UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Faculty of Health Sciences

Group Sequential Trials for Survival Data

Paul Blanche

Section of Biostatistics, University of Copenhagen

October 28, 2021

Acknowledgments:
many slides are based of those of Prof. Chris Jennison, from the shortcourse organized by Danish Society of Biopharmaceutical Statistics, in September 2019.

Material (figures, examples) also taken from:

- Baayen, Volteau, Flamant & P. Blanche. "Sequential trials in the context of competing risks: Concepts and case study, with R and SAS code." Statistics in medicine 38, no. 19 (2019): 3682-3702.
- Jennison, Turnbull. "Group sequential designs for survival data". In: Klein JP, van Houwelingen HC, Ibrahim JG, Scheike TH, eds. Handbook ofsurvival analysis. Boca Raton, FL: Chapman and Hall/CRC; 2013:595-613

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Outline

Motivation: Phase 3 clinical trials

Phase III trials are conducted as the last stage in the drug development process.

Two positive studies are usually required to confirm that a new treatment is superior to the current standard treatment.

Regulators customarily require a hypothesis test to reach significance at the one-sided 2.5% level.

Studies may recruit hundreds, or even thousands, of subjects at a cost of as much as 10k to 50k euros per patient.

The time taken to reach a conclusion eats into the limited patent lifetime remaining to the company developing the drug.

Thus, there are strong incentives to reach an early conclusion for either a positive or negative decision.

What are GST?

"Group sequential trials (GST) allow researchers to evaluate accumulating data at preplanned time intervals during a trial, without compromising the validity of the final analysis results. These interim looks at the data allow for early stopping of a trial in case of a clear effect of the treatment or a clear lack thereof."

(Baayen et al, SiM 2019)

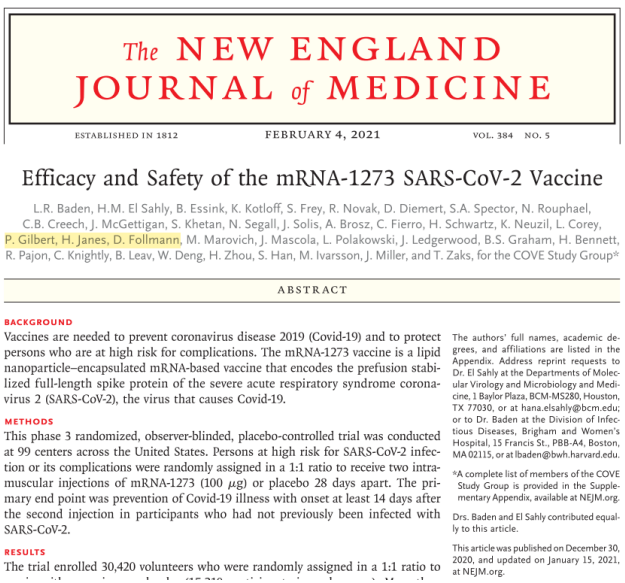
Why GST?

Group sequential trials can be:

- more ethical
  - stop early if clearly effective / ineffective treatment: patients are better treated faster!
    (e.g. covid-19 vaccine available earlier)
- economically interesting
  - in average, smaller sample size (reduced costs)
  - shorter trial if stop early implies longer period the drug can be marketed with patent protection.

GST are becoming increasingly popular in the pharmaceutical industry and regulatory agencies (EMA, FDA) now tend to consider them as the "default approach".

Example: Moderna covid-19 vaccine



The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 FEBRUARY 4, 2021 VOL. 384 NO. 5

Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine

L.R. Baden, H.M. El Sahly, B. Essink, K. Kotloff, S. Frey, R. Novak, D. Diemert, S.A. Spector, N. Rouphael, C.B. Creech, J. McGettigan, S. Khetan, N. Segall, J. Solis, A. Brosz, C. Fierro, H. Schwartz, K. Neuzil, L. Corey, P. Gilbert, H. Janes, D. Follmann, M. Marovich, J. Mascola, L. Polakowski, J. Ledgerwood, B.S. Graham, H. Bennett, R. Pajon, C. Knightly, B. Leav, W. Deng, H. Zhou, S. Han, M. Ivarsson, J. Miller, and T. Zaks, for the COVE Study Group*

ABSTRACT

BACKGROUND

Vaccines are needed to prevent coronavirus disease 2019 (Covid-19) and to protect persons who are at high risk for complications. The mRNA-1273 vaccine is a lipid nanoparticle–encapsulated mRNA-based vaccine that encodes the prefusion stabilized full-length spike protein of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the virus that causes Covid-19.

METHODS

This phase 3 randomized, observer-blinded, placebo-controlled trial was conducted at 99 centers across the United States. Persons at high risk for SARS-CoV-2 infection or its complications were randomly assigned in a 1:1 ratio to receive two intramuscular injections of mRNA-1273 (100 μg) or placebo 28 days apart. The primary end point was prevention of Covid-19 illness with onset at least 14 days after the second injection in participants who had not previously been infected with SARS-CoV-2.

RESULTS

The trial enrolled 30,420 volunteers who were randomly assigned in a 1:1 ratio to

The authors' full names, academic degrees, and affiliations are listed in the Appendix. Address reprint requests to Dr. El Sahly at the Departments of Molecular Virology and Microbiology and Medicine, 1 Baylor Plaza, BCM-MS280, Houston, TX 77030, or at hana.elsahly@bcm.edu; or to Dr. Baden at the Division of Infectious Diseases, Brigham and Women's Hospital, 15 Francis St., PBB-A4, Boston, MA 02115, or at lbaden@bwh.harvard.edu.

*A complete list of members of the COVE Study Group is provided in the Supplementary Appendix, available at NEJM.org.

Drs. Baden and El Sahly contributed equally to this article.

This article was published on December 30, 2020, and updated on January 15, 2021, at NEJM.org.

(2 interim analyses, Cox model, powered for modest effect VE=1-HR=0.6, found was 0.94)

Another: Lin et al (Biometrics, 1996)

- Wilms tumor (in kidney) in children under the age of 15 years.
- Current treatment: chemotherapy after surgery.
- Experimental treatment: no further treatment after surgery for patients under two years of age with a tumor of "favorable" histology.
- Criterion for evaluating this new treatment (among eligible children) is the proportion of patients who remain continuously disease-free and alive for two years following surgery ("cured"). Should be $\geq 95\%$.
- If interim results turn out worse than anticipated, it will be mandatory to stop further patient entry immediately and to commence chemotherapy for those already enrolled who may be in danger of relapse.

(one arm trial, 5 interim analyses, i.e. every year, Kaplan-Meier)

Survival data and GST

GST methodology is available for continuous outcome (e.g. blood pressure), binary outcome (e.g. pain-free 2 hours after taking a drug) or survival outcomes (time to event).

For survival data, one major challenge is that is difficult to know in advance how much information we will be available at each interim analysis. This is problematic because this plays an important for accounting for the multiple testing induced by the interim analyses, to control the type-I error ($\rightarrow$).

Hence the simplest methods for GST are usually not well suited to survival data and specific methods are needed.

Statistical framework

Suppose a new treatment (Treatment A) is to be compared to a placebo or positive control (Treatment B) in a Phase III trial.

The treatment effect θ for the primary endpoint represents the advantage of Treatment A over Treatment B.

If $\theta > 0$, Treatment A is more effective.

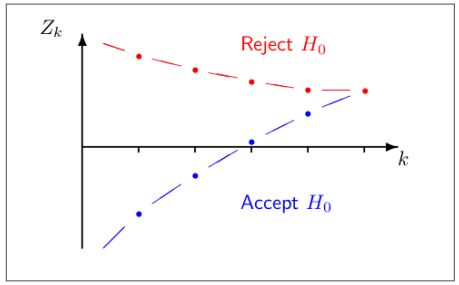
We wish to test the null hypothesis $H_0 : \theta \leq 0$ against $\theta > 0$ with

- $P_{\theta=0}\{\text{Reject } H_0\} = \alpha$ (type-I error)
- $P_{\theta=\delta}\{\text{Reject } H_0\} = 1 - \beta$ (power).

In a group sequential trial, data are examined on a number of occasions, $\boxed{k = 1, \ldots, K}$, to see if an early decision may be possible.

The main statistical challenge

What we want (most) is to find boundaries. Typical boundaries for a one-sided test, expressed in terms of standardised test statistics $Z_1, \ldots, Z_K$ (with $K - 1$ interim analyzes), have the form:



Crossing the upper boundary leads to early stopping for a positive outcome, rejecting H_0 in favour of $\theta > 0$.

(Rk: unlike in most observational studies, in properly powered trials we can "accept" H_0).

Outline

Canonical joint distribution

Let $\widehat{\theta}_k$ denote the estimate of θ based on data at analysis k.

The information for θ at analysis k is

$$\mathcal{I}_k = \left\{ \text{Var}(\widehat{\theta}_k) \right\}^{-1}, \qquad k = 1, \ldots, K.$$

In many situations, $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$ are approximately multivariate normal, with the **canonical joint distribution** defined by

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

and

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

The simplest (but important) example

The joint distribution of $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$ can be derived "directly" when θ is a single normal mean, we observe $X_1, X_2, \ldots$ i.i.d $N\left(\theta, \sigma^2\right)$ and $\hat{\theta}_k$ is the MLE

$$\widehat{\theta}_k = \frac{X_1 + \ldots + X_{n_k}}{n_k},$$

for $k = 1, \ldots, K$, with $n_1 < n_2 < \cdots < n_K$.

Warm-up Exercise: check that $(\widehat{\theta}_1, \ldots, \widehat{\theta}_K)$ follows the canonical joint distribution.

Exercise: check the same for the two sample case. That is, suppose observations on Treatments A and B, respectively, are

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

and $\theta = \mu_A - \mu_B$. At analysis k, with n_{Ak} observations on Treatment A and n_{Bk} on Treatment B,

$$\widehat{\theta}_k = \widehat{\mu}_{A,k} - \widehat{\mu}_{B,k} = \frac{1}{n_{Ak}} \sum_{i=1}^{n_{Ak}} X_{Ai} - \frac{1}{n_{Bk}} \sum_{i=1}^{n_{Bk}} X_{Bi}$$

Sequential distribution theory

Remarkably, almost always, $(\widehat{\theta}_1, \ldots, \widehat{\theta}_K)$ follows, asymptotically, the canonical joint distribution.

This is especially the case when θ is:

- a parameter in a general parametric model (eg GLM) fitted by MLE
- a hazard ratio in a Cox model
- a subdistribution hazard ratio in Fine-Gray model
- a difference or ratio of t-year risk or survival estimated using Kaplan-Meier or Aalen-Johansen
- an odds ratio fitted by solving MLE score equations, using an inverse probability of censoring weighted binary outcome.
- ....

Where to find the maths?

Important work about the above statement include, for general results:

- Scharfstein, Tsiatis & Robins (JASA, 1997)
- Jennison & Turnbull (JASA, 1997)

For results with survival data:

- Tsiatis, Rosner & Tritchler (Bka, 1985)
- Logan & Zhang (SiM, 2013)
- Martens & Logan (Bcs, 2018)
- Martens & Logan (LIDA, 2020)

A simple, nice, result!

Jennison & Turnbull (1997) provide a simple heuristic explanation of why the canonical joint distribution almost always (asymptotically) holds.

Result: all efficient estimators have the canonical covariance. That is, if $\widehat{\theta}_1$ at analysis 1 and $\widehat{\theta}_2$ at analysis 2 are efficient, unbiased estimators of θ, then $\mathrm{Cov}\left(\widehat{\theta}_1, \hat{\theta}_2\right) = \mathrm{Var}\left(\widehat{\theta}_2\right)$.

Proof (by contradiction): Suppose $\mathrm{Cov}\left(\hat{\theta}_1, \hat{\theta}_2\right) \neq \mathrm{Var}\left(\hat{\theta}_2\right)$, so $\mathrm{Cov}\left(\widehat{\theta}_1 - \hat{\theta}_2, \hat{\theta}_2\right) \neq 0$. Consider an unbiased estimator of the form $\widehat{\theta}_2^* = \widehat{\theta}_2 + \epsilon\left(\widehat{\theta}_1 - \widehat{\theta}_2\right)$. For ϵ small and of the opposite sign to $\mathrm{Cov}\left(\widehat{\theta}_1 - \widehat{\theta}_2, \hat{\theta}_2\right)$, then $\mathrm{Var}\left(\hat{\theta}_2^*\right) < \mathrm{Var}\left(\widehat{\theta}_2\right)$, which contradicts the assumption that $\widehat{\theta}_2$ is an efficient estimator of θ.

Canonical joint distribution of Z-statistics

In testing $H_0 : \theta = 0$, the standardised statistic at analysis k is

$$Z_k = \frac{\widehat{\theta}_k}{\sqrt{\mathrm{Var}\left(\widehat{\theta}_k\right)}} = \widehat{\theta}_k \sqrt{\mathcal{I}_k}$$

For these statistics, asymptotically,

$$(Z_1, \ldots, Z_K) \quad \text{is multivariate normal,}$$
$$Z_k \sim N\left(\theta\sqrt{\mathcal{I}_k}, 1\right), \quad k = 1, \ldots, K$$
$$\mathrm{Cov}\left(Z_{k_1}, Z_{k_2}\right) = \sqrt{\mathcal{I}_{k_1}/\mathcal{I}_{k_2}} \quad \text{for } k_1 < k_2$$

This is super nice!

Why is it so nice?

In short: whatever the data we collect, the parameter θ of interest, the variance $\mathrm{Var}\left(\widehat{\theta}_k\right)$ or equivalently the information $\mathcal{I}_k^{-1}$, what is important to know is only how much information is acquired between successive analyses.

Outline

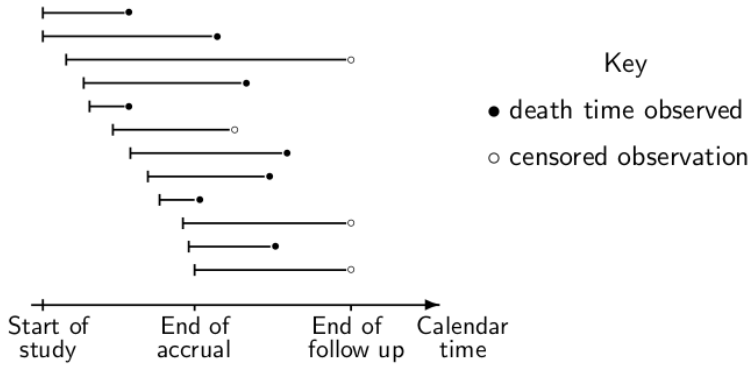
Why are survival data challenging?

Usually, with simple continuous or binary outcome, the variance and information is proportional to the sample size and we have

$$\mathrm{Cov}\left(Z_{k_1}, Z_{k_2}\right) = \sqrt{\mathcal{I}_{k_1}/\mathcal{I}_{k_2}} = \sqrt{\frac{n_{k_1}}{n_{k_2}}}$$

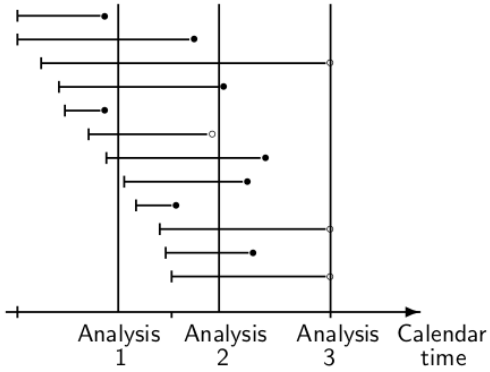
Hence, the canonical distribution depends only on the sample size of newly included patients at each analysis (and of one additional parameter, say the variance at the last planned analysis).

With survival data, this (usually) does not hold anymore. This is because the information accumulated between two successive analyses also depends on the increase in follow-up of the patients already included in the study at the time of the first analysis.



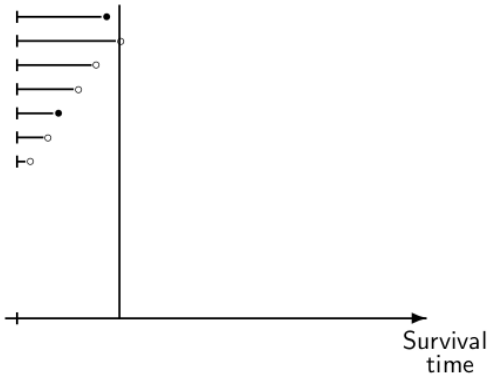
Subjects are randomised to a treatment as they enter the study.

Survival is measured from entry to the study.



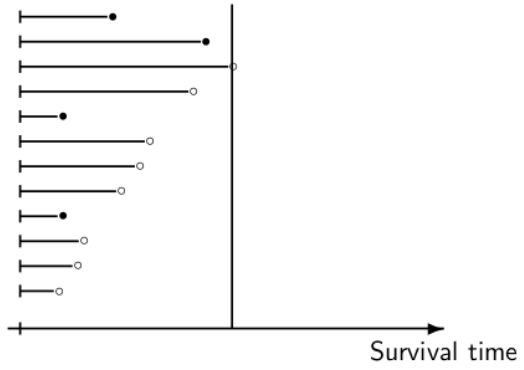
At an interim analysis, subjects are censored if they are still alive.

Information on such patients continues to accrue at later analyses.



We analyse data on survival from time of randomisation.

Survival times start at zero and censoring occurs at the analysis time for subjects surviving past this first analysis.



At interim analysis 2, there is further follow-up of subjects who were censored at analysis 1.

In addition, there is initial information on the survival times of subjects entering the trial since analysis 1.

Information and Cox regression

For Cox proportional-hazards models, where $\theta = \log(HR)$, with HR the hazard ratio for the randomized treatment (adjusted or not), we have, asymptotically (Tsiatis et al. 1985),

$$\mathcal{I}_k = \left\{\text{expected number of } \textbf{observed events} \text{ at analysis } k\right\} \cdot p_0 \cdot p_1$$

where p_0 and p_1 are the proportions of patients randomized to the treatment and control groups (often $=1/2$).

The expected number of observed events depends on both the **sample size** and **length of follow up**. (censoring, accrual rate...)

It also works for:

- log-rank test (=score test in Cox model)
- Fine-Gray subdistribution hazard (Latouche & Porcher, 2007)

Canonical joint distribution of score statistics

The score statistics, $S_k = Z_k\sqrt{\mathcal{I}_k}$, are also multivariate normal with

$$S_k \sim N\left(\theta\mathcal{I}_k, \mathcal{I}_k\right), \quad k = 1, \ldots, K$$

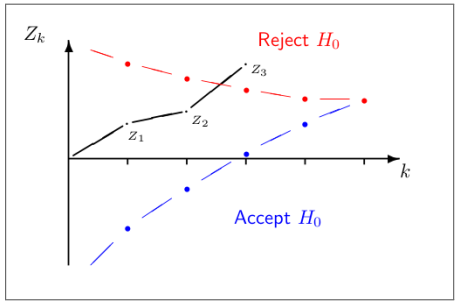
The score statistics possess the "independent increments" property,

$$\text{Cov}\left(S_k - S_{k-1}, S_{k'} - S_{k'-1}\right) = 0 \quad \text{for } k \neq k'$$

It can be helpful to know that the score statistics behave as Brownian motion with drift θ observed at times $\mathcal{I}_1, \ldots, \mathcal{I}_K$, and some papers refer to that extensively.

Outline

Computation for group sequential tests



Today, probabilities such as $P_\theta\{a_1 < Z_1 < b_1, a_2 < Z_2 < b_2, Z_3 > b_3\}$ can easily be computed (e.g. `R` package `mvtnorm`).

Combining these probabilities yields type I error rate, power, expected sample size, etc., of a group sequential design.

Boundaries and group sizes can be chosen to define a test with a specified type I error probability and power.

Type-I error, with binding futility boundary:

$$
\begin{aligned}
&P_{\theta=0}\{\text{Reject } H_0\} = \alpha \\
&= \sum_{k=1}^{K} P_{\theta=0}\{\text{Reject } H_0 \text{ at analysis } k\} \\
&= \sum_{k=1}^{K} P_{\theta=0}\left\{a_1 < Z_1 < b_1, \ldots, a_{k-1} < Z_{k-1} < b_{k-1}, Z_k > b_k\right\}
\end{aligned}
$$

with non binding:

$$
= \sum_{k=1}^{K} P_{\theta=0}\left\{Z_1 < b_1, \ldots, Z_{k-1} < b_{k-1}, Z_k > b_k\right\}
$$

Power (any case, assuming we stop for futility if data suggest so): same as for type-I error but $P_{\theta=\delta}$ instead of $P_{\theta=0}$.

Binding or non-binding futility boundaries?

"the investigators sometimes want to have the opportunity to continue the trial while the data suggest stopping for futility. In that case, the futility boundaries are used for guidance only but not for defining a ***binding*** *rule."* (Baayen et al 2019)

E.g. collecting more data on secondary endpoints can be interesting. It can be allowed if continuing the trial after crossing the futility boundary is not thought unethical.

If a futility boundary is deemed to be non-binding, the type I error rate should be computed ignoring the futility boundary. (worst case scenario)

However, investigators will wish to know power and expected sample size when the futility boundary is obeyed. (expected behavior)

Hands on exercise (R)

Consider a group sequential test with $K = 2$, i.e. 1 interim analysis. Using the R package `mvtnorm` and the function `pmvnorm`, check that the following boundaries are appropriate for the information levels, with type-I error rate $\alpha = 5\%$, power $1 - \beta = 80\%$ and expected effect $\delta = \log(2)$, when stopping for futility is **non-binding**.

k	a_k (futility)	b_k (efficacy)	$\mathcal{I}_k$
1	0.16	2.24	6.81
2	1.70	1.70	13.62

Example from Baayen et al (2019). Note that `qnorm(0.95)`=1.64 < 1.70 < `qnorm(0.975)`=1.96.

Information for "fixed" testing (i.e. non-sequential)

In order to test $H_0 : \theta \leq 0$ against $\theta > 0$ with type-I error probability α and power $1 - \beta$ at $\theta = \delta$, a fixed sample size study needs information

$$\boxed{\mathcal{I}_f = \frac{(z_\alpha + z_\beta)^2}{\delta^2}}$$

Here the subscript f stands for "fixed" and z_γ is the $(1 - \gamma)$-th quantile of the standard normal distribution, i.e. $P(Z > z_\gamma) = \gamma$.

Simple proof

This is because

- $P_{\theta=0}(Z > z_\alpha) = \alpha$
- $P_{\theta=\delta}(Z > z_\alpha) = 1 - \beta$
- $Z \sim N\left(\theta\sqrt{\mathcal{I}}, 1\right)$

and therefore

$$P_{\theta=\delta}\Big(\underbrace{Z - \delta\sqrt{\mathcal{I}}}_{N(0,1)} > \underbrace{z_\alpha - \delta\sqrt{\mathcal{I}}}_{z_{1-\beta}}\Big) = 1 - \beta$$

hence

$$z_\alpha - \delta\sqrt{\mathcal{I}} = z_{1-\beta} \quad \text{and} \quad \mathcal{I} = \frac{(z_\alpha + z_\beta)^2}{\delta^2}$$

as $z_\beta = -z_{1-\beta}$, by symmetry of the normal distribution.

Information for group sequential testing

Because of the multiple testing issue (K hypothesis tests), necessarily,

$$a_K = b_K = z_{\alpha'} > z_\alpha \ , \quad \text{where} \quad \alpha' < \alpha \ .$$

otherwise the type-I error is above α, because of the multiple testing issue.

Hence we need the information at the final analysis ($k = K$) to be larger than that for a fixed test, i.e.,

$$\boxed{\mathcal{I}_K = R \cdot \mathcal{I}_f} \quad \text{for some } R > 1 \ .$$

We call R the "inflation factor".

Example: hands on exercise cont'

1. Which information $\mathcal{I}_f$ would we need for a fixed test, with type-I error rate $\alpha = 5\%$, power $1 - \beta = 80\%$ and effect effect $\delta = \log(2)$?
2. In the previous example of GST, with $K = 2$, we had $\mathcal{I}_K = 13.62$. What was the value of the inflation factor R is that case?

   In other words, how much more information do we need at the final analysis to compensate for the fact that we chose to have an interim analysis, with the option to stop the trial early for efficacy?

Use R !

Average information (and sample size)

Because of the inflation factor $R > 1$, we might end-up needing more information –and hence a larger sample size– to conclude with a GST than with a fixed test.

However, in average, we need less and the gains can be substantial.

More importantly, we have a (possibly) non negligible chance to stop the trial earlier.

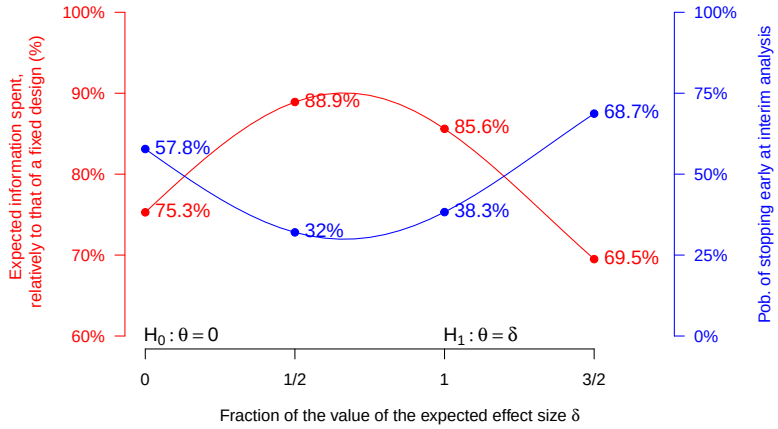
Example: hands on exercise cont'

Assume that we stop for futility when the data suggest so and consider these cases for the true treatment effect:

- $\delta = \log(2)$ (as assumed for the power calculation)
- $\delta = 0$ (no effect)
- $\delta = 0.5 \cdot \log(2)$ (smaller than expected)
- $\delta = 1.5 \cdot \log(2)$ (larger than expected)

How much information will we spend before ending the trial, in average, in each case? How does this compares to $\mathcal{I}_f$, the information that we would spend when using a fixed test? What are our chances to stop early, i.e., at the interim analysis?

Use R !



Note that in this example, when we stop early, we end up collecting only $0.5 \times R = 52.9\%$ of the information needed for a fixed test (here the inflation factor is $R = 1.058$).

Defining (planned) boundaries and inflation factor (1/2)

Recall that:

- $P_{\theta=0}\{\text{Reject } H_0\} = \alpha$ (type-I error rate)
- $P_{\theta=\delta}\{\text{Reject } H_0\} = 1-\beta$ (power).
- $P\{\text{Reject } H_0\} = \sum_{k=1}^{K} P\{\text{Reject } H_0 \text{ at analysis } k\}$
- $\mathcal{I}_f = (z_\alpha + z_\beta)^2/\delta^2$

Further:

1. Assume that $\mathcal{I}_k = (k/K)\cdot\mathcal{I}_K = (k/K)\cdot R\cdot\mathcal{I}_f$ (equally spaced information levels)
2. Define how much of α and β to "spend" at each analysis k, via the choice of a parameter $\rho > 0$ for the spending functions $f(\cdot)$ and $g(\cdot)$:

$$f(\mathcal{I}) = \alpha\min\{1,(\mathcal{I}/\mathcal{I}_K)^\rho\} \quad \text{and} \quad g(\mathcal{I}) = \beta\min\{1,(\mathcal{I}/\mathcal{I}_K)^\rho\}$$

The smaller ρ the easier to stop early for efficacy, but this might lead to higher average information spent

Defining (planned) boundaries and inflation factor (2/2)

This means that, for each analysis k, we want:

$$P_{\theta=0}\{\text{Reject } H_0 \text{ at analysis } k\} = f(\mathcal{I}_k) - f(\mathcal{I}_{k-1})$$
$$P_{\theta=\delta}\{\text{Reject } H_0 \text{ at analysis } k\} = g(\mathcal{I}_k) - g(\mathcal{I}_{k-1})$$

3. Remember that we want to impose $a_K = b_K$, i.e., we want to stop and conclude at analysis K at the latest.
4. Conclude that this defines the (unique) value of the inflation factor R, which depends on α, β, K and ρ.

Indeed, for given α, β, K, ρ and $\mathcal{I}_K$:

- At $k=1$, boundaries a_1 and b_1 are uniquely defined by

$$P_{\theta=0}\{Z_1 > b_1\} = \alpha\cdot(1/K)^\rho \quad \text{(stop for efficacy)}$$
$$P_{\theta=\delta}\{Z_1 < a_1\} = \beta\cdot(1/K)^\rho \quad \text{(stop for futility)}$$

- At $k=2$, boundaries a_2 and b_2 are uniquely defined (given a_1 and b_1) by

$$P_{\theta=0}\{Z_1 < b_1, Z_2 > b_2\} = \alpha\left[(2/K)^\rho - (1/K)^\rho\right] \quad \text{(non binding case)}$$
$$P_{\theta=\delta}\{a_1 < Z_1 < b_1, Z_2 < a_2\} = \beta\left[(2/K)^\rho - (1/K)^\rho\right] \quad \text{(stop as suggested)}$$

- etc....
- But we need to end up with $a_K = b_K$, which uniquely defines $\mathcal{I}_K$ (given α, β, K, ρ) and so the inflation factor R, as $\mathcal{I}_K = R\cdot\mathcal{I}_f$.

  Note that although not explicit in the above, the probabilities computed under the alternative ($\theta=\delta$), unlike those computed under the null ($\theta=0$), depends on $\mathcal{I}_K$ (and hence R) via $Z_k \sim N\big(\theta\sqrt{\mathcal{I}_k},1\big)$ with $\mathcal{I}_k = (k/K)\mathcal{I}_K$.

Software will take care :-)

Of course many software (e.g. `R`, `SAS`) provide boundaries $(a_k,b_k)_{k=1,\ldots,K}$ and inflation factor R for given α, β, K and ρ.

Note that they do not depend on δ, as $Z_k \sim N(\delta\sqrt{\mathcal{I}_k},1)$ under the alternative and

$$\begin{aligned}\delta\sqrt{\mathcal{I}_k} &= \delta\sqrt{(k/K)\cdot\mathcal{I}_K}\\ &= \delta\sqrt{(k/K)\cdot R\cdot\mathcal{I}_f}\\ &= \delta\sqrt{(k/K)\cdot R\cdot(z_\alpha+z_\beta)^2/\delta^2}\\ &= (z_\alpha+z_\beta)\sqrt{(k/K)\cdot R}\end{aligned}$$

Hence you can find the values of R in some books, for common choices of α, β, K and ρ.

R package gsDesign (1/2)

R code:

```
library(gsDesign)
gsDesign(k=2,           # K=2 analyses
         test.type = 4, # non-binding futility boundaries
         alpha=0.05,    # type-I error
         beta=0.2,      # 1-power
         sfu=sfPower,   # spending function (upper boundaries)
         sfl=sfPower,   # spending function (lower boundaries)
         sfupar=2,  # rho=2 (upper)
         sflpar=2)  # rho=2 (lower)
```

Output (partial):

```
          Sample
             Size   ----Lower bounds----  ----Upper bounds-----
 Analysis Ratio*  Z   Nominal p Spend+  Z   Nominal p Spend++
        1  0.529 0.16    0.5652   0.05 2.24    0.0125  0.0125
        2  1.059 1.70    0.9554   0.15 1.70    0.0446  0.0375
    Total                        0.2000                0.0500
```

And we recognize key values from our example ... (1.059 instead of 1.058 due to minor roundin

approximations used in previous slides)

R package gsDesign (2/2)

R code:

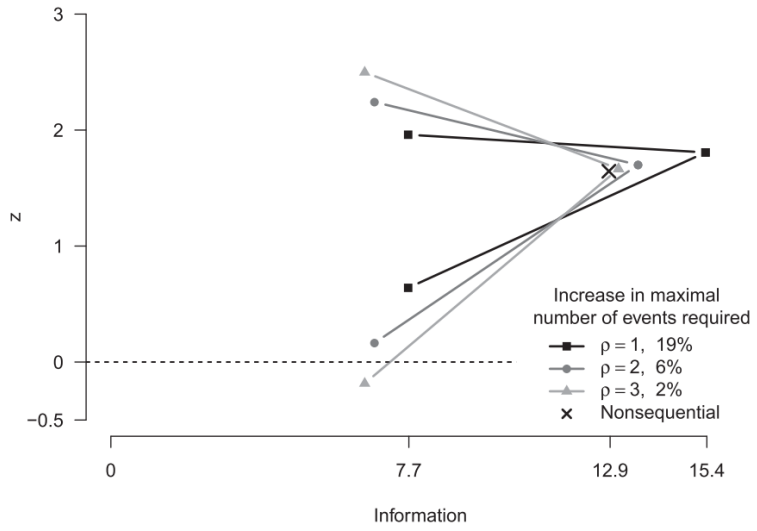
```
library(gsDesign)
gsDesign(k=2,           # K=2 analyses
         test.type = 4, # non-binding futility boundaries
         alpha=0.05,    # type-I error
         beta=0.2,      # 1-power
         n.fix=12.868,  # If
         sfu=sfPower,   # spending function (upper boundaries)
         sfl=sfPower,   # spending function (lower boundaries)
         sfupar=2,      # rho=2 (upper)
         sflpar=2)$n.I  # rho=2 (lower)
```

Output:

```
[1]  6.811048 13.622095
```

And we recognize key values from our example ...

Impact of ρ (Example cont')



(Baayen et al 2019)

Unpredictable information: a challenge!

The amount of information $\mathcal{I}_K$ at analysis k that we will observed is usually impossible to predict "exactly". This is especially problematic with survival data, because the information will generally depends not only on the sample size but also on the length of follow-up, which depends on random inclusion times (and possibly dropout), as already discussed.

This difficulty in predicting the information levels at each analysis will necessarily impact the power of the study, but we need to find a solution to analyze the data such that this does not impact the type-I error control.

Unpredictable information: a solution

To control the type-I error with unpredictable information levels, we need to recalculate the boundaries at each interim analysis using the observed level of information instead of the predicted/planned one used for planning the trial.

This is possible because, as we have (briefly) seen before, the computation of the boundaries (a_k, b_k) at analysis k depends only on information $\mathcal{I}_k$ and past information levels and boundaries related to previous analyses, i.e. $(a_i, b_i, \mathcal{I}_i)$ for $i < k$ (and of pre-specified α, β, K, ρ and also consequently of the chosen inflation factor R).

In short, recalculation is possible as the calculation needed for each analysis k depends only on quantities observed by that time.

Unpredictable information: case study cont'

Consider again $K = 2$, $\alpha = 5\%$, $\beta = 20\%$, $\delta = \log(2)$, non-binding futility boundaries, which led to $\mathcal{I}_f = 12.868$ and $R = 1.0584$, and the planned boundaries and information levels

k	a_k (futility)	b_k (efficacy)	$\mathcal{I}_k$
1	0.16	2.24	6.81
2	1.70	1.70	13.62

Now, suppose that when running the trial we actually observe, at interim analysis $(k = 1)$:

- $\widehat{\mathcal{I}}_1 = \left\{\widehat{\text{Var}}(\widehat{\theta}_1)\right\}^{-1} = \{0.36\}^{-1} = 7.52$

How should we update the boundaries $(a_1, b_1) = (0.16, 2.24)$?

Example from Baayen et al (2019)

We just solve the same equations as for planning the boundaries, but replace the "planned" information level $(\mathcal{I}_1 = 6.81)$ by the observed level $(\widehat{\mathcal{I}}_1 = 7.52)$, while using the **pre-specified spending functions**, defined via $\rho = 2$ and the planned $\mathcal{I}_K$, called the "maximum information", $\mathcal{I}_{max} = R \cdot \mathcal{I}_f = 1.058 \cdot 12.868 = 13.62$ as

$$f(\mathcal{I}) = \alpha \min\left\{1, (\mathcal{I}/\mathcal{I}_{max})^\rho\right\} \quad \text{and} \quad g(\mathcal{I}) = \beta \min\left\{1, (\mathcal{I}/\mathcal{I}_{max})^\rho\right\}$$

Hence, we solve

$$\begin{aligned} P_{\theta=0}\{Z_1 > b_1\} &= f(\widehat{\mathcal{I}}_1) && \text{(stop for efficacy)} \\ &= \alpha \cdot \left(\widehat{\mathcal{I}}_1/\mathcal{I}_{max}\right)^\rho \\ &= 0.05 \cdot (7.52/13.62)^2 \\ P_{\theta=\delta}\{Z_1 < a_1\} &= 0.20 \cdot (7.52/13.62)^2 && \text{(stop for futility)} \end{aligned}$$

This leads to the updated boundaries $(a_1, b_1) = (0.35, 2.16)$, using the fact that, under the alternative $\theta = \delta$, then $Z_1 \sim N(\delta\sqrt{\widehat{\mathcal{I}}_1}, 1)$.

R code:

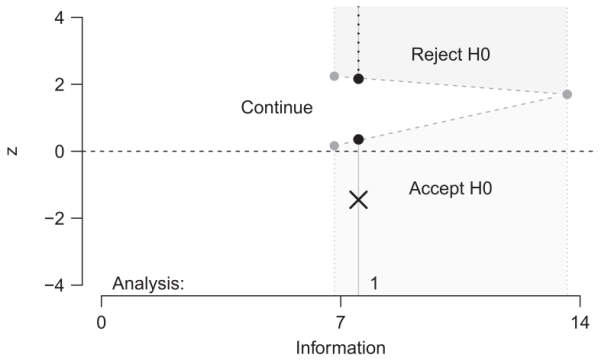
```
b1.Updated <- qnorm(p=0.05*(7.52/13.61)^2,
                    mean=0,
                    lower.tail=FALSE)
#--
a1.Updated <- qnorm(p=0.2*(7.52/13.61)^2,
                    mean=log(2)*sqrt(7.52),
                    lower.tail=TRUE)
#--
round(c(a1.Updated,b1.Updated),2)
```

Output:

```
[1] 0.35 2.16
```

Of course for $k > 1$ the code is a bit more complicated...

Visualizing planned vs used boundaries



Example from Baayen et al (2019)

R package gsDesign

R code:

```
library(gsDesign)
gsDesign(k=2,
         test.type = 4,
         alpha=0.05,
         beta=0.2,
         sfu=sfPower,
         sfl=sfPower,
         sfupar=2,
         sflpar=2,
         n.fix=12.868,        # If
         n.I=c(7.52,13.62),   # Observed I_1 and I_max
         maxn.IPlan = 13.62) # I_max
```

Output (partial):

```
                ----Lower bounds----  ----Upper bounds-----
 Analysis N   Z   Nominal p Spend+  Z   Nominal p Spend++
        1  8 0.35    0.6384  0.061 2.16    0.0152  0.0152
        2 14 1.70    0.9553  0.139 1.71    0.0437  0.0348
    Total                    0.2000                0.0500
```

And we recognize key values from our example ...

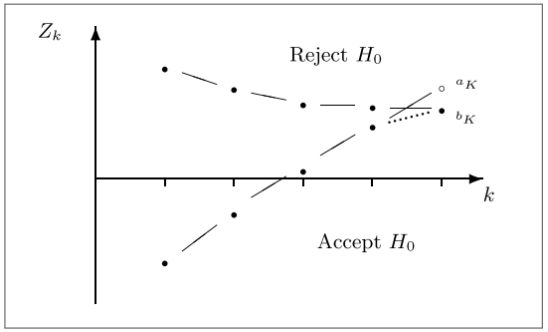
Over-running

If we never decide to stop at interim analysis, we can end-up observing more information at the last analysis ($k = K$) than expected. This is called **over-running**.

Problem:
In that case, solving the equations (as done above) will lead to $a_K > b_K$ instead of the desired $a_K = b_K$.

Solution:
Compute b_K as usual, to control the type-I error at α, but simply define $a_K = b_K$. In that case, the power attained under $\theta = \delta$ will be greater than the planned $1 - \beta$.

Over-running graphically



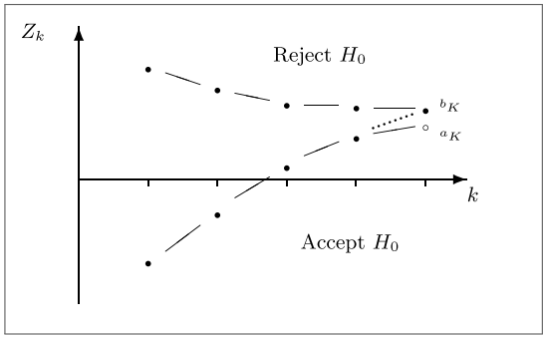
Under-running

If we never decide to stop at interim analysis, we can end-up observing less information at the last analysis ($k = K$) than expected. This is called **under-running**.

Problem:
In that case, solving the equations (as done above) will lead to $a_K < b_K$ instead of the desired $a_K = b_K$.

Solution:
Compute b_K as usual, to control the type-I error at α, but simply define $a_K = b_K$. In that case, the power attained under $\theta = \delta$ will be **smaller** than the planned $1 - \beta$.

Under-running graphically



Outline

New case study: Carinoma example (log-rank)

- Plan a group-sequential trial
- Analyze the data (reconstructed)

(Example from Jennison & Turnbull (2013))

Carinoma example: planning (first considerations)

Assume that we want to plan a randomized clinical trial with:

- Survival data, with death as main outcome
- θ= log hazard ratio to compare the two treatments, estimated via a (prespecified) adjusted Cox model (common practice).
- $\alpha = 2.5\%$ (most common, for one-sided tests)
- $1 - \beta = 80\%$
- $\delta = 0.5$ (expected log HR)
- usual randomization 1:1
- one-sided test $(H_0 : \theta \leq 0)$

Schoenfeld formula

Textbook formula for sample sizes/ power calculation for (Cox) proportional-hazards models in the context of nonsequential randomized clinical trials (Schoenfeld, 1983)

$$\#\text{observed events} = \frac{(z_\alpha + z_\beta)^2}{p_0 \cdot p_1 \cdot \{\log(\text{HR})\}^2} = \frac{\mathcal{I}_f}{p_0 \cdot p_1}$$

where p_0 and p_1 are the proportions of patients randomized to the treatment and control groups (often $=1/2$).

The number of observed events depends on both the sample size and the length of follow up.

It also works for:

- log-rank test (=score test in Cox model)
- Fine-Gray subdistribution hazard (Latouche & Porcher, 2007)

Exercise: part 1 (fixed design)

1. How much information $\mathcal{I}_f$ do we need?
2. How many observed events do we need to observe?
3. Check that this corresponds to including $n = 177$ given that:
   - we expect survival time to be approx. exponentially distributed with rate $\lambda_0 = 0.00167$ and $\lambda_1 = \lambda_0 \exp(-0.5)$ (here the time unit is day)
   - we expect to include the patients "uniformly" during an accrual period of 1735 days i.e. time from study start to inclusion $\sim \mathcal{U}([0, 1735])$.
   - we plan to stop the study, and thus stop the follow-up of all patients, approx. 1994 day (i.e. $\approx$ 5.5 years) after study start.
   - no dropout is expected

   You can proceed by simulation.

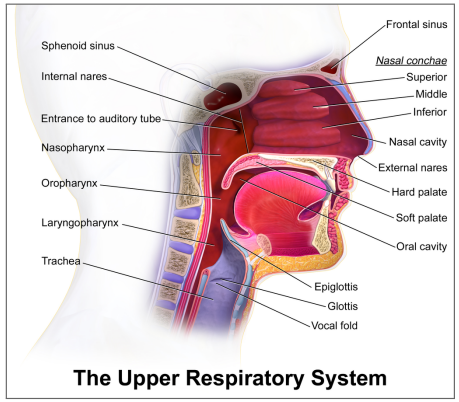
Exercise: part 2 (group-sequential design)

4. Using the `gsDesign` package, check that using $\rho = 2$, non-binding futility boundaries and $K = 5$, the inflation factor is $R = 1.133$, suggesting to include $n = 200$ subjects.

Note: here $K = 5$ such that the data that can be studied by a monitoring committee meeting at dates approx. 2, 3, 4 and 6 years after study start (accounting for a delay, say a few months, dues to data management etc...).

Carcinoma trial data

Data from a clinical trial conducted to investigate treatments of carcinoma of the oropharynx (in 1968-1973).



The Upper Respiratory System

Carcinoma is a cancer that begins in a tissue that lines the inner or outer surfaces of the body.

(Data from Kalbfleisch and Prentice, 2002, Appendix II)

Carcinoma data: baseline variables

- institution ($\approx$ hospitals, `inst`): 1,2,3,4,5 or 6.
- gender (`sex`): 1=male, 2=female
- initial condition (`cond`): 1= no disability, 2= restricted work, 3= require assistance with self care, 4= bed confined
- T-staging (`T`): 1=primary tumor $\leq$ 2 cm in largest diameter, 2=primary tumor 2 to 4 cm in largest diameter, minimal infiltration in depth, 3=primary tumor $>$ 4 cm, 4=massive invasive tumor.
- N-staging (`N`): 0= no clinical evidence of node metastases, 1= single positive node $\leq$ 3cm in diameter, not fixed, 2= single positive node $>$ 3cm in diameter, not fixed, 3=multiple positive nodes or fixed positive nodes
- tumor site (`site`): 1= fausial arch, 2= tonsillar fossa, 4= pharyngeal tongue
- treatment (`trt`): 0= "new", 1= "standard"
- `time`: time from inclusion to death or end of follow-up (days)
- `status`: 1=death, 0= alive at end of follow-up (censored) of follow-up (days)
- `dateEntry`: date of inclusion

(Kalbfleisch and Prentice, 2002, Appendix II)

Exercise: part 3 (analysis group-sequential data)

We assume to have prespecified to analyze the data by adjusting on these variables: institution (`inst`), gender (`sex`), initial condition (`cond`), T-staging (`T`), N-staging (`N`) and tumor site (`site`), to test for the treatment effect (`trt`). We further assume to have pre-specified the use a Cox model with baseline hazard stratified by institution.

We aim to show that the new treatment (`trt`=0) leads to a lower hazard rate that the standard treatment (`trt`=1).

5. What does suggest the 1st interim analysis? Why?
   (data Carcinoma1.csv)

5. Check the difference between the data available at the 1st and 2nd analysis (sample size, follow-up). Does it make sense? What does suggest the 2nd interim analysis? Why?
   (data Carcinoma2.csv)
6. Assume that we continue to the 3rd interim analysis. What does it suggest? Why?
   (data Carcinoma3.csv)
7. Assume that we continue to the 4th interim analysis. What does it suggest? Why?
   (data Carcinoma4.csv)
8. Assume that we continue to the 5th and final analysis. What does it suggest? Why?
   (data Carcinoma5.csv)

Outline

Inference after stopping

"the Food and Drug Administration and European Medicines Agency guidelines E9, 'Statistical principles for clinical trials', recommend presenting confidence intervals for treatment effects, and it is common to present a p-value for testing $H_0 : \theta \leq 0$."

(Hampson & Jennison, JRRS-B, 2013)

p-value

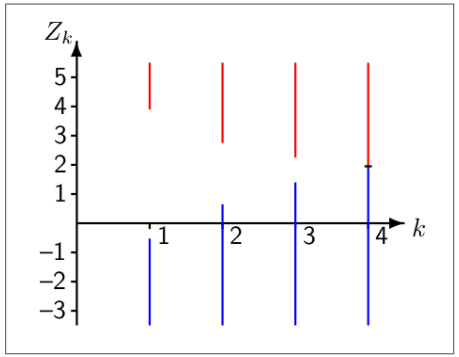
How should we compute a p-value when the GS trial terminates?

Reminder: the p-value is the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct.

What could "*at least as extreme*" mean here?

p-value: the sampe space

Unlike for fixed trials, we do not only observe one variable Z but a pair (k, Z_k). The sample space is:



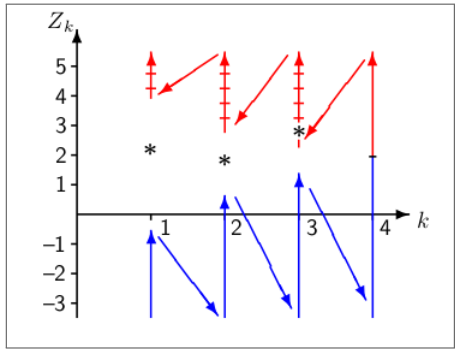
(not just $\mathbb{R}$, unlike when we observe only Z with a fixed test)

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

p-value: stagewise ordering the sampe space

For $H_0 : \theta \leq 0$ on observing (k^*, z^*), we use:

$$\text{p-value} = P_{\theta=0}\{\text{Obtain } (k, z) \text{ at least as extreme as } (k^*, z^*)\}$$



when using stagewise ordering as depicted above.

Confidence interval

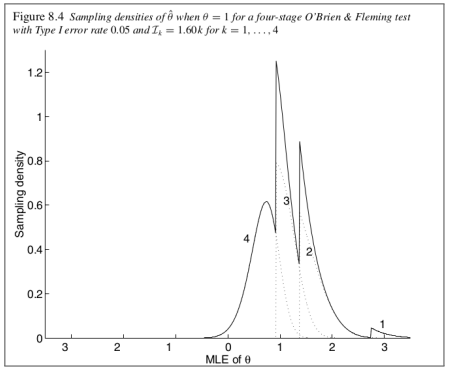
"it is straightforward to create a $100(1-\alpha)\%$ confidence set for θ by inverting a family of hypothesis tests for each possible θ-value." (Hampson & Jennison, JRRS-B, 2013)

Roughly speaking, the confidence interval is the set of values θ which do not lead to p-values smaller than α.

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Unbiased pointwise estimate

Because of the possibility to stop the trial early when the treatment effect looks either "very good" or "very bad", the estimate $\widehat{\theta}_k$ obtained from the available data after stopping the trial will generally be biased and "too extreme" (and a weird, non-normal, distribution) .



(Jennison, C. and Turnbull, B. W. (2000) *Group Sequential Methods with Applications to Clinical Trials.*)

It is generally difficult (possible?) to obtained an unbiased estimator wi survival data (as the information is not predictable...).

Median unbiased estimator

Instead of an usual unbiased pointwise estimate, it is possible to report a median unbiased estimator.

The interpretation is that the median unbiased estimator underestimates just as often as it overestimates $(\text{median}(\widehat{\theta}) = \theta)$.

This can be computed as the lower (or upper) limit of a 50% confidence interval.

(See e.g. Whitehead, *The Design and Analysis of Sequential Clinical Trials* (1997), section 5.2)

UNIVERSITY OF COPENHAGEN DEPARTMENT OF BIOSTATISTICS

Further (important) topics

- Pipeline data

  E.g. Data are locked before each interim analysis. Time passes as data are cleaned, the DMC meets, and –at one analysis– the DMC recommends to the Steering Committee that the trial be stopped. When stopping actually happens, more events will have occurred.

  How should the additional data be analyzed?

  (Hampson & Jennison, JRSS-B, 2013)

- Hierarchical testing of primary and secondary endpoint

  E.g. progression-free survival and overall survival in oncology.

  (Glimm & Bretz, SiM, 2009)

- Simultaneous testing of co-primary endpoints or subgroups

  E.g. recurrence of atrial fibrillation and quality of life.

  E.g. subgroup "biomarker positive" and broader population.

  (Ye, Li & Yao, SiM, 2011)

A few take home messages

- Group Sequential Trials are increasingly common and often advantageous.
- Specific must be used with survival data due to the unpredictable information levels.
- Both methods and software are available and expanding fast.
- Still some (many) open problems: the current theory does not fully cover the current needs of medical research.