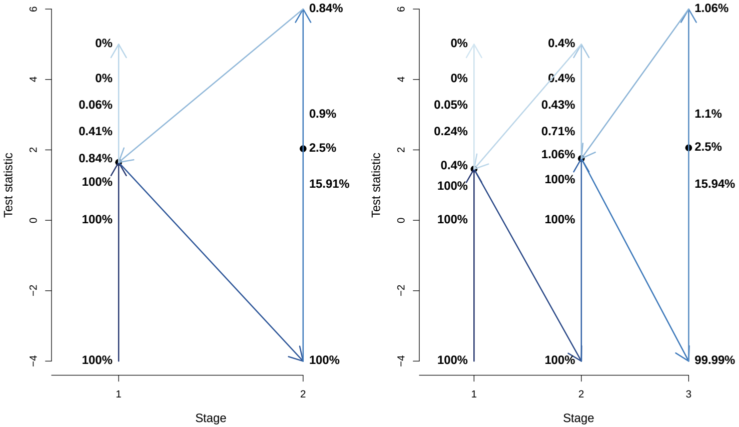
Conservative p-value: some path are counted twice, e.g.

- stage 1 stop for futility and conclude efficacy
- stage 1 continue despite futility, conclude efficacy at stage 2

P-value space (Method 1, non-binding)



P-value space (Method 1, non-binding)



After numerical integration, p-values greater than 1 are set to 1.

Simulation results (n $\approx$ 500, 10000 datasets)

non-Binding futility	Method 1	
	2 stages	3 stages
Type 1 error	2.53%	2.54%
Power	80.50%	80.52%
CI=[NA;NA]	**5.93%**	**8.11%**
Coverage	95.92%	96.09%

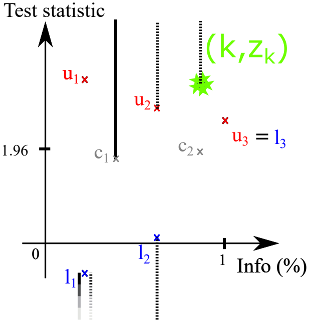
When rejecting early for futility,

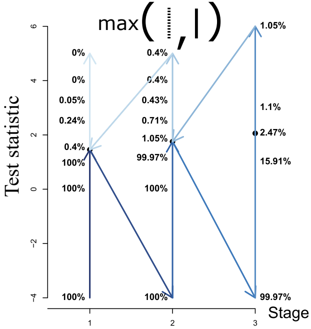
$$x \mapsto \mathbb{P}_x \left[\left(\kappa, \widetilde{Z}_\kappa \right) \succeq (k, \widetilde{z}_k) \right]$$

is essentially constant equal to 1 so the CI cannot be estimated.

A possible improvement?

Instead of summing the rejection probablity of duplicated paths, evaluate the maximal rejection probability.





- all p-values $\in [0,1]$ otherwise very similar values

Conclusion

P-values and CI can be estimated in a GSD with delayed outcome

- extension of Armitage ordering
- median unbiased estimator also available
- implemented in `DelayedGSD`, R package available on Github

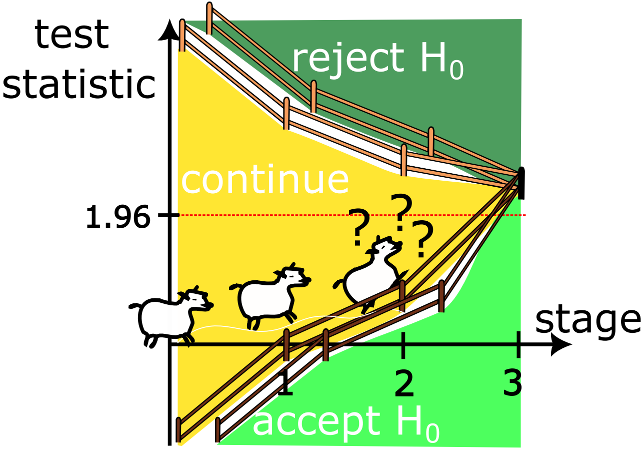
Difficulties:

- non-binding futility stopping rule
  → conservative p-value
  → no reliable CI when concluding early for futility
- constraints on boundaries (e.g. $c_k \geq 1.96$)

More to come:

- Corine Baayen's talk Friday morning
- Upcoming paper

Questions



Reference I

Agency, E. M. (2007). Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design.

Hampson, L. V. and Jennison, C. (2013). Group sequential tests for delayed responses (with discussion). *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 75(1):3–54.

Jennison, C. (2022). The Design of Group Sequential and Adaptive Clinical Trials. *Course Slides*.

Planned information and sample size

$$\mathcal{I}_{\max} = R\frac{\left(\Phi^{-1}(1-\alpha)+\Phi^{-1}(1-\beta)\right)^2}{\theta_0^2}$$

where θ_0 is the expected effect under the alternative hypothesis Φ is the cumulative distribution function of a standard normal distribution. R is the inflation factor which depends on $\alpha_1, \alpha_2, \alpha$, β_1, β_2, β.

Denoting by n the sample size in one arm and nw in the other arm:

$$n = R(1+1/w)\sigma^2\frac{\left(\Phi^{-1}(1-\alpha)+\Phi^{-1}(1-\beta)\right)^2}{\theta_0^2}$$

where σ^2 is the variance of the outcome.

Efficient estimators have canonical covariance

Suppose $\mathbb{C}ov\left[\widehat{\theta}_1, \widehat{\theta}_2\right] \neq \mathbb{V}ar\left[\widehat{\theta}_1\right]$ so $\mathbb{C}ov\left[\widehat{\theta}_1 - \widehat{\theta}_2, \widehat{\theta}_2\right] \neq 0$.

Let ε be a small number of sign opposite to $\mathbb{C}ov\left[\widehat{\theta}_1 - \widehat{\theta}_2, \widehat{\theta}_2\right]$. Consider a new estimator:

- $\widetilde{\theta}_2 = \widehat{\theta}_2 + \varepsilon(\widehat{\theta}_1 - \widehat{\theta}_2)$

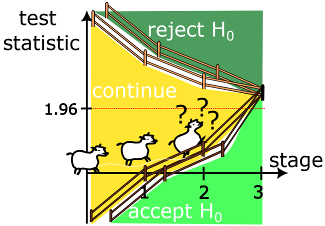
$$
\begin{aligned}
\mathbb{E}\left[\widetilde{\theta}_2\right] &= \theta + \varepsilon(\theta - \theta) = \theta \\
\mathbb{V}ar\left[\widetilde{\theta}_2\right] &= \mathbb{V}ar\left[\widehat{\theta}_2\right] + \varepsilon^2 \mathbb{V}ar\left[\widehat{\theta}_1 - \widehat{\theta}_2\right] + 2\varepsilon \mathbb{C}ov\left[\widehat{\theta}_1 - \widehat{\theta}_2, \widehat{\theta}_2\right] \\
&\approx \mathbb{V}ar\left[\widehat{\theta}_2\right] + 2\varepsilon \mathbb{C}ov\left[\widehat{\theta}_1 - \widehat{\theta}_2, \widehat{\theta}_2\right] \\
&< \mathbb{V}ar\left[\widetilde{\theta}_1\right]
\end{aligned}
$$

contradicting that $\widehat{\theta}_2$ is efficient.

EMA (Agency, 2007)

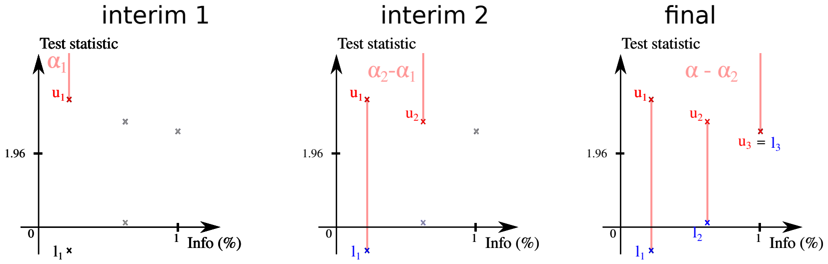
"If a trial is to be terminated as a result of an interim analysis it is always important to carry out an additional analysis including all of these further patients that did not contribute to the interim analysis.

It may be that when this analysis is carried out, the null hypothesis can no longer be rejected [. . .]. In such a situation, it is accepted regulatory practice to base decision making on the final results of the trial (not the interim analysis)."



Estimating boundaries

Boundaries should be such that the probablity of rejection under the null equals the type 1 error to be spent



interim 1 interim 2 final

x-axis: information fraction $\mathcal{I}^{\%} = \mathcal{I}_k / \mathcal{I}_{\max}$

Key result: $\left(\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\theta}_3\right)$ are asymptotically multivariate normal
→ numerical integration of the multivariate Gaussian density

Boundary with method 1 (stage 1)

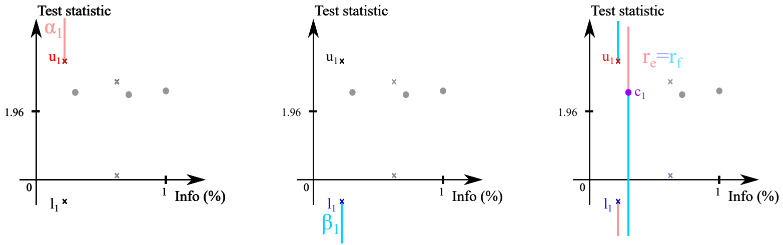
At any interim analysis 1

At the decision analysis 1

Efficacy boundary

Futility boundary

Decision boundary



$$\mathbb{P}_0[Z_1 \geq u_1] = f(\mathcal{I}_1^{\%})$$

$$\mathbb{P}_{\theta_0}[Z_1 \leq l_1] = g(\mathcal{I}_1^{\%})$$

$$\mathbb{P}_0[Z_1 \geq u_1, \widetilde{Z}_1 < c_1] = \mathbb{P}_0[Z_1 \leq l_1, \widetilde{Z}_1 \geq c_1]$$

Boundary with method 1 (stage 2)

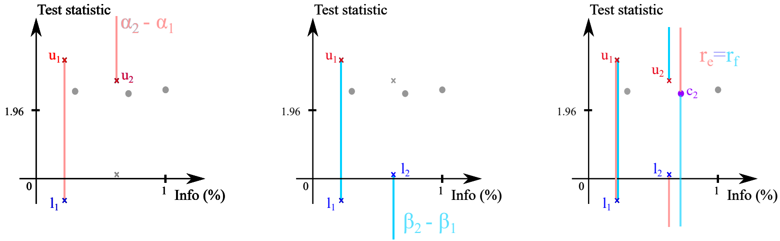
At any interim analysis 2

At the decision analysis 2

Efficacy boundary

Futility boundary

Decision boundary



$$\mathbb{P}_0[Z_1 \in]l_1, u_1[, Z_2 \geq u_2] = f(\mathcal{I}_2^{\%}) - f(\mathcal{I}_1^{\%})$$

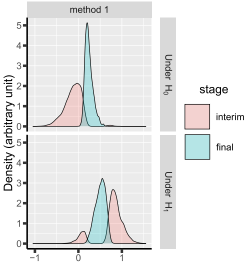
$$\mathbb{P}_{\theta_0}[Z_1 \in]l_1, u_1[, Z_2 \leq l_2] = g(\mathcal{I}_2^{\%}) - g(\mathcal{I}_1^{\%})$$

$$\mathbb{P}_0[Z_1 \in]l_1, u_1[, Z_2 \geq u_2, \widetilde{Z}_2 < c_2] = \mathbb{P}_0[Z_1 \in]l_1, u_1[, Z_2 \leq l_2, \widetilde{Z}_2 \geq c_2]$$

Corrected point estimate

The standard maximum likelihood estimator (MLE) is biased. Typically:

- positive bias for $\theta > 0$
- negative bias for $\theta < 0$



There is no uniformly minimum variance unbiased estimator for θ

Median unbiased estimator (MUE) is easy to compute

$$\text{Find } \theta' \text{ solving } \mathbb{P}_{\theta'}\left[\left(\kappa, \widetilde{Z}_{\kappa}\right) \succeq (k^*, \widetilde{z}^*)\right] = 0.5$$

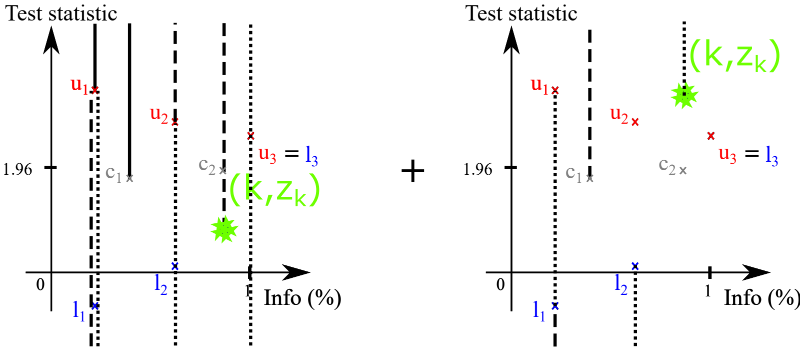
Simulation setting

Inspired by a real dataset from Lundbeck

- $K = 2$ or $K = 3$
- $n_K \in [491, 557]$,
  from a power calculation to get 80% power with method 3
- $\widehat{\theta}$: MLE from a linear mixed model
- 3 measurements: (Y_0, Y_1, Y_2), correlation range $[-0.15, 0.68]$
- group difference: $(0, 0.3, 0.6)$
- about 10% observations with one or more missing values
- Information rate:
  2 stages $\mathcal{I}_1^{\%} = 58\%$, $\widetilde{\mathcal{I}}_1^{\%} = 68\%$
  3 stages: $\mathcal{I}_1^{\%} = 40\%$, $\widetilde{\mathcal{I}}_1^{\%} = 50\%$, $\mathcal{I}_2^{\%} = 65\%$, $\widetilde{\mathcal{I}}_2^{\%} = 75\%$
- 10 000 datasets

Modification for non-binding futility

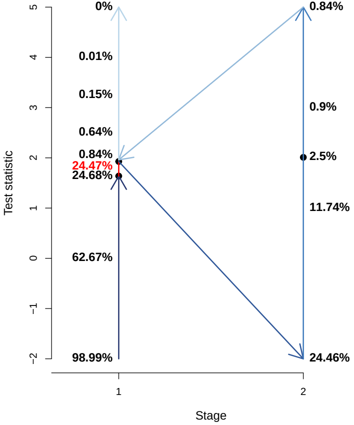
When concluding futility at stage 2

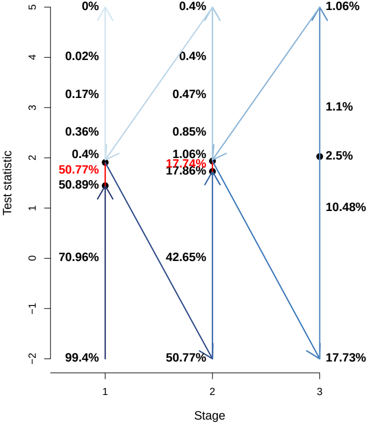


Modification for $c_k \geq \Phi^{-1}(1-\alpha)$

With decision boundary $\underline{c}_k = \max(c_k, \Phi^{-1}(1-\alpha))$,

$\mathbb{P}_0\left[\left(\kappa, \widetilde{Z}_\kappa - (\underline{c}_k - c_k)\right) \succeq (k, \widetilde{z}_k)\right]$ is a conservative p-value.


Test statistic
Stage


Test statistic
Stage