Key Multiplicity Issues in Clinical Trials (Part I)

Alex Dmitrienko (Mediana Inc)
admitrienko@medianainc.com

Introduction

Multiplicity in clinical trials

Drug development challenges

Drug development costs have been increasing steadily

More sophisticated trial designs are used to improve efficiency of drug development programs

Example: Designs with increasingly more complex objectives

Multiplicity issues

Multiple objectives induce multiplicity and increase false-positive rates

Multiplicity issues in clinical trials

Multiplicity adjustment

Multiplicity adjustment methods are required in trials with multiple objectives

Regulatory guidance documents

U.S. Food and Drug Administration (FDA)

European Medicines Agency (EMA)

Multiplicity issues in clinical trials

FDA guidance

Draft guidance on multiplicity issues in clinical trials (January 2017)

EMA guidance

Points to consider on multiplicity issues in clinical trials (September 2002)

Draft guideline on multiplicity issues in clinical trials (April 2017)

What is new in this course?

New topics

Multiple updates throughout the modules included in the course and a new module on power calculations

Updated case studies

New guidance documents

Detailed discussion of FDA and EMA guidance documents

New software tools

Updated software implementation sections (using Mediana package)

General outline

Part I

Traditional multiplicity problems

Clinical trials with a single source of multiplicity

Part II

Advanced multiplicity problems

Clinical trials with multiple sources of multiplicity

Part I: Traditional multiplicity problems

Single family of null hypotheses

$$H_1, \ldots, H_m$$

Trials with a single source of multiplicity

Single source of multiplicity

Multiple endpoints

Better characterize efficacy/safety of a new treatment

Multiple dose-placebo comparisons

Evaluate the dose-response relationship

Multiple patient populations

Evaluate efficacy/safety in different groups of patients and help develop targeted agents

Part II: Advanced multiplicity problems

Multiple families of null hypotheses

Family 1 $H_1, \ldots, H_{k_1}$

$\ldots$

Family m $H_{k_{m-1}+1}, \ldots, H_{k_m}$

Trials with multiple sources of multiplicity

Multiple sources of multiplicity

Multiple primary and secondary endpoints

Evaluate primary objective (primary endpoints) and provide useful supportive information (secondary endpoints)

Multiple endpoints and multiple dose-control comparisons

Evaluate efficacy/safety at different dose levels

Multiple patient populations and multiple dose-control comparisons

Evaluate efficacy/safety at different dose levels in different groups of patients

Outline

Module A

Clinical trial examples

Module B

Key concepts (inferential goals, error rate definitions, classification of multiple testing procedures)

Module C

Nonparametric and semiparametric multiple testing procedures I (data-driven hypothesis ordering)

Outline

Module D

Nonparametric multiple testing procedures II (pre-specified hypothesis ordering)

Module E

Parametric multiple testing procedures

Module F

Simultaneous confidence intervals

Module G

Power calculations

Books

Analysis of Clinical Trials Using SAS

Edited by Alex Dmitrienko (Mediana) and Gary Koch (UNC-Chapel Hill)

Published by SAS Press in 2017

Chapter 5: Multiplicity adjustment methods

Introduction to multiplicity problems arising in clinical trials, popular multiple testing procedures and gatekeeping procedures

Books

Multiple Testing Problems in Pharmaceutical Statistics

Edited by Alex Dmitrienko (Eli Lilly), Ajit Tamhane (Northwestern University), Frank Bretz (Novartis, Hannover Medical School)

Published by Chapman and Hall/CRC Press in 2009

Comprehensive summary of methodological, regulatory and practical issues related to multiplicity problems in pre-clinical research and clinical trials

Review papers

Recent review papers and tutorials

Dmitrienko, D'Agostino and Huque. (2013). Key multiplicity issues in clinical drug development.

Dmitrienko and D'Agostino. (2013). Tutorial in Biostatistics: Traditional multiplicity adjustment methods in clinical trials.

Alosh, Bretz and Huque (2014). Advanced multiplicity adjustment methods in clinical trials.

Web site

Instant Training web site

http://sprmm.com

Supplementary materials

Presentation slides

SAS and R code

References

Phase II and Phase III trials

Multiplicity problems

Focus on multiplicity problems in confirmatory Phase III trials but general approaches can also be applied to Phase II trials

EMA guidance (EMA, 2017)

"The main scope is to provide guidance on the confirmatory conclusions which are usually based on the results from pivotal Phase III trials and, to a lesser extent, on Phase II studies"

Conventions

Multiple tests and procedures

Multiple testing procedure is a tool for testing multiple null hypotheses

Multiple test is a tool for testing a single null hypothesis

One-sided and two-sided testing

Testing problems, unless otherwise stated, are defined as one-sided problems

Module A
Clinical trial examples

Module A outline

A1. Clinical trial examples

Clinical trials with multiple endpoints, multiple doses and multiple patient populations to motivate key concepts

A1. Clinical trial examples

Multiple endpoints

Example 1: Prostate cancer trial

Example 2: Alzheimer's disease trial

Example 3: Fracture healing trial

Multiple doses

Example 4: Type 2 diabetes trial

Multiple populations

Example 5: Non-small-cell lung cancer trial

Example 1

Clinical trial with multiple endpoints

Example 1: Prostate cancer trial

Objective

Evaluate the effects of an experimental treatment (enzalutamide) on progression-free and overall survival (Beer at al., 2014)

Design

Experimental treatment versus placebo

Clinical trial with multiple endpoints

Example 1: Prostate cancer trial

Two primary endpoints

Endpoint 1: Radiographic progression-free survival (rPFS)

Endpoint 2: Overall survival (OS)

Overall analysis

At least one endpoint must be significant

Example 2

Clinical trial with multiple endpoints

Example 2: Alzheimer's disease trial

Objective

Evaluate the effects of a treatment (rivastigmine) on cognition and global changes in patients with mild to moderate Alzheimer's disease (IDEAL study, Winblad et al., 2007)

Design

Treatment versus placebo

Clinical trial with multiple endpoints

Example 2: Alzheimer's disease trial

Two co-primary endpoints

Endpoint 1: Cognition endpoint (Alzheimer's Disease Assessment Scale-Cognitive subscale)

Endpoint 2: Clinical global scale (Alzheimer's Disease Cooperative Study-Clinical Global Impression of Change)

Overall analysis

Both endpoints must be significant

Example 3

Clinical trial with multiple endpoints

Example 3: Fracture healing trial

Objective

Evaluate treatment effect on functional recovery in patients with osteoporosis

Design

Treatment versus placebo

Clinical trial with multiple endpoints

Example 3: Fracture healing trial

Three endpoints

Endpoint 1: Timed up-and-go test

Endpoint 2: Six-minute walking distance test

Endpoint 3: Pain score

Overall analysis

Overall treatment effect on all endpoints must be significant

Example 4

Clinical trial with multiple doses

Example 4: Type 2 diabetes trial

Objective

Evaluate the efficacy of an experimental treatment (saxagliptin) in treatment-naive patients with Type 2 diabetes (Rosenstock et al., 2009)

Primary endpoint

HbA1c change from baseline to Week 24

Design

Three dose groups versus placebo

Example 5

Clinical trial with multiple patient populations

Example 5: Non-small-cell lung cancer trial

Objective

Evaluate the effects of a treatment (erlotinib) in advanced non-small-cell lung cancer (SATURN trial, Cappuzzo et al., 2010)

Primary endpoint

Progression-free survival (PFS)

Design

Treatment versus placebo

Clinical trial with multiple patient populations

Example 5: Non-small-cell lung cancer trial

Tailored therapy approach is implemented in this trial

Two patient populations

General population

Subpopulation of patients with EGFR (epidermal growth factor receptor) immunohistochemistry-positive tumors

Multiple patient populations

EMA guidance (EMA, 2017)

"Multiplicity issues arise if study success is defined by the demonstration of a beneficial effect of the treatment in the whole study population or in a pre-defined subgroup (or in one of several subgroups)"

Module B
Key concepts

Module B outline

B1. Inferential goals

At-least-one testing, all-or-none testing and global testing

B2. Error rate definitions for multiple testing procedures

Familywise error rate

B3. Selection of multiple testing procedures

Guidelines for selecting multiple testing procedures

Section B1
Inferential goals

Inferential goals

Multiple testing problem

Inferences used in a multiple testing problem depend on the inferential goal

Three inferential goals

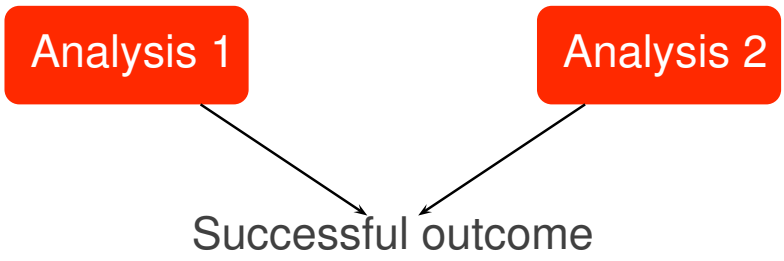
Individual analyses separately lead to a successful outcome (at-least-one procedures, also known as multiple testing procedures)

Individual analyses jointly lead to a successful outcome (all-or-none procedures)

Overall analysis leads to a successful outcome (global procedures)

At-least-one procedures

Each analysis is independently clinically relevant



Each analysis independently provides a proof of efficacy

The trial's outcome is declared positive if at least one analysis is significant

At-least-one procedures

Union-intersection problem

Problem is known as the union-intersection problem and requires a multiplicity adjustment

At-least-one procedures (multiple testing procedures) will be discussed in this course

Examples

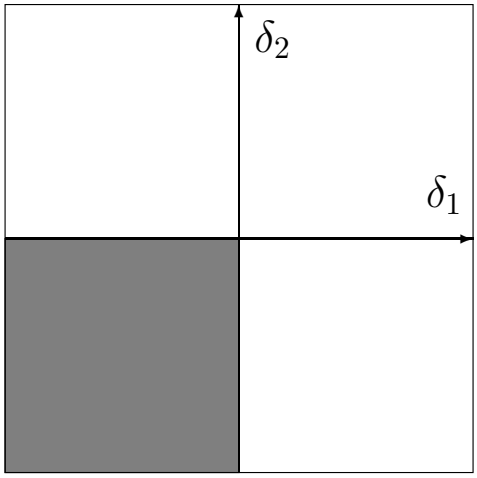
Example 1: Prostate cancer trial

Example 4: Type 2 diabetes trial

Example 5: Non-small-cell lung cancer trial

Example 1: Prostate cancer trial

Global null hypothesis (shaded region)

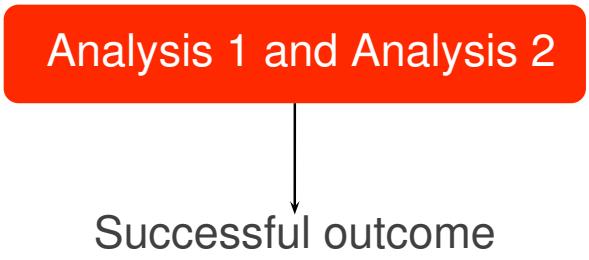


δ_1, True treatment difference (overall survival)

δ_2, True treatment difference (progression-free survival)

All-or-none procedures

All analyses must show benefit



The trial's outcome is positive if all analyses produce a significant outcome

All-or-none procedures

Intersection-union problem

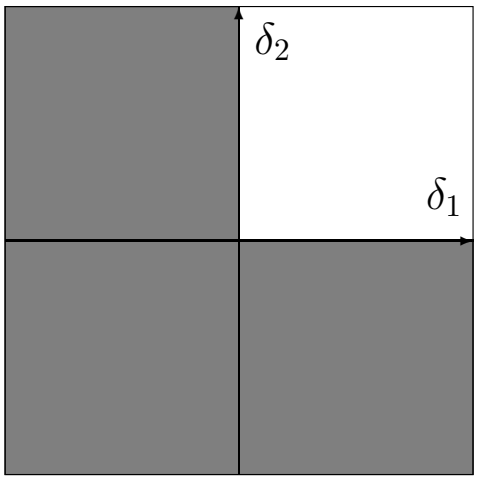
Problem is known as the intersection-union problem and does not require a multiplicity adjustment

Example

Example 2: Alzheimer's disease trial with co-primary endpoints

Example 2: Alzheimer's disease trial

Global null hypothesis (shaded region)



δ_1, True treatment difference (cognition endpoint)

δ_2, True treatment difference (clinical global scale)

All-or-none procedures

Notation

$H_1, \ldots, H_m$, null hypotheses

$p_1, \ldots, p_m$, p-values

α, Type I error rate, e.g., $\alpha = 0.025$

Intersection-union problem

All null hypotheses are rejected if $p_1 \leq \alpha$, …, $p_m \leq \alpha$

This problem will not be discussed further in this course

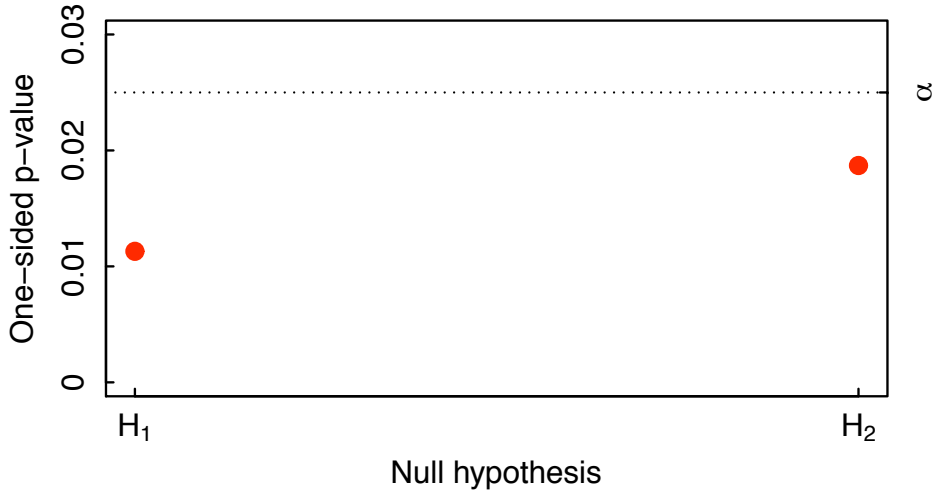
All-or-none procedures

FDA guidance (FDA, 2017)

"There have been suggestions that the statistical testing criteria for each co-primary endpoint could be relaxed (e.g., testing at an alpha of 0.06 or 0.07)... Relaxation of alpha is generally not acceptable because doing so would undermine the assurance of an effect on each disease aspect considered essential to showing that the drug is effective in support of approval."

Example 2: Alzheimer's disease trial

Decision rule for $\alpha = 0.025$



One-sided p-values: $p_1 = 0.0113$ and $p_2 = 0.0187$

All-or-none procedure rejects H_1 and H_2

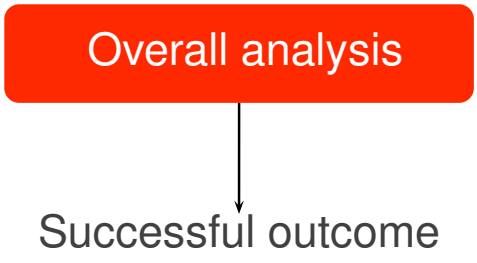
All-or-none procedures

FDA guidance (FDA, 2017)

"The use of two or more endpoints for which demonstration of an effect on each is needed to support regulatory approval (called co-primary endpoints) increases the Type II error rate and decreases study power"

Global procedures

Individual analyses are components of an overall analysis



Treatment effect is defined in terms of a combination of individual effects across multiple analyses

The trial's outcome is positive if the overall effect is significant

Global procedures

Example

Example 3: Fracture healing trial

Individual analyses (timed up-and-go test, six-minute walking distance test and pain score) are components of an overall analysis

Global testing

Very rarely used in Phase III trials but may be used in Phase II trials (will not be discussed further in this course)

See *Multiple Testing Problems in Pharmaceutical Statistics* (Chapter 4) for more information

Section B2
Error rate definitions

Error rate definitions

Multiple testing procedures

To choose an appropriate testing method, it is critical to select the definition of correct and incorrect decisions

Preferred definition

Familywise error rate (FWER)

Other definitions

Generalized familywise error rate and false discovery rate are not used in confirmatory clinical trials

Importance of addressing multiplicity issues

EMA guidance (EMA, 2017)

"A clinical study that requires no adjustment of the significance level of elementary hypothesis tests (i.e. single statistical tests on one parameter only) is one that consists of two treatment groups, which uses a single primary variable, and has a confirmatory statistical strategy that pre-specifies just one single null hypothesis relating to the primary variable and no interim analysis"

Familywise error rate

Notation

$H_1, \ldots, H_m$, Null hypotheses

Definition

Familywise error rate is controlled in the strong sense at α level if the probability of incorrectly rejecting at least one true null hypothesis is $\leq \alpha$ regardless of which and how many other hypotheses are true

Familywise error rate

Example 4: Type 2 diabetes trial

H_1, H_2 and H_3, null hypotheses

Consider all combinations of true null hypotheses and show that Type I error rate $\leq \alpha$ for any combination

Example: Suppose that H_1 and H_2 are true and H_3 is false, then

$$P\left(\text{Reject } H_1 \text{ or } H_2\right) \leq \alpha$$

Familywise error rate control

Properties

This definition enables clinical trial sponsors to make specific claims

Regulatory position

Strong FWER control for primary objectives is mandated by regulators in all confirmatory clinical trials

Multiple testing procedures

Procedures introduced in this course provide FWER control in the strong sense

Section B3
Selection of multiple testing procedures

Selection of multiple testing procedures

1. Define hypothesis testing problem

2. Define relationships among null hypotheses

It is important to account for trial-specific information

Clinical information: Logical restrictions, e.g., are the null hypotheses ordered?

Statistical information: Distributional information, e.g., is the joint distribution of the hypothesis test statistics known?

Selection of multiple testing procedures

3. Define candidate multiple testing procedures

Procedures consistent with requirements defined in Step 2

4. Select an optimal multiple testing procedure

Most powerful procedure consistent with requirements defined in Step 2

Step 1: Hypothesis testing problem

Notation

$H_1, \ldots, H_m$, Null hypotheses of interest

α, Familywise error rate, e.g., one-sided $\alpha = 0.025$

Assumptions

Null hypotheses are equally important (extensions to the case of unequally important null hypotheses are easily constructed)

Step 1: Hypothesis testing problem

Hypothesis test statistics

$t_1, \ldots, t_m$, Statistics used for testing the hypotheses $H_1, \ldots, H_m$

$t_{(1)} > \ldots > t_{(m)}$, Ordered test statistics

Hypothesis p-values

$p_1, \ldots, p_m$, Original p-values

$p_{(1)} < \ldots < p_{(m)}$, Ordered p-values

Step 2: Classification schemes

Clinical information

Classification scheme based on clinically relevant logical relationships among the null hypotheses

Single-step and stepwise procedures

Statistical information

Classification scheme based on distributional relationships, i.e., the joint distribution of the hypothesis test statistics

Nonparametric, semiparametric and fully parametric procedures

Step 2: Logical restrictions

Basic single-step testing approach

Null hypotheses are tested simultaneously or in a single step

Clinically meaningful relationships among null hypotheses are not taken into account

Stepwise testing approach

Null hypotheses are ordered using clinical importance or using significance of test statistics/p-values

Step 2: Logical restrictions

Single-step procedures

Null hypotheses are tested in a single step, i.e., each null hypothesis is rejected independently of other null hypotheses

H_1 H_2 $\cdots$ H_m

Example: Bonferroni and Dunnett procedures

Step 2: Logical restrictions

Pre-specified testing sequence

Null hypotheses are ordered at the design stage to reflect clinical importance or probability of success for associated objectives

Example 4: Type 2 diabetes trial

Strong evidence of a positive dose-response relationship

H_1, H_2, H_3 are tested sequentially beginning with the highest dose

Step 2: Logical restrictions

Pre-specified testing sequence

Null hypotheses are ordered and tested sequentially beginning with H_1

H_1 H_2 $\cdots$ H_m

Example: Fixed-sequence, fallback and chain procedures

Step 2: Logical restrictions

Data-driven testing sequence

Null hypotheses are not ordered at the design stage

Example 4: Type 2 diabetes trial

Difficult to assume a positive dose-response relationship

H_1, H_2, H_3 are tested in the order determined by significance of test statistics

Step 2: Logical restrictions

Data-driven testing sequence

Null hypotheses are tested in the order determined by significance of test statistics

H_2 H_m $\ldots$ H_1

$p_2 < p_m < \ldots < p_1$, Ordered p-values

Example: Holm, Hommel, Hochberg and step-down Dunnett procedures

Step 2: Distributional relationships

Three classes of procedures

Nonparametric	Semiparametric	Parametric
Test 1	Test 1	Test 1
Test 2	Test 2	Test 2

Step 2: Distributional relationships

Nonparametric procedures

Based on univariate p-values and impose no distributional assumptions

Bonferroni, Holm, fixed-sequence, fallback and chain procedures

Properties

Very popular due to their simplicity

Tend to perform poorly with too many null hypotheses or strongly correlated hypothesis test statistics

Step 2: Distributional relationships

Semiparametric procedures

Based on univariate p-values and impose some distributional assumptions (multivariate normal distribution with non-negative correlations)

Hochberg and Hommel procedures

Properties

More powerful than nonparametric procedures

Step 2: Distributional relationships

Parametric procedures

Based on multivariate p-values computed from a pre-specified joint distribution (multivariate normal or t distribution)

Single-step, step-down and step-up Dunnett procedures

Properties

More powerful than nonparametric and semiparametric procedures

Step 2: Distributional relationships

Resampling-based procedures

Do not make distributional assumptions and approximate true joint distribution of test statistics using bootstrap or permutation methods

Not used in Phase III trials but may be used in early-phase trials (will not be discussed further in this course)

FDA guidance (FDA, 2017)

"Resampling methods are not recommended as primary analysis methods for adequate and well-controlled trials in drug development"

Example 1: Prostate cancer trial

Two primary endpoints

Endpoint 1: Radiographic progression-free survival (rPFS)

Endpoint 2: Overall survival (OS)

What class of procedures can be used in this example?

Nonparametric, semiparametric or parametric procedures?

Example 1: Prostate cancer trial

Important considerations

Correlation information can be taken into account only if it is known at the trial's design stage

Correlations can be estimated but use of sample correlations in multiple testing procedures may result in FWER inflation

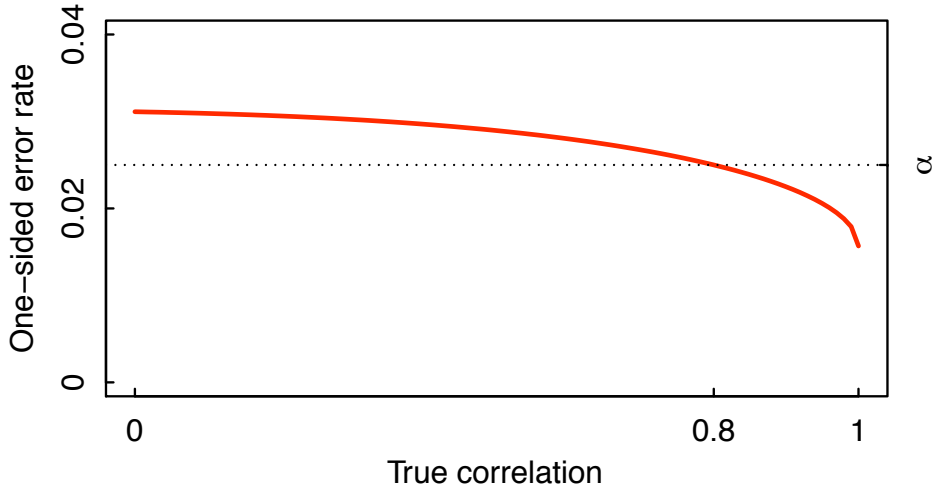
Example 1: Prostate cancer trial

Important considerations

"The use of correlation for alpha allocation may be challenged when the trial is not prospectively planned with a sample size to detect a prespecified treatment effect in the subset, in which case the sample size fraction is unknown at the trial design stage and is not determined until the end of the study " (Wang, O'Neill and Hung, 2007)

Example 1: Prostate cancer trial

Actual Type I error rate with $\alpha = 0.025$



Based on a parametric procedure (endpoint test statistics follow a bivariate normal distribution with sample correlation $\rho = 0.8$)

Example 1: Prostate cancer trial

Parametric procedure

A parametric procedure may be used if the correlation between the endpoints is known

Semiparametric procedure

A semiparametric procedure may be used if the correlation between the endpoints is known to be non-negative

Noparametric procedure

A nonparametric procedure will need to be used if the correlation between the endpoints is unknown

Step 3: Candidate procedures

Candidate multiple testing procedures

Identify a candidate set of procedures consistent with the clinical and statistical requirements defined in Step 2

Single-step procedures

Nonparametric procedures

Bonferroni (Module C)

Semiparametric procedures

Simes and Šidák (Module C)

Parametric procedures

Dunnett (Module E)

Stepwise procedures: Pre-specified testing sequence

Nonparametric procedures

Fixed-sequence, fallback and a class of chain procedures (Module D)

Semiparametric procedures

No procedures

Parametric procedures

Parametric fallback and parametric chain (will not be discussed)

Stepwise procedures: Data-driven testing sequence

Nonparametric procedures

Holm (Module C)

Semiparametric procedures

Hochberg and Hommel (Module C)

Parametric procedures

Step-down Dunnett and step-up Dunnett (Module E)

Step 4: Optimal multiple testing procedure

Stepwise procedures

Stepwise procedures are more powerful than single-step procedures

Stepwise procedures with a data-driven testing sequence are most flexible

Semiparametric procedures

Strike a balance between nonparametric and parametric approaches

More powerful than nonparametric and more robust than parametric procedures

Module C
Nonparametric and semiparametric multiple testing procedures I

Module C outline

C1. Basic procedures

Bonferroni and Simes

C2. Closure principle

Powerful tool for constructing stepwise multiple testing procedures

C3. Stepwise procedures with a data-driven testing sequence

Examples: Holm procedure (Holm, 1979), Hommel procedure (Hommel, 1988) and Hochberg procedure (Hochberg, 1988)

Section C1
Basic multiple testing procedures

Basic procedures

Single-step procedure

Bonferroni procedure (nonparametric procedure)

Global procedure

Simes global test (semiparametric test)

Note

Introduced to provide a foundation for more powerful multiple testing procedures

Bonferroni procedure

A little bit of history

Bonferroni procedure was most likely proposed by Sir Ronald Fisher

Based on the Bonferroni inequality named after Carlo Emilio Bonferroni (1892–1960) but Bonferroni's research actually focused on extensions of this inequality

Bonferroni inequality goes back to the work of George Boole (1815–1864)

Bonferroni procedure

Decision rule

Bonferroni procedure rejects H_i if $p_i \leq \alpha/m$

Procedure controls FWER for any joint distribution of test statistics due to Bonferroni inequality

$$\begin{aligned} & P(p_1 \leq \alpha/m \text{ or } \ldots \text{ or } p_m \leq \alpha/m) \\ \leq \quad & \sum_{i=1}^{m} P(p_i \leq \alpha/m) = \sum_{i=1}^{m} \alpha/m = \alpha \end{aligned}$$

since p_i follows Uniform $(0, 1)$ distribution, $i = 1, \ldots, m$, under the global null hypothesis

Simes global test

Decision rule

Simes test (Simes, 1986) examines the global hypothesis of no treatment effect

$$H_I = \bigcap_{i=1}^{m} H_i$$

Simes global test rejects H_I if

$$p_{(i)} \leq i\alpha/m \text{ for at least one } i = 1, \ldots, m,$$

where $p_{(1)} < \ldots < p_{(m)}$ are ordered p-values

Example 4: Type 2 diabetes trial

Three dose-placebo comparisons

Comparison	P-value
Dose 1 vs Placebo (H_1)	$p_1 = 0.0111$
Dose 2 vs Placebo (H_2)	$p_2 = 0.0065$
Dose 3 vs Placebo (H_3)	$p_3 = 0.0293$

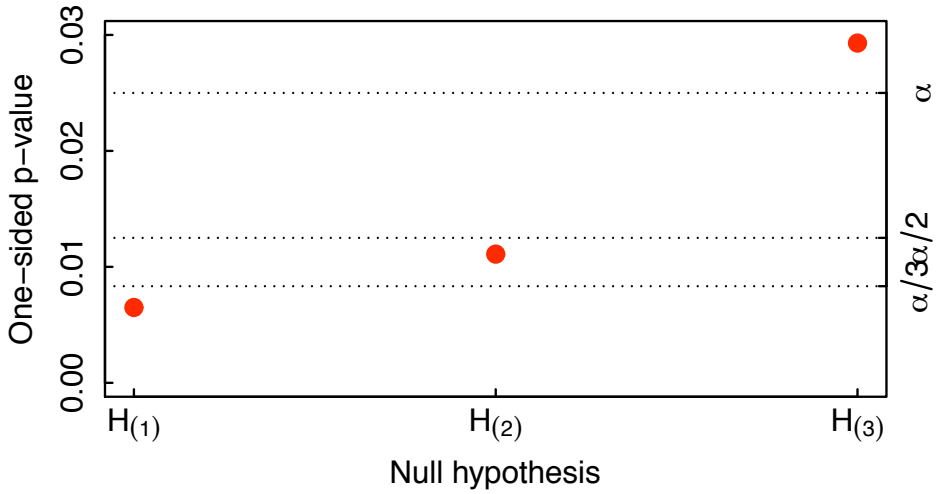
Evidence of treatment effect at Doses 1 and 2

$p_{(1)} = p_2 = 0.0065$, $p_{(2)} = p_1 = 0.0111$ and $p_{(3)} = p_3 = 0.0293$, ordered p-values

$H_{(1)}$, $H_{(2)}$ and $H_{(3)}$, ordered null hypotheses

Bonferroni and Simes procedures

Decision rules in Example 4 ($\alpha = 0.025$)



Bonferroni rejects $H_{(1)} = H_2$

Simes rejects the global null hypothesis

Type I error rate control

Bonferroni procedure

Bonferroni is conservative if the number of hypotheses is large or test statistics are strongly positively correlated

Simes global test

Simes controls Type I error rate for some joint distributions, e.g., if test statistics are independent or positively dependent

Simes may lead to Type I error rate inflation and its properties will be discussed in Section C3

Type I error rate control

Example

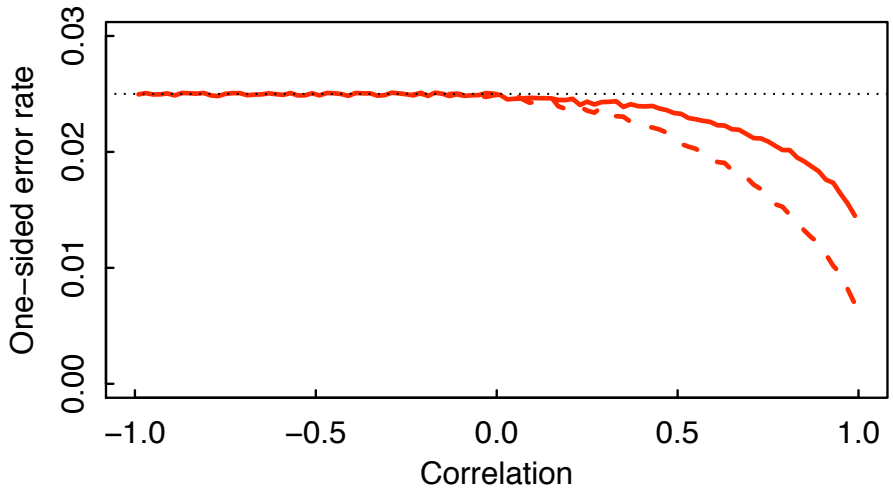
Clinical trial with $m = 2$ and $m = 5$ endpoints

Test statistics follow a multivariate normal distribution and are equally correlated
($-1 < \rho \leq 1$ if $m = 2$, $-1/4 < \rho \leq 1$ if $m = 5$)

$\alpha = 0.025$, Familywise error rate

Bonferroni procedure

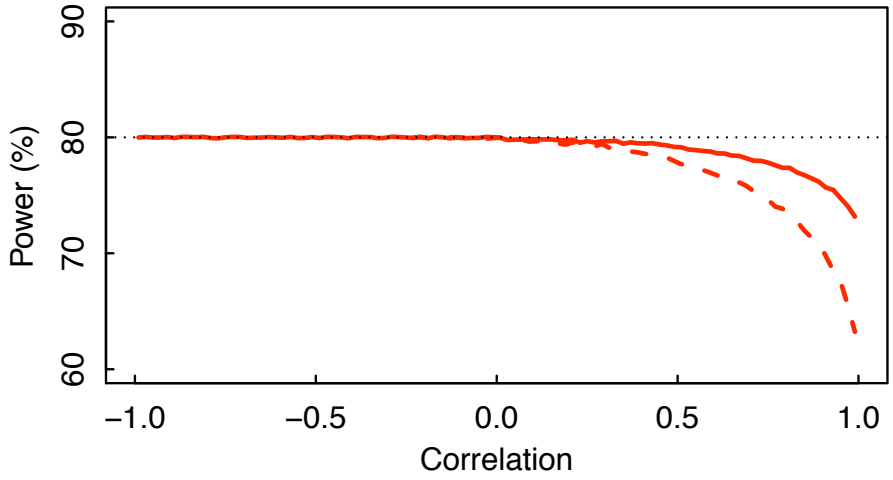
Actual Type I error rate with $\alpha = 0.025$



— 2 endpoints, - - - 5 endpoints

Bonferroni procedure

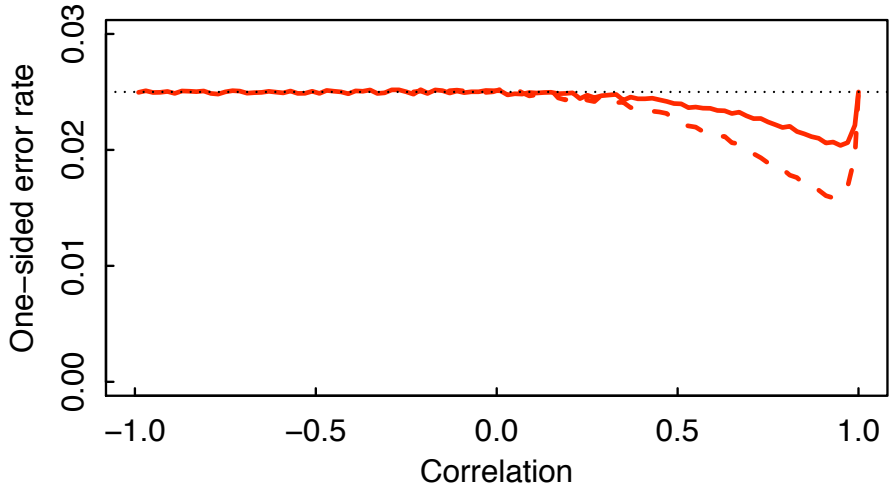
Power



— 2 endpoints, - - - 5 endpoints

Simes global test

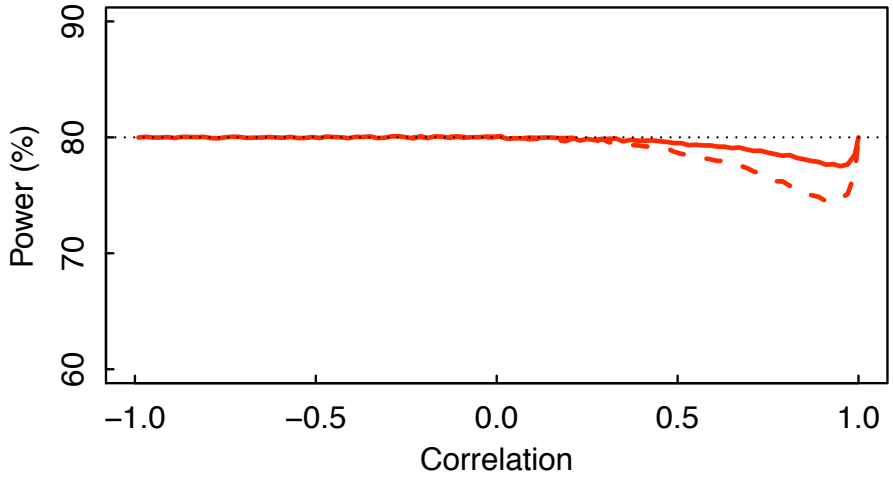
Actual Type I error rate with $\alpha = 0.025$



— 2 endpoints, - - - 5 endpoints

Simes global test

Power



— 2 endpoints, - - - 5 endpoints

Section C2
Closure principle

Closure principle

Closed testing procedures

Closure principle (Marcus, Peritz and Gabriel, 1976) is a powerful tool for building powerful multiple testing procedures (closed testing procedures)

Provides a foundation for virtually all multiple testing methods used in clinical trial applications

Closure principle

Example 1: Prostate cancer trial

$H_1: \ \delta_1 \leq 0$ (Overall survival)

$H_2: \ \delta_2 \leq 0$ (Radiographic progression-free survival)

One-sided p-values: $p_1 = 0.0102$ and $p_2 = 0.0181$

Closed testing procedure

Set up a procedure for testing H_1 and H_2 which controls FWER at α level

Example 1: Prostate cancer trial

Closed testing procedure

Define closed family of hypotheses which includes all intersections of H_1 and H_2, i.e., H_1, H_2 and $H_1 \cap H_2$

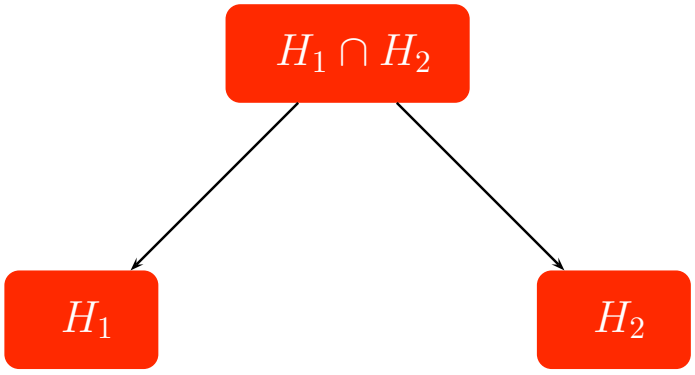
Establish implication relationships, i.e., $H_1 \cap H_2$ implies H_1 and H_2

Define α-level local tests for H_1, H_2 and $H_1 \cap H_2$

Reject a null hypothesis if all intersection hypotheses implying this hypothesis are rejected by local tests

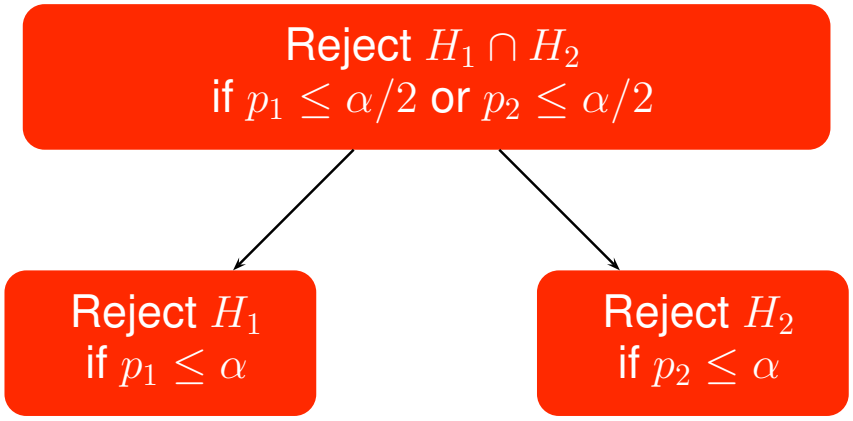
Example 1: Prostate cancer trial

Closed family



Example 1: Prostate cancer trial

Bonferroni-based closed testing procedure



Familywise error rate control

Closure principle

Closed testing procedures control FWER in the strong sense at α since α-level local tests are used for all intersection hypotheses

Example 1: Prostate cancer trial

Notation

$p_{(1)} = p_1 = 0.0102$, $p_{(2)} = p_2 = 0.0181$, ordered p-values

$H_{(1)}$, $H_{(2)}$, ordered null hypotheses

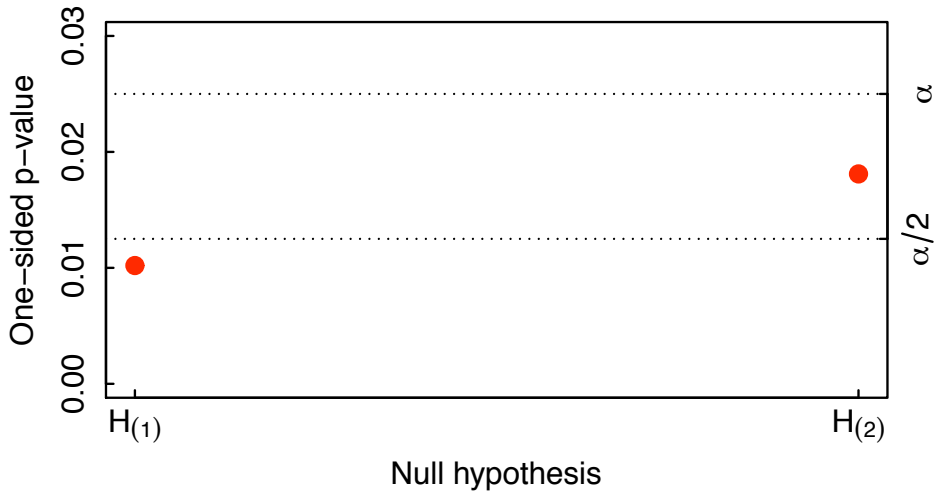
Bonferroni-based closed testing procedure (Holm procedure)

Decision rule

- Reject $H_{(1)}$ if $p_{(1)} \leq \alpha/2$
- Reject $H_{(2)}$ if $p_{(2)} \leq \alpha$ and $H_{(1)}$ is rejected

Bonferroni and Holm procedures

Decision rules in Example 1 ($\alpha = 0.025$)



Bonferroni procedure rejects $H_{(1)} = H_1$

Holm procedure rejects both null hypotheses

Summary

Familywise error rate control

Closed testing procedures protect FWER in the strong sense

Power

Closed testing procedures are more powerful than procedures they are derived from

For example, Bonferroni-based closed testing procedure (Holm procedure) always rejects as many null hypotheses as or more null hypotheses than Bonferroni procedure

Section C3
Stepwise procedures with a data-driven testing sequence

Data-driven testing sequence

Hypothesis testing problem

$H_1, \ldots, H_m$, Null hypotheses

$\alpha = 0.025$, Familywise error rate (one-sided)

Distributional relationships

Nonparametric procedures (Holm): No assumptions about the joint distribution of hypothesis test statistics

Semiparametric procedures (Hochberg and Hommel): Flexible distributional assumptions

Data-driven testing sequence

Logical relationships

Stepwise procedures with a data-driven testing sequence

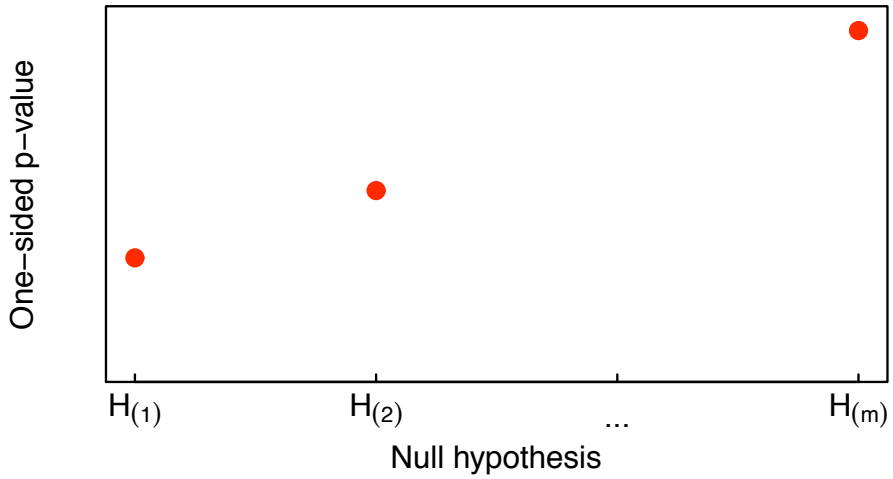
Two types of stepwise procedures

Step-down procedures (Holm):

Step-up procedures (Hochberg and Hommel)

Step-down procedures

Testing begins with the smallest p-value

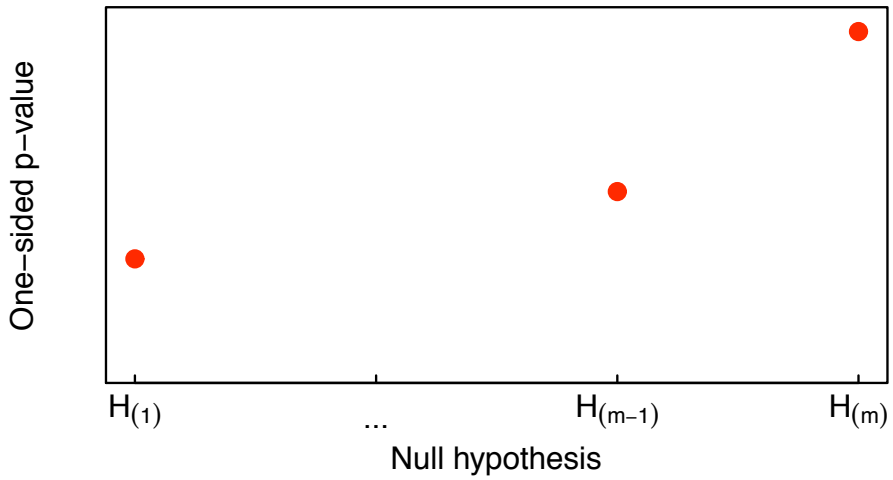


$p_{(1)} < \ldots < p_{(m)}$, ordered p-values

$H_{(1)}, \ldots, H_{(m)}$, ordered null hypothesis

Step-up procedures

Testing begins with the largest p-value



$p_{(1)} < \ldots < p_{(m)}$, ordered p-values

$H_{(1)}, \ldots, H_{(m)}$, ordered null hypothesis

Holm procedure

Holm procedure

General decision rules (step-down algorithm)

Step 1: If $p_{(1)} \leq \alpha/m$, reject $H_{(1)}$ and go to Step 2, otherwise accept all hypotheses and stop

Steps $i = 2, \ldots, m-1$: If $p_{(i)} \leq \alpha/(m-i+1)$, reject $H_{(i)}$ and go to Step $i+1$, otherwise accept $H_{(i)}, \ldots, H_{(m)}$ and stop

Step m: If $p_{(m)} \leq \alpha$, reject $H_{(m)}$, otherwise accept $H_{(m)}$

Holm procedure

Type I error rate

Holm procedure is a nonparametric procedure and controls FWER for any joint distribution of hypothesis test statistics

Power

Holm procedure is uniformly more powerful than Bonferroni procedure, i.e., it rejects all null hypotheses rejected by Bonferroni procedure and potentially more null hypotheses

Multiplicity-adjusted p-values

Multiplicity adjustments

Multiplicity adjustments are applied by adjusting the significance level downward or adjusting the p-value upward

Null hypothesis H_i is rejected if $p_i \leq \tilde{\alpha}_i$ or $\tilde{p}_i \leq \alpha$

It is generally more convenient to work with adjusted p-values since they can be used with any α level

Example 4: Type 2 diabetes trial

Raw p-values

Comparison	P-value
Dose 1 vs Placebo (H_1)	$p_1 = 0.0111$
Dose 2 vs Placebo (H_2)	$p_2 = 0.0065$
Dose 3 vs Placebo (H_3)	$p_3 = 0.0293$

Example 4: Type 2 diabetes trial

Multiplicity-adjusted p-values

Hypothesis	Adjusted p-values	
	Bonferroni procedure	Holm procedure
H_1	0.0333	0.0222
H_2	0.0195	0.0195
H_3	0.0879	0.0293

Bonferroni procedure rejects H_2, Holm procedure rejects H_1 and H_2 at $\alpha = 0.025$

Hommel procedure

Hommel procedure

Simes-based step-up procedure

Hommel procedure is a semiparametric procedure derived from Simes global test

Using the closure principle, each intersection hypothesis is tested using Simes global test

Hommel procedure is based on a rather complicated step-up algorithm

Hommel procedure

General decision rules (step-up algorithm)

Step 1: If $p_{(m)} > \alpha$, accept $H_{(m)}$ and go to Step 2, otherwise reject all null hypotheses and stop

Steps $i = 2, \ldots, m-1$: If $p_{(m-j+1)} > j\alpha/i$ for all $j = 1, \ldots, i$, accept $H_{(m-i+1)}$ and go to Step $i+1$, otherwise reject all remaining null hypotheses $H_{(j)}$ with $p_{(j)} \leq \alpha/(i-1)$ and stop

Step m: If $p_{(j)} > j\alpha/m$ for all $j = 1, \ldots, m$, accept $H_{(1)}$, otherwise reject $H_{(1)}$ if $p_{(1)} \leq \alpha/(m-1)$

Example 4: Type 2 diabetes trial

Scenario 2

Comparison	P-value
Dose 1 vs Placebo	0.0291
Dose 2 vs Placebo	0.0095
Dose 3 vs Placebo	0.0153

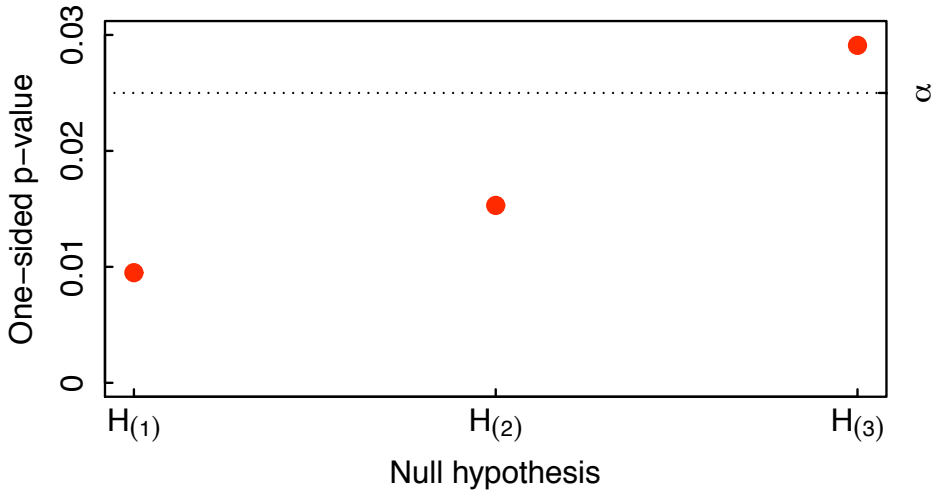
No significant effect at Dose 1 at $\alpha = 0.025$ due to tolerability problems

$p_{(1)} = p_2 = 0.0095$, $p_{(2)} = p_3 = 0.0153$ and $p_{(3)} = p_1 = 0.0291$, ordered p-values

$H_{(1)}$, $H_{(2)}$ and $H_{(3)}$, ordered null hypotheses

Hommel procedure

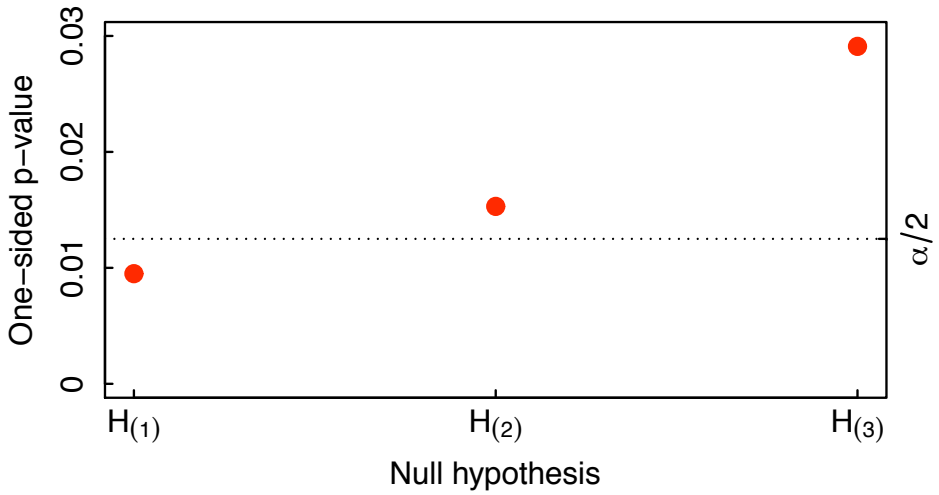
Decision rules in Example 4 ($\alpha = 0.025$)



Step 1: $H_{(3)}$ is accepted since $p_{(3)} > \alpha$

Hommel procedure

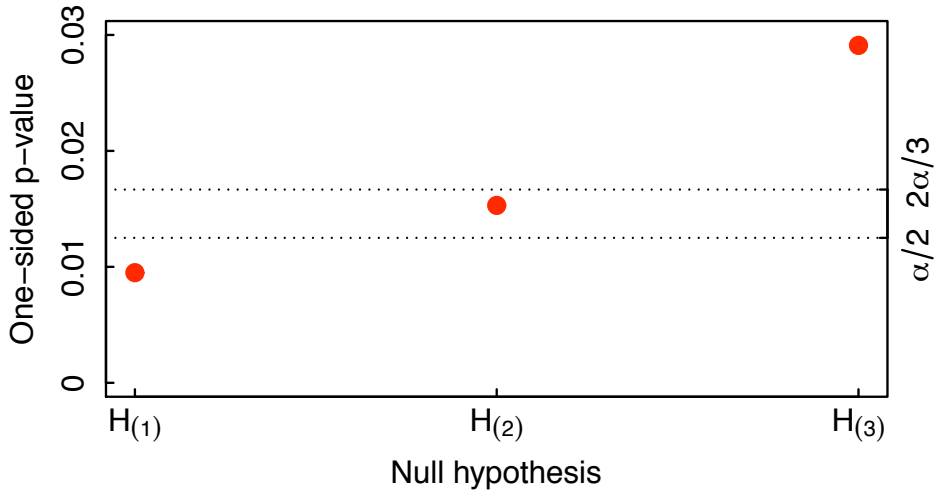
Decision rules in Example 4 ($\alpha = 0.025$)



Step 2: $H_{(2)}$ is accepted since $p_{(3)} > \alpha$ and $p_{(2)} > \alpha/2$

Hommel procedure

Decision rules in Example 4 ($\alpha = 0.025$)



Step 3: $H_{(1)}$ is rejected since $p_{(3)} > \alpha$, $p_{(2)} \leq 2\alpha/3$ and $\alpha/3 < p_{(1)} \leq \alpha/2$

Hochberg procedure

Hochberg procedure

Simes-based step-up procedure

Hochberg procedure is a semiparametric derived from simplified Simes global test (which is less powerful than regular Simes global test)

Hochberg procedure is based on a straightforward step-up algorithm

Hochberg procedure

General decision rules (step-up algorithm)

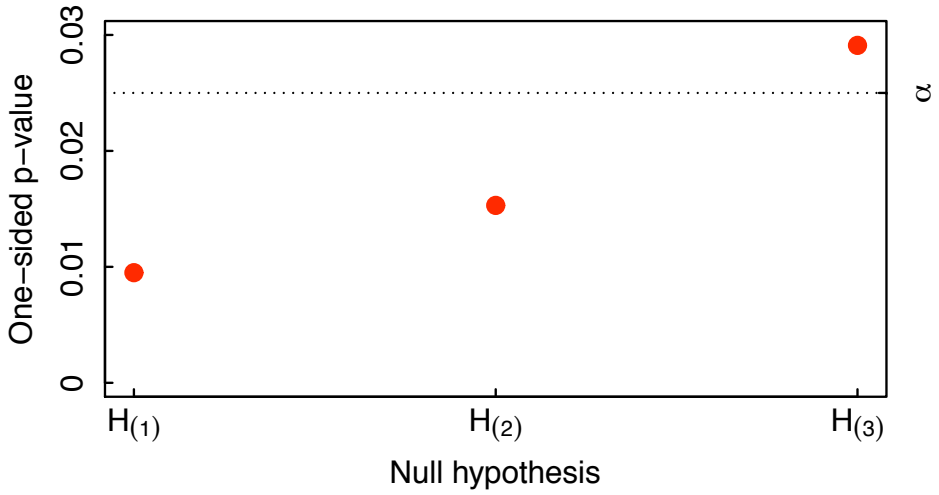
Step 1: If $p_{(m)} > \alpha$, accept $H_{(m)}$ and go to Step 2, otherwise reject all hypotheses and stop

Steps $i = 2, \ldots, m-1$: If $p_{(m-i+1)} > \alpha/i$, accept $H_{(m-i+1)}$ and go to Step $i+1$, otherwise reject all remaining hypotheses and stop

Step m: If $p_{(1)} > \alpha/m$, accept $H_{(1)}$, otherwise reject $H_{(1)}$

Hochberg procedure

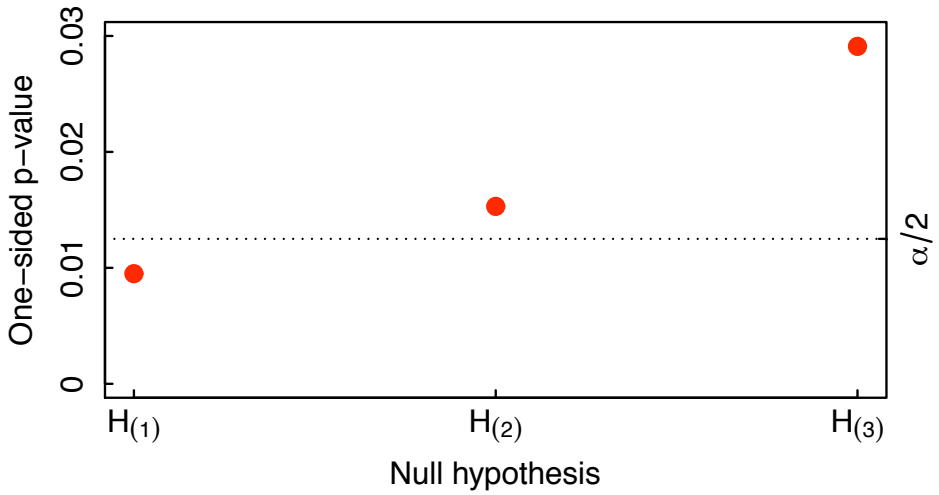
Decision rules in Example 4 ($\alpha = 0.025$)



Step 1: $H_{(3)}$ is accepted since $p_{(3)} > \alpha$

Hochberg procedure

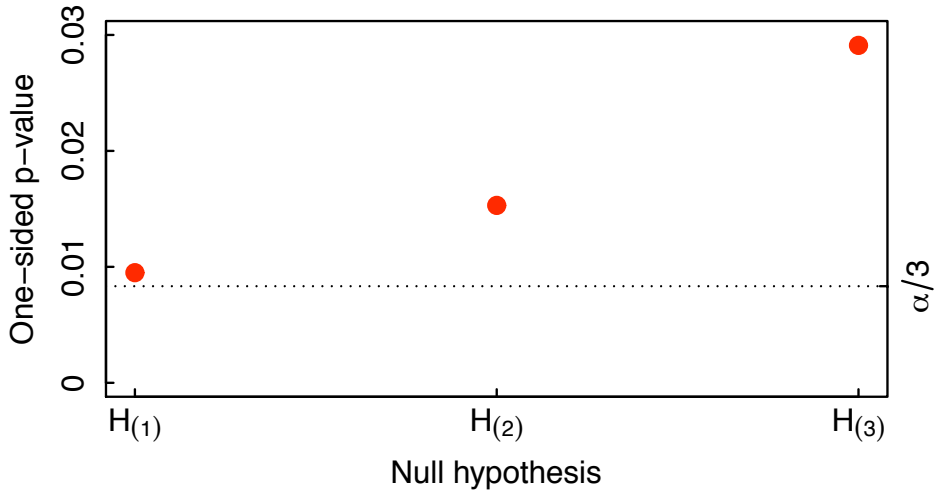
Decision rules in Example 4 ($\alpha = 0.025$)



Step 2: $H_{(2)}$ is accepted since $p_{(2)} > \alpha/2$

Hochberg procedure

Decision rules in Example 4 ($\alpha = 0.025$)



Step 3: $H_{(1)}$ is accepted since $p_{(1)} > \alpha/3$

Hommel and Hochberg procedures

Properties

Both Hommel and Hochberg procedures reject all null hypotheses if all raw p-values $\leq \alpha$

Power

Hochberg procedure is uniformly more powerful than Holm procedure

Hommel procedure is uniformly more powerful than Hochberg procedure

Power comparison

Hochberg procedure

Very popular in clinical trial applications even though it is less powerful than Hommel procedure

Based on a simple algorithm and is easy to explain to non-statisticians

Hommel procedure

Recommended since it provides a uniform improvement in power

Hommel and Hochberg procedures

FWER control

Hommel and Hochberg procedures are semiparametric and control FWER only under additional assumptions on the joint distribution of test statistics (when Simes global test controls Type I error rate)

Worst-case scenario

In general multiplicity problems with m null hypotheses, the upper limit on FWER is $(1 + 1/2 + \ldots + 1/m)\alpha$, i.e., 1.5α if $m = 2$ and 2.1α if $m = 4$

Hommel and Hochberg procedures

FWER inflation

Upper limit on FWER inflation is achieved only in pathological cases

With normally distributed test statistics, FWER inflation tends to be trivial

Example

Comprehensive simulation-based evaluation of FWER inflation in problems with 3, 5 and 10 hypotheses (Sarkar and Chang, 1997)

Highest error rate was 0.0254 (one-sided)

Hommel and Hochberg procedures

Positive dependence condition

Simes global test controls Type I error rate when the positive dependence condition is satisfied, i.e., the joint distribution of hypothesis test statistics is multivariate totally positive of order two (MTP2) (Sarkar and Chang, 1997; Sarkar, 1998; see also Huque, 2016)

Example

Positive dependence condition is satisfied for multivariate normal test statistics with non-negative pairwise correlations (Sarkar, 2008)

Positive dependence condition

Example 1: Prostate cancer trial

Condition is satisfied if the two endpoints are positively correlated

Example 4: Type 2 diabetes trial

Condition is satisfied since the treatment arms are compared to a common control

Example 5: Non-small-cell lung cancer trial

Condition is satisfied since the subpopulation is a subset of the overall population

Hommel and Hochberg procedures

FDA guidance (FDA, 2017)

"Beyond the aforementioned cases where the Hochberg procedure is known to be valid, its use is generally not recommended for the primary comparisons of confirmatory clinical trials unless it can be shown that adequate control of Type I error rate is provided."

Exercise

Exercise

Cardiovascular clinical trial

Study start

A single primary endpoint

Interim analysis

A co-primary endpoint was added

FWER control

Multiple testing procedure is required to control familywise error rate

Exercise

Proposal

Bonferroni procedure is too conservative

Alternative approaches: Holm, Hommel, and Hochberg procedures

Power comparison: Holm $<$ Hommel $<$ Hochberg

Hommel procedure does not always control FWER whereas Hochberg procedure always does

Hochberg procedure is superior to Hommel procedure and will be used in the study

Multiple testing procedures

Power comparison

Less powerful More powerful

Bonferroni Holm Hochberg Hommel

Software implementation

Software implementation in SAS

SAS/STAT module

Nonparametric and semiparametric procedures: MULTTEST procedure

Custom macros

Nonparametric and semiparametric procedures: PVALPROC macro

Software implementation in R

MultComp package

Parametric procedures for linear and semi-parametric models

http://cran.r-project.org/web/packages/multcomp/index.html

Mediana package

AdjustPvalues function: Nonparametric, semiparametric and fully parametric procedures

http://biopharmnet.com/mediana

Example 4: Type 2 diabetes trial

Three dose-placebo comparisons

Comparison	P-value
Dose 1 vs Placebo (H_1)	$p_1 = 0.0111$
Dose 2 vs Placebo (H_2)	$p_2 = 0.0065$
Dose 3 vs Placebo (H_3)	$p_3 = 0.0293$

Holm procedure in SAS

Data set

```
data ex4;
  input raw_p;
  weight=1/3;
  datalines;
  0.0111
  0.0065
  0.0293
run;
```


Holm procedure in SAS

MULTTEST procedure

```
proc multtest pdata=ex4 holm;
  run;
```

Other options

bonferroni hochberg hommel

Holm procedure in SAS

MULTTEST procedure: Output

Test	Raw	Stepdown Bonferroni
1	0.0111	0.0222
2	0.0065	0.0195
3	0.0293	0.0293

Holm procedure in SAS

PVALPROC macro

```
%pvalproc(in=ex4,out=adjp);
proc print data=adjp noobs label;
  format raw holm 6.4;
  var test raw holm;
  run;
```

Other options

bonferroni hochberg hommel

Example 1: Prostate cancer trial

Two primary endpoints

Endpoint 1: Overall survival (OS)

Endpoint 2: Radiographic progression-free survival (rPFS)

Unequal weights

Endpoint 1: 0.8

Endpoint 2: 0.2

Weighted Holm procedure in SAS

Data set and PVALPROC macro

```
data ex1;
  input raw_p weight;
  datalines;
  0.0102 0.8
  0.0181 0.2
run;
%pvalproc(in=ex1,out=adjp);
proc print data=adjp noobs label;
  format raw holm 6.4;
  var test raw holm;
  run;
```

Weighted Holm procedure in SAS

PVALPROC macro: Output

Test	Raw	Holm
1	0.0102	0.0128
2	0.0181	0.0181

Weighted Holm procedure in R

Mediana package (AdjustPvalues function)

```
rawp=c(0.0102,0.0181)
weight=c(0.8,0.2)
adjp=AdjustPvalues(rawp, proc="HolmAdj",
  par=parameters(weight=weight))
round(adjp, 4)
```

Output

```
0.0128 0.0181
```

Other options

BonferroniAdj HochbergAdj HommelAdj

Module D
Nonparametric multiple testing procedures II

Module D outline

D1. Stepwise procedures with a pre-specified testing sequence

Commonly used stepwise procedures

Fixed-sequence procedure (Maurer et al., 1995)

Fallback procedure (Wiens, 2003; Wiens and Dmitrienko, 2005)

Class of chain procedures/graphical procedures (Bretz et al., 2009; Burman, Sonesson and Guilbaud, 2009; Millen and Dmitrienko, 2011)

Section D1
Stepwise procedures with a pre-specified testing sequence

Pre-specified testing sequence

Distributional relationships

Nonparametric procedures that make no assumptions about the joint distribution of test statistics

Logical relationships

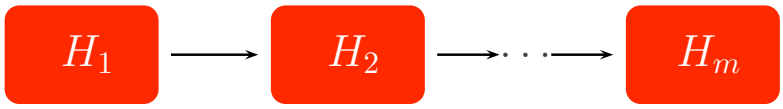
Stepwise procedures with a pre-specified testing sequence

Null hypotheses are ordered to reflect clinical importance or probability of success for the associated objectives

Fixed-sequence procedure

Fixed-sequence procedure

Sequentially rejective method



— null hypothesis is rejected

Each null hypothesis is tested at α level

Single-strike rule is applied: Stop testing after first non-significant outcome

Fixed-sequence procedure

General decision rules

Step 1: If $p_1 \leq \alpha$, reject H_1 and go to Step 2, otherwise accept all hypotheses and stop

Steps $i = 2, \ldots, m-1$: If $p_i \leq \alpha$, reject H_i and go to Step $i+1$, otherwise accept $H_i, \ldots, H_m$ and stop

Step m: If $p_m \leq \alpha$, reject H_m, otherwise accept H_m

Exercise

Clinical trial with two interim analyses and final analysis

Can the fixed-sequence procedure be used in this trial to control overall Type I error rate?

Example 4: Type 2 diabetes trial

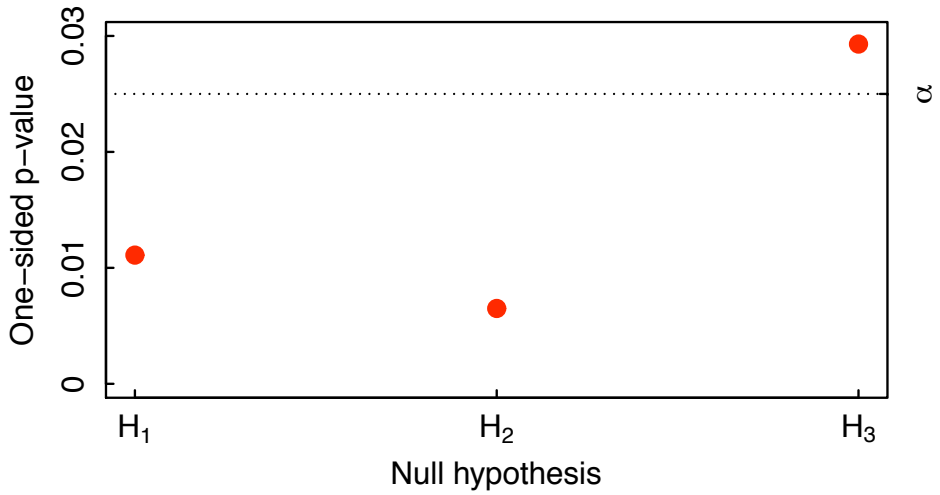
Scenario 1

Comparison	P-value
Dose 1 vs Placebo (H_1)	0.0111
Dose 2 vs Placebo (H_2)	0.0065
Dose 3 vs Placebo (H_3)	0.0293

Evidence of treatment effect at Doses 1 and 2

Fixed-sequence procedure

Decision rules in Example 4 ($\alpha = 0.025$)



Fixed-sequence procedure rejects H_1 and H_2

Fixed-sequence procedure

Type I error rate

Fixed-sequence procedure is nonparametric and controls FWER for any joint distribution of hypothesis test statistics

Power

Power is maximized under the monotonicity assumption (null hypotheses are ordered from the largest effect size to the smallest effect size)

Power loss is likely when the monotonicity assumption is violated

Example 4: Type 2 diabetes trial

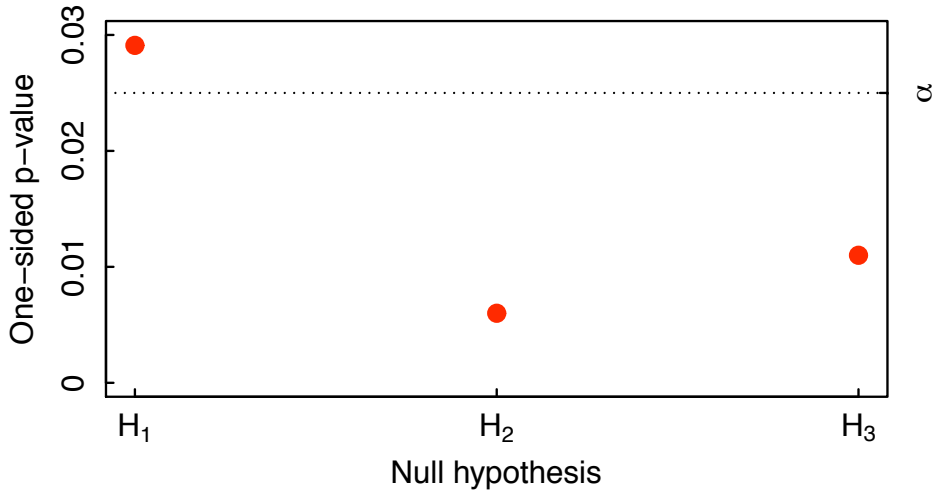
Scenario 3

Comparison	P-value
Dose 1 vs Placebo (H_1)	0.0291
Dose 2 vs Placebo (H_2)	0.0060
Dose 3 vs Placebo (H_3)	0.0110

No significant effect at Dose 1 at $\alpha = 0.025$

Fixed-sequence procedure

Decision rules in Example 4 ($\alpha = 0.025$)



Fixed-sequence procedure rejects no null hypotheses

Power evaluation

Clinical trial

Three endpoints are tested sequentially using the fixed-sequence procedure

Responses are equicorrelated and follow a multivariate normal distribution

Power

e_1, e_2, e_3, effect sizes for endpoint tests

80% power for each endpoint test with 98 patients per group if $e_1 = e_2 = e_3 = 0.4$

Effect size assumptions

Case 1

All tests are adequately powered, $e_1 = 0.4$, $e_2 = 0.4$, $e_3 = 0.4$

Case 2

First test is underpowered, $e_1 = 0.3$, $e_2 = 0.4$, $e_3 = 0.4$

Case 3

First test is overpowered, $e_1 = 0.5$, $e_2 = 0.4$, $e_3 = 0.4$

Summary of power evaluation

Power of three endpoint tests (%)

Correlation	Fixed-sequence procedure
Case 1	
0	$(80, 63, 51)$
0.5	$(80, 68, 61)$
Case 2	
0	$(55, 44, 35)$
0.5	$(55, 49, 46)$
Case 3	
0	$(94, 75, 60)$
0.5	$(94, 77, 67)$

Fixed-sequence procedure

EMA guidance (EMA, 2017)

"Two or more endpoints ranked according to clinical relevance: No numerical adjustment of each single hypothesis test is necessary. However, no confirmatory claims can be based on endpoints that have a rank lower than or equal to that variable whose null hypothesis was the first that could not be rejected."

Fixed-sequence procedure

FDA guidance (FDA, 2017)

"Carefully selecting the ordering of the tests of hypotheses is essential. A test early in the sequence that fails to show statistical significance will render the remainder of the endpoints not statistically significant. It is often not possible to determine a priori the best order for testing, and there are other methods for addressing the multiplicity problem..."

Fallback procedure

Fallback procedure

Flexible fixed-sequence procedure

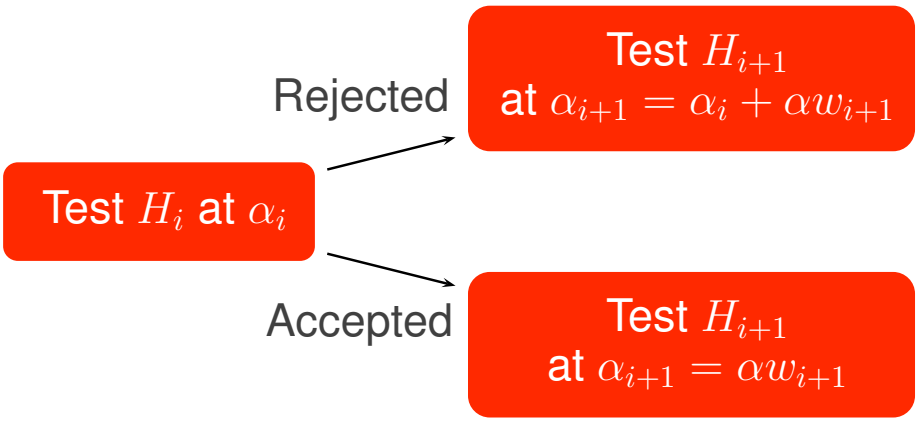
Fallback procedure is an extension of the fixed-sequence procedure derived from the Bonferroni procedure

Notation

$w_1, \ldots, w_m$, weights assigned to $H_1, \ldots, H_m$
($w_i \geq 0$, $i = 1, \ldots, m$, and $w_1 + \ldots + w_m = 1$)

Fallback procedure

Testing method with a fallback option



No single-strike rule

Fallback procedure

General decision rules

Step 1: Let $\alpha_1 = \alpha w_1$. If $p_1 \leq \alpha_1$, reject H_1, otherwise accept H_1. Go to Step 2

Steps $i = 2, \ldots, m-1$: Let $\alpha_i = \alpha_{i-1} + \alpha w_i$ if H_{i-1} is rejected and $\alpha_i = \alpha w_i$ if H_{i-1} is accepted. If $p_i \leq \alpha_i$, reject H_i, otherwise accept H_i. Go to Step $i+1$

Step m: Let $\alpha_m = \alpha_{m-1} + \alpha w_m$ if H_{m-1} is rejected and $\alpha_m = \alpha w_m$ if H_{m-1} is accepted. If $p_m \leq \alpha_m$, reject H_m, otherwise accept H_m

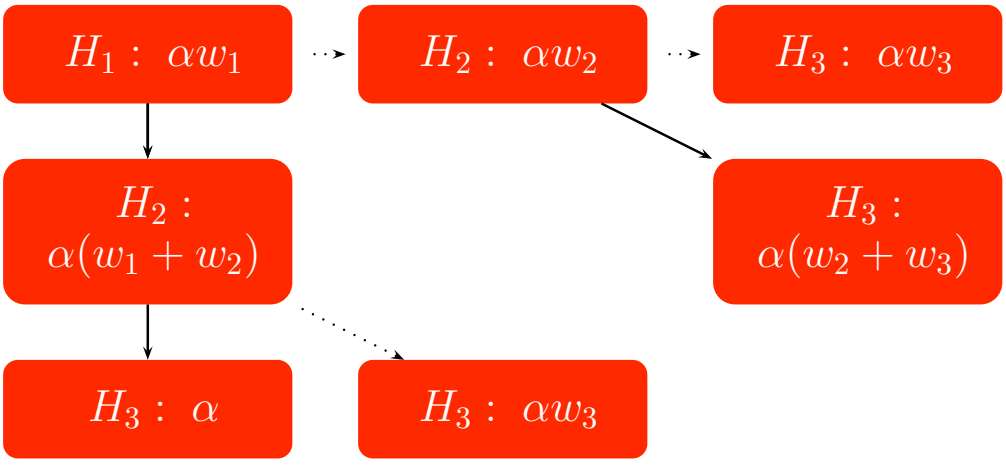
Example 4: Type 2 diabetes trial

Scenario 3

Comparison	P-value	Weight
Dose 1 vs Placebo (H_1)	0.0291	$1/2$
Dose 2 vs Placebo (H_2)	0.0060	$1/4$
Dose 3 vs Placebo (H_3)	0.0110	$1/4$

Greater weight is assigned to Dose 1 since it is expected to be more effective than Dose 2 or Dose 3

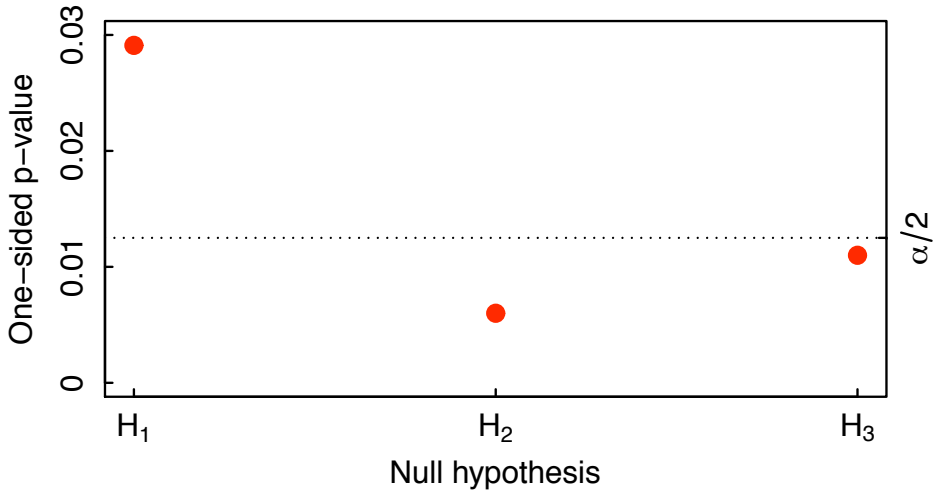
Fallback procedure



— H_i is rejected, $\cdots$ H_i is accepted

Fallback procedure

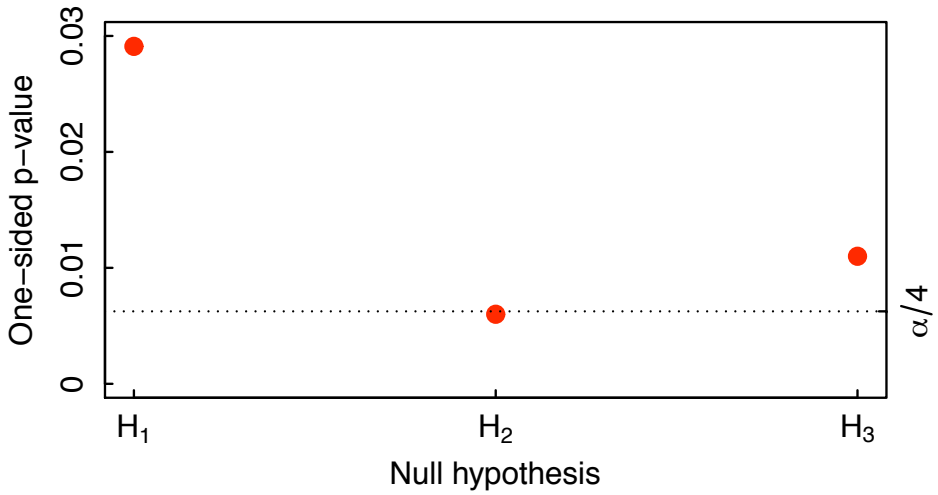
Decision rules in Example 4 ($\alpha = 0.025$)



Step 1: H_1 is accepted since $p_1 > \alpha_1 = \alpha w_1 = \alpha/2$

Fallback procedure

Decision rules in Example 4 ($\alpha = 0.025$)

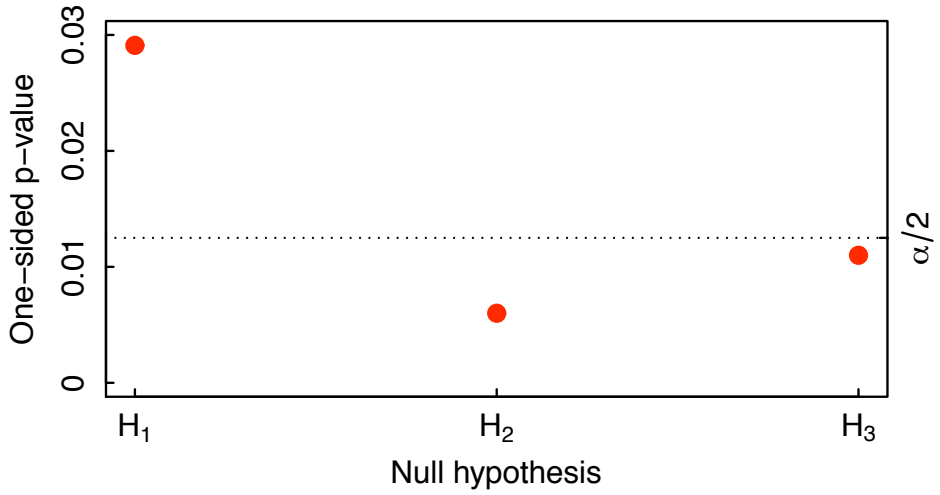


Step 2: H_2 is rejected since $p_2 < \alpha_2 = \alpha w_2 = \alpha/4$

α_2 is carried over to H_3

Fallback procedure

Decision rules in Example 4 ($\alpha = 0.025$)



Step 3: H_3 is rejected since $p_3 < \alpha_3 = \alpha_2 + \alpha w_3 = \alpha/4 + \alpha/4 = \alpha/2$

Fallback procedure

Type I error rate

Fallback procedure is nonparametric and controls FWER for any joint distribution of hypothesis test statistics

Power

Fallback is uniformly more powerful than weighted Bonferroni but is uniformly less powerful than weighted Holm with the same set of weights

Fallback may be more powerful or less powerful than weighted Hommel with the same set of weights

Fallback procedure

Extension

Fallback procedure is an extension of fixed-sequence procedure (fallback procedure simplifies to fixed-sequence procedure if $w_1 = 1$ and $w_2 = \ldots = w_m = 0$)

Fallback procedure is an attractive alternative to fixed-sequence procedure when the monotonicity assumption is violated

Fallback and fixed-sequence procedures

Clinical trial

Same set of assumptions as before

Fallback procedure

$w_1 = 0.5$, $w_2 = 0.25$, $w_3 = 0.25$, weights assigned to three endpoint tests

Summary of power evaluation

Power of three endpoint tests (%)

Correlation	Fixed-sequence procedure	Fallback procedure
	Case 1	
0	$(80, 63, 51)$	$(70, 72, 73)$
0.5	$(80, 68, 61)$	$(70, 71, 72)$
	Case 2	
0	$(55, 44, 35)$	$(43, 68, 71)$
0.5	$(55, 49, 46)$	$(43, 66, 70)$
	Case 3	
0	$(94, 75, 60)$	$(90, 75, 75)$
0.5	$(94, 77, 67)$	$(90, 75, 74)$

Chain procedures

Class of chain procedures

Stepwise procedures with flexible decision rules

Take general logical relationships among null hypotheses into account

Known as graphical procedures or as chain procedures since testing algorithm is similar to a chain

Cyclical and serial chain algorithms

Cyclical algorithms rely on a data-driven testing sequence

Focus on serial algorithms with a pre-specified testing sequence

Example 4: Type 2 diabetes trial

Logical relationships

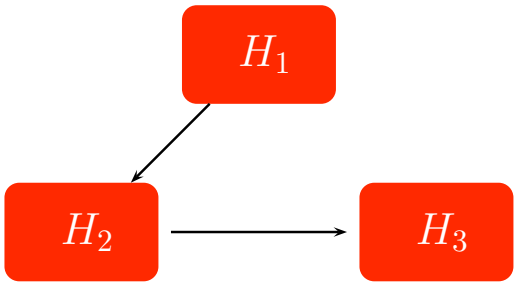
Dose 1 is expected to more effective than the other two doses (Doses 2 and 3)

Dose 2 (H_2) and Dose 3 (H_3) are tested after Dose 1 (H_1)

Dose 2 (H_2) is more important than Dose 3 (H_3)

Fallback procedure

Logical relationships

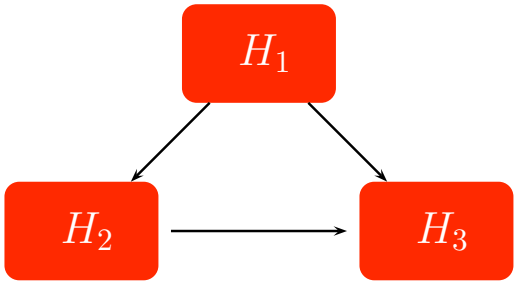


Null hypotheses are tested sequentially

Logical relationships among three null hypotheses are not taken into account

Serial chain procedure

Logical relationships



Most important null hypothesis (H_1) is tested first followed by less important null hypotheses (H_2 and H_3)

Logical relationships among three null hypotheses are taken into account

Chain procedures

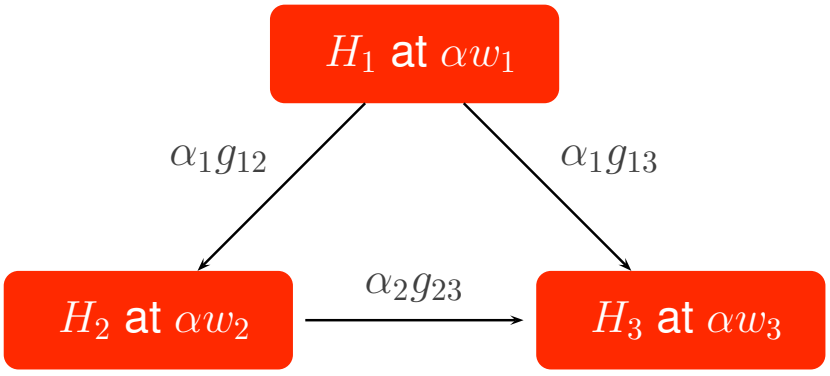
Flexible testing procedures

α allocation rule: Initial distribution of the error rate across the hypotheses

α propagation rule: Transfer of the error rate from a rejected hypothesis to non-rejected hypotheses

Serial chain procedure

α allocation and propagation rules



Serial chain procedure

Parameters

w_1, w_2, w_3, weights of null hypotheses (α allocation rule)

g_{12}, g_{13}, g_{23}, transition parameters with $g_{12} + g_{13} = 1$ and $g_{23} = 1$ (α propagation rule)

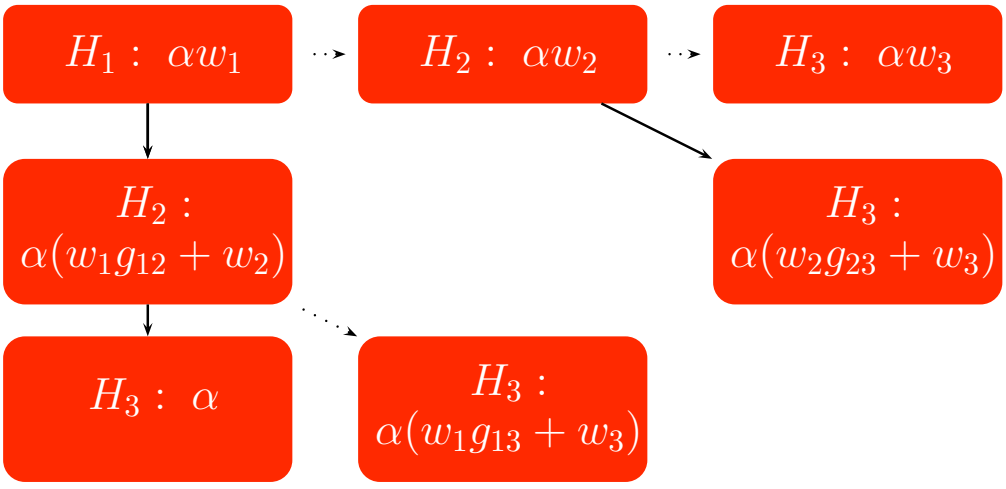
Transition parameters

g_{12}, fraction of α_1 carried forward from H_1 to H_2

g_{13}, fraction of α_1 carried forward from H_1 to H_3

g_{23}, fraction of α_2 carried forward from H_2 to H_3

Example 4: Type 2 diabetes trial



— H_i is rejected, $\cdots$ H_i is accepted

Example 4: Type 2 diabetes trial

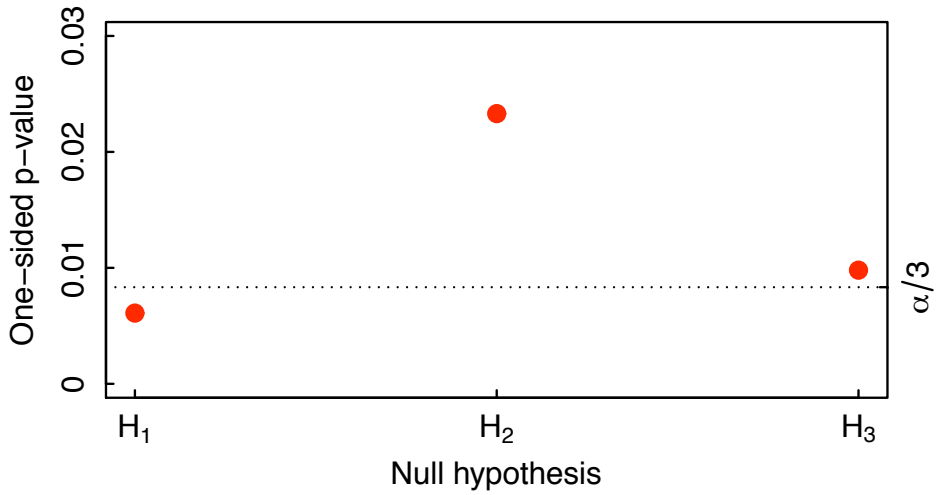
Scenario 4

Comparison	P-value	Weight
Dose 1 vs Placebo (H_1)	0.0061	$1/3$
Dose 2 vs Placebo (H_2)	0.0233	$1/3$
Dose 3 vs Placebo (H_3)	0.0098	$1/3$

Transition parameters, $g_{12} = 1/2$, $g_{13} = 1/2$ and $g_{23} = 1$

Serial chain procedure

Decision rules in Example 4 ($\alpha = 0.025$)

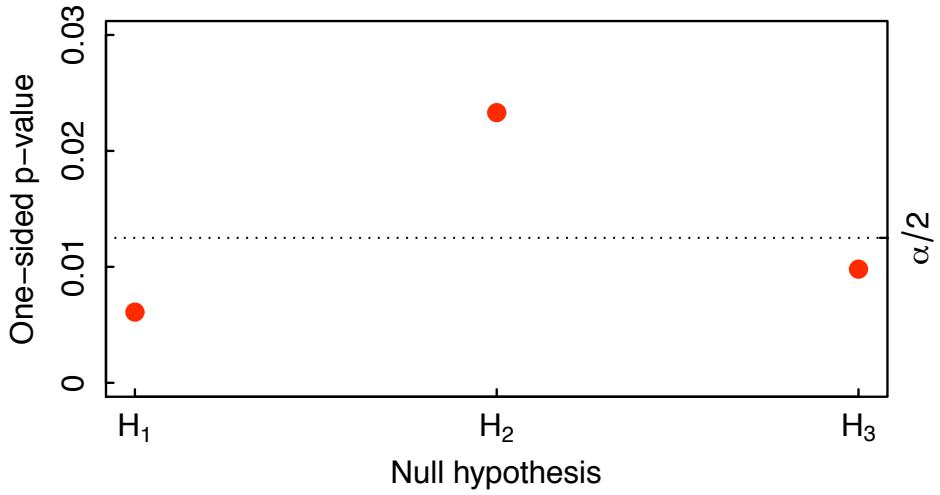


Step 1: H_1 is rejected since $p_1 < \alpha_1 = \alpha w_1 = \alpha/3$

$\alpha_1 g_{12} = \alpha_1/2$ is carried over to H_2 and $\alpha_1 g_{13} = \alpha_1/2$ to H_3

Serial chain procedure

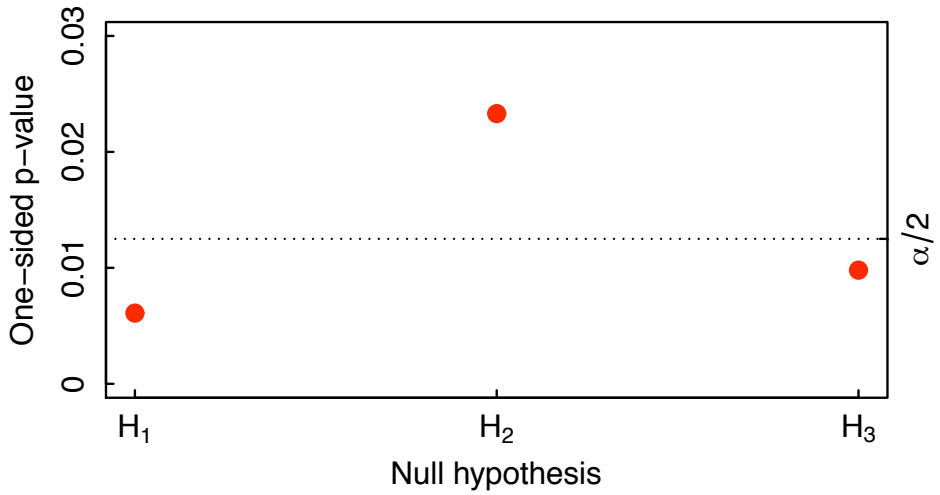
Decision rules in Example 4 ($\alpha = 0.025$)



Step 2: H_2 is accepted since $p_2 > \alpha_2 = \alpha w_2 + \alpha_1 g_{12}$
$= \alpha/3 + \alpha/3(1/2) = \alpha/2$

Serial chain procedure

Decision rules in Example 4 ($\alpha = 0.025$)



Step 3: H_3 is rejected since $p_3 < \alpha_3 = \alpha w_3 + \alpha_1 g_{13}$
$= \alpha/3 + \alpha/3(1/2) = \alpha/2$

Serial chain procedure

General testing problem

$H_1, \ldots, H_m$, null hypotheses

$w_1, \ldots, w_m$, weights of null hypotheses

g_{ij}, $i = 1, \ldots, m-1$, $j = i+1, \ldots, m$, transition parameters ($g_{ij} \geq 0$ and $\sum_{j=i+1}^{m} g_{ij} = 1$)

g_{ij}, fraction of α_i carried from H_i to H_j when H_i is rejected

General α propagation rule

If H_i is rejected, $\alpha_i g_{ij}$ is carried over to H_j, $i = 1, \ldots, m-1$, $j = i+1, \ldots, m$

Chain procedures

Type I error rate

Chain procedures are nonparametric and control FWER for any joint distribution of hypothesis test statistics

Power

Chain procedures are uniformly more powerful than weighted Bonferroni procedure and may be more powerful or less powerful than Hommel procedure with the same set of weights

Multiple testing procedures

Power comparison

Less powerful — More powerful

Bonferroni | Holm | Hochberg | Hommel

Fallback/Chain

Software implementation

Software implementation in SAS

Custom macros

Fallback and fixed-sequence procedures: PVALPROC macro

Serial chain procedure: CHAINSER macro

Software implementation in R

Mediana package

AdjustPvalues function: Supports commonly used nonparametric, semiparametric and fully parametric procedures

Web site

http://biopharmnet.com/mediana

Example 4: Type 2 diabetes trial

Scenario 4

Comparison	P-value	Weight
Dose 1 vs Placebo (H_1)	0.0061	$1/3$
Dose 2 vs Placebo (H_2)	0.0233	$1/3$
Dose 3 vs Placebo (H_3)	0.0098	$1/3$

Fallback procedure in SAS

Data set

```
data ex4;
  input raw_p weight;
  datalines;
  0.0061 0.3333
  0.0233 0.3333
  0.0098 0.3334
run;
```

Fallback procedure in SAS

PVALPROC macro

```
%pvalproc(in=ex4,out=adjp);
proc print data=adjp noobs label;
  format raw fallback 6.4;
  var test raw fallback;
  run;
```

Other options

fixedseq

Fallback procedure in SAS

PVALPROC macro: Output

Test	Raw	Fallback
1	0.0061	0.0183
2	0.0233	0.0350
3	0.0098	0.0294

Chain procedure

α allocation rule

Hypothesis weights

$$W = (1/3, 1/3, 1/3)$$

α propagation rule

Transition parameters

$$G = \begin{bmatrix} 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

Chain procedure in SAS

Data set

```
data ex4chain;
  input raw_p weight g1-g3;
  datalines;
  0.0061 0.3333 0.0 0.5 0.5
  0.0233 0.3333 0.0 0.0 1.0
  0.0098 0.3334 0.0 0.0 0.0
run;
```

Chain procedure in SAS

CHAINSER macro

```
%chainser(in=ex4chain,out=adjp);
proc print data=adjp noobs label;
  format raw chain 6.4;
  var test raw chain;
  run;
```


Chain procedure in SAS

CHAINSER macro: Output

Test	Raw	Chain
1	0.0061	0.0183
2	0.0233	0.0466
3	0.0098	0.0196

Fallback procedure

α allocation rule

Hypothesis weights

$$W = (1/3, 1/3, 1/3)$$

α propagation rule

Transition parameters

$$G = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

Fallback procedure in R

Mediana package (AdjustPvalues function)

```
rawp=c(0.0061,0.0233,0.0098)
weight=c(1/3,1/3,1/3)
transition = matrix(c(0, 1, 0,
                      0, 0, 1,
                      0, 0, 0),
             3, 3, byrow = TRUE)
adjp=AdjustPvalues(rawp,
  proc = "ChainAdj",
  par = parameters(weight = weight,
  transition = transition))
round(adjp, 4)
```

Fallback procedure in R

AdjustPvalues function: Output

```
0.0183 0.0350 0.0294
```


Chain procedure in R

Mediana package (AdjustPvalues function)

```
rawp=c(0.0061,0.0233,0.0098)
weight=c(1/3,1/3,1/3)
transition = matrix(c(0, 0.5, 0.5,
                      0, 0, 1,
                      0, 0, 0),
             3, 3, byrow = TRUE)
adjp=AdjustPvalues(rawp,
  proc = "ChainAdj",
  par = parameters(weight = weight,
  transition = transition))
round(adjp, 4)
```

Chain procedure in R

AdjustPvalues function: Output

```
0.0183 0.0466 0.0196
```


Module E
Parametric procedures

Module E outline

E1. Parametric multiple testing procedures

Dunnett family of parametric procedures

Other parametric procedures

Section E1
Parametric multiple testing procedures

Parametric procedures

Distributional relationships

Make explicit distributional assumptions, e.g., hypothesis test statistics follow a multivariate normal or t distribution

More powerful than nonparametric or semiparametric procedures because they account for the correlations among test statistics

Logical relationships

Single-step procedures

Stepwise procedures with a data-driven or pre-specified hypothesis ordering

Dunnett family of parametric procedures

Single-step Dunnett procedure

Parametric version of Bonferroni procedure (Dunnett, 1955)

Step-down Dunnett procedure

Parametric version of Holm procedure (Naik, 1975; Marcus, Peritz and Gabriel, 1976; Dunnett and Tamhane, 1991)

Step-up Dunnett procedure

Parametric version of Hochberg procedure (Dunnett and Tamhane, 1992)

Parametric testing problem

Dose-finding clinical trial

Several doses or regimens are compared to a common control (placebo)

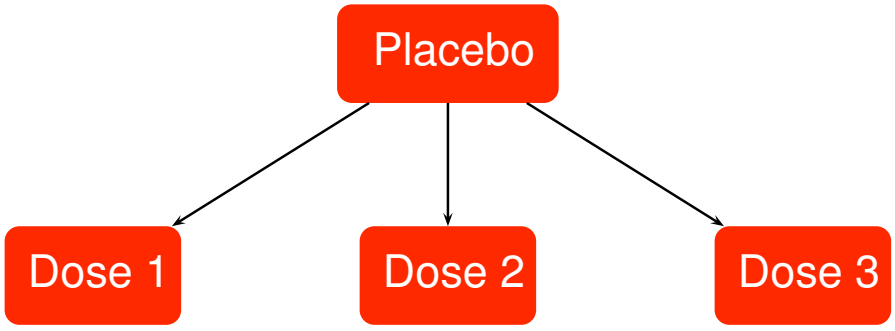
Assumptions

Responses are normally distributed

Balanced design (equal number of patients across treatment groups)

Example 4: Type 2 diabetes trial

Three doses compared to placebo



Parametric testing problem

ANOVA model

$y_{ij} = \mu_i + \varepsilon_{ij}$

where $i = 0, \ldots, m$ ($i = 0$ denotes placebo group) and $j = 1, \ldots, n$

y_{ij}, response of jth patient in ith group

μ_i, $i = 1, \ldots, m$, mean response in ith group

ε_{ij}, $i = 0, \ldots, m$, $j = 1, \ldots, n$, normally distributed errors

Parametric testing problem

Null hypotheses

$H_i: \ \theta_i \leq 0$, $i = 1, \ldots, m$

where $\theta_i = \mu_i - \mu_0$, $i = 1, \ldots, m$, are mean treatment differences

Hypothesis test statistics

$t_1, \ldots, t_m$, test statistics

t_i follows a t distribution with $\nu = 2(n-1)$ degrees of freedom

P-values

$p_1, \ldots, p_m$, p-values

P-value-based procedure

Bonferroni procedure

Single-step p-value-based procedure

Adjusted significance level

Assume that $H_1, \ldots, H_m$ are true

Adjusted significance level for p-values, $\tilde{\alpha} = \alpha/m$, is found from Bonferroni inequality

$$P(p_1 \leq \tilde{\alpha} \text{ or } \ldots \text{ or } p_m \leq \tilde{\alpha}) \leq \sum_{i=1}^{m} P(p_i \leq \tilde{\alpha}) = \alpha$$

Dunnett procedure

Parametric procedure

Dunnett procedure

Single-step parametric procedure

Adjusted significance level

Assume that $H_1, \ldots, H_m$ are true

Adjusted significance level for p-values, $\tilde{\alpha}$, is found from

$$P(p_1 \leq \tilde{\alpha} \text{ or } \ldots \text{ or } p_m \leq \tilde{\alpha}) = \alpha$$

using joint distribution of p-values

Parametric procedure

Adjusted critical value

Adjusted critical value for test statistics, c, is found from

$$\begin{aligned} \alpha &= P(t_1 \geq c \text{ or } \ldots \text{ or } t_m \geq c) \\ &= P(\max(t_1, \ldots, t_m) \geq c) = P(T \geq c) \end{aligned}$$

using joint distribution of test statistics

$T = \max(t_1, \ldots, t_m)$ follows Dunnett distribution with m and $\nu = (m+1)(n-1)$ degrees of freedom

Parametric procedure

Dunnett distribution

Maximum of m test statistics that follow a multivariate t distribution with $\nu = (m+1)(n-1)$ degrees of freedom and are equicorrelated with a common correlation coefficient $\rho = 1/2$ (balanced design)

Adjusted critical value of Dunnett procedure is $(1-\alpha)$ quantile of Dunnett distribution, i.e., $c = d_\alpha(m, \nu)$

Decision rule

Dunnett procedure rejects H_i if $t_i \geq c$, $i = 1, \ldots, m$

Example 4: Type 2 diabetes trial

Scenario 5

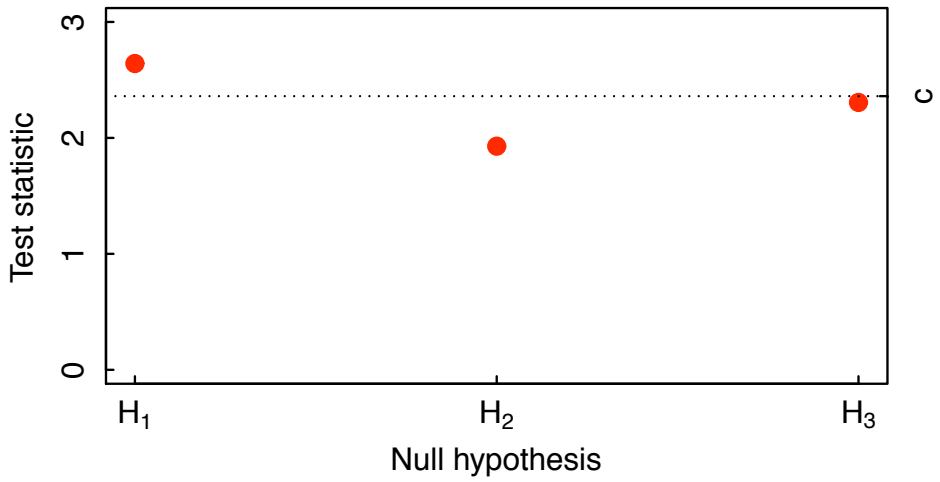
Comparison	Mean difference	Test statistic
Dose 1 vs Placebo (H_1)	0.63%	2.64
Dose 2 vs Placebo (H_2)	0.46%	1.93
Dose 3 vs Placebo (H_3)	0.55%	2.31

Sample size per group is 90 patients

Pooled standard deviation is 1.6

Dunnett procedure

Decision rules in Example 4 ($\alpha = 0.025$)

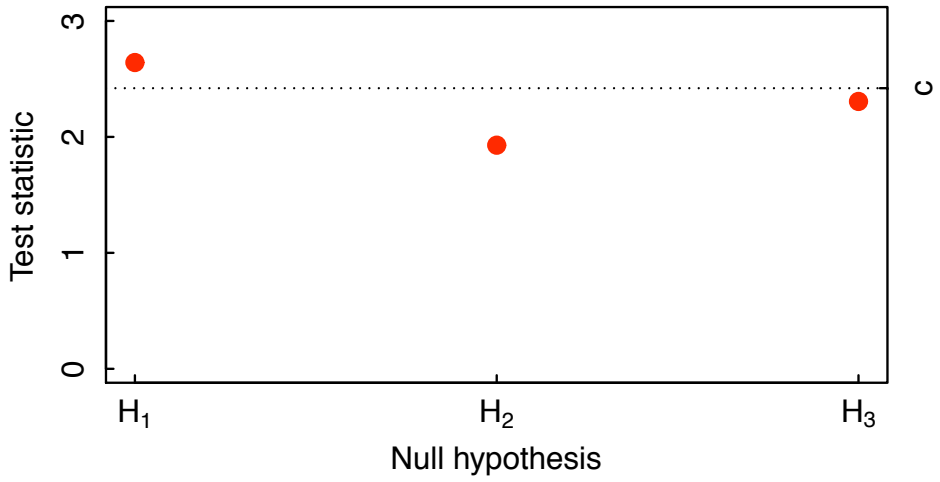


Dunnett critical value is $c = d_\alpha(m, \nu) = d_{0.025}(3, 356) = 2.36$

Dunnett procedure rejects H_1

Bonferroni procedure

Decision rules in Example 4 ($\alpha = 0.025$)



Bonferroni critical value is $c = t_{\alpha/3}(2(n-1)) = t_{0.0083}(178) = 2.42$

Bonferroni procedure rejects H_1

Dunnett procedure

Type I error rate

Dunnett procedure is parametric and controls FWER when hypothesis test statistics follow a multivariate t distribution, e.g., in trials with dose-control comparisons for normally distributed responses

Dunnett procedure can also be used with non-normally distributed responses, see Hothorn, Bretz and Westfall (2008)

Power

Dunnett procedure is uniformly more powerful than Bonferroni procedure

Step-down Dunnett procedure

Step-down Dunnett procedure

Data-driven hypothesis ordering

Null hypotheses are not ordered

Step-down procedure

Parametric version of Holm procedure, i.e., null hypotheses are tested sequentially beginning with the largest t statistic

Notation

$t_{(1)} > \ldots > t_{(m)}$, ordered test statistics

$H_{(1)}, \ldots, H_{(m)}$, ordered null hypotheses

Step-down Dunnett procedure

General decision rule

Step 1: If $t_{(1)} \geq c_1$, where $c_1 = d_\alpha(m, \nu)$, reject $H_{(1)}$ and go to Step 2, otherwise accept all hypotheses and stop

Steps $i = 2, \ldots, m-1$: If $t_{(i)} \geq c_i$, where $c_i = d_\alpha(m - i + 1, \nu)$, reject $H_{(i)}$ and go to Step $i + 1$, otherwise accept $H_{(i)}, \ldots, H_{(m)}$ and stop

Step m: If $t_{(m)} \geq c_m$, where $c_m = d_\alpha(1, \nu)$, reject $H_{(m)}$, otherwise accept $H_{(m)}$

Example 4: Type 2 diabetes trial

Scenario 5

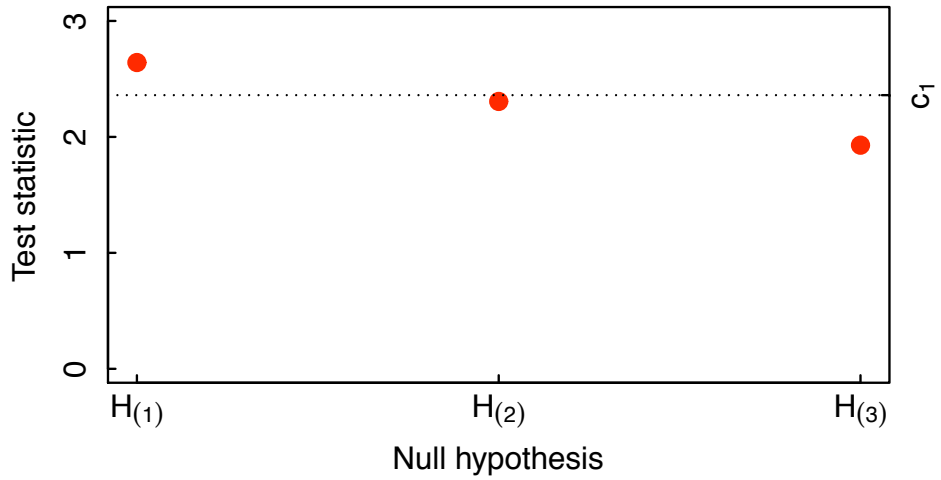
Comparison	Mean difference	Test statistic
Dose 1 vs Placebo (H_1)	0.63%	2.64
Dose 2 vs Placebo (H_2)	0.46%	1.93
Dose 3 vs Placebo (H_3)	0.55%	2.31

$t_{(1)} = t_1 = 2.64$, $t_{(2)} = t_3 = 2.31$, $t_{(3)} = t_2 = 1.93$, ordered test statistics

$H_{(1)}$, $H_{(2)}$ and $H_{(3)}$, ordered null hypotheses

Step-down Dunnett procedure

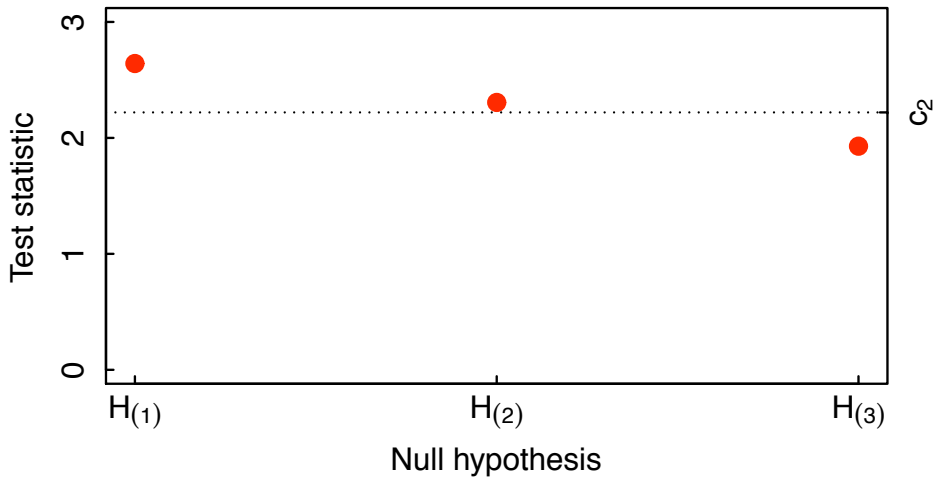
Decision rules in Example 4 ($\alpha = 0.025$)



Step 1: $H_{(1)}$ is rejected since $t_{(1)} > c_1 = d_\alpha(m, \nu) = d_{0.025}(3, 356) = 2.36$

Step-down Dunnett procedure

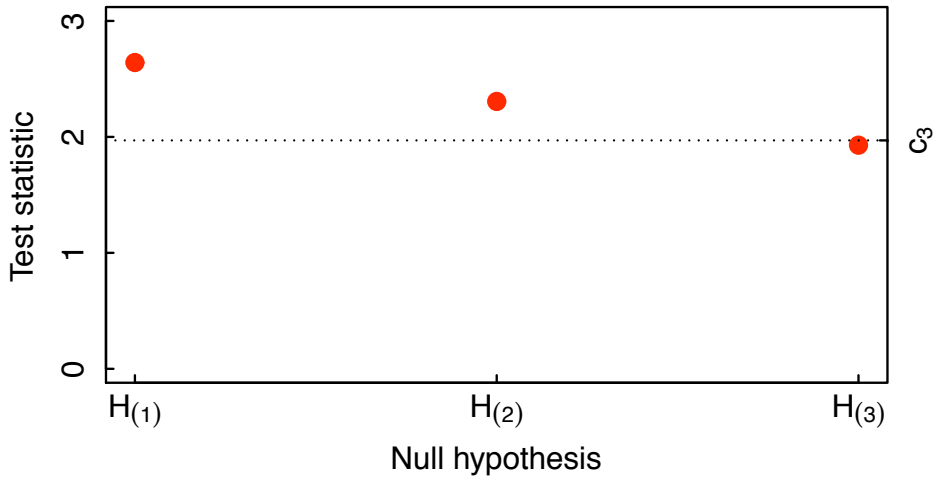
Decision rules in Example 4 ($\alpha = 0.025$)



Step 2: $H_{(2)}$ is rejected since $t_{(2)} > c_2 = d_\alpha(m-1,\nu) = d_{0.025}(2,356) = 2.22$

Step-down Dunnett procedure

Decision rules in Example 4 ($\alpha = 0.025$)



Step 3: $H_{(3)}$ is accepted since $t_{(3)} < c_3 = d_\alpha(m-2,\nu) = d_{0.025}(1,356) = 1.97$

Step-down Dunnett procedure

Type I error rate

Step-down Dunnett procedure is parametric and controls FWER when test statistics follow a multivariate t distribution

Power

Step-down Dunnett procedure is uniformly more powerful than Holm procedure and single-step Dunnett procedure

Step-up Dunnett procedure

Step-up Dunnett procedure

Data-driven hypothesis ordering

Null hypotheses are not ordered

Step-up procedure

Parametric version of Hochberg procedure, i.e., null hypotheses are tested beginning with the smallest t statistic

Based on a complicated algorithm and no software implementation is currently available

Not widely used in clinical trial applications and will not be discussed further in this course

Step-up Dunnett procedure

Type I error rate

Step-up Dunnett procedure is parametric and controls FWER when test statistics follow a multivariate t distribution

Power

Step-up Dunnett procedure is uniformly more powerful than Hochberg and single-step Dunnett procedure

Step-up Dunnett procedure is not always more powerful than step-down Dunnett procedure

Other parametric procedures

Other parametric procedures

Parametric fallback procedure

Extension of fallback procedure for a pre-specified hypothesis ordering (Huque and Alosh, 2008)

Parametric chain procedures

Extension of chain procedures for a data-driven or pre-specified testing sequence (Millen and Dmitrienko, 2011)

Feedback procedures

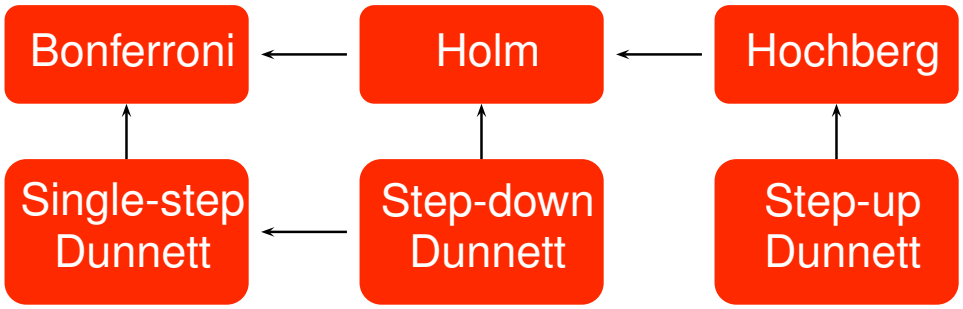
Class of parametric procedures with a pre-specified testing sequence (Zhao, Dmitrienko and Tamura, 2010)

Parametric procedures

Power comparison

Less powerful

More powerful



More powerful

Software implementation

Software implementation in SAS

Custom macros

Single-step and step-down Dunnett procedures: PARPROC macro

Software implementation in R

Mediana package

AdjustPvalues function: Supports commonly used nonparametric, semiparametric and fully parametric procedures

Web site

http://biopharmnet.com/mediana

Example 4: Type 2 diabetes trial

Scenario 5

Comparison	Mean difference	Test statistic
Dose 1 vs Placebo (H_1)	0.63%	2.64
Dose 2 vs Placebo (H_2)	0.46%	1.93
Dose 3 vs Placebo (H_3)	0.55%	2.31

Sample size per group is 90 patients

Dunnett procedures in SAS

Data set

```
data ex4;
    input t;
    datalines;
    2.64
    1.93
    2.31
run;
```

Dunnett procedures in SAS

PARPROC macro

```
%parproc(in=ex4,n=90,out=adjp);
proc print data=adjp noobs label;
    format raw dunnett 6.4;
    var test raw dunnett;
    run;
```

Other options

stepdunnett

Dunnett procedures in SAS

PARPROC macro: Output

Test	Raw	Single-step Dunnett
1	0.0045	0.0118
2	0.0276	0.0677
3	0.0110	0.0283

Dunnett procedures in R

Mediana package (AdjustPvalues function)

```
stat=c(2.64,1.93,2.31)
n=90
p=1-pt(stat, df=2*(n-1))
adjp=AdjustPvalues(p, proc="DunnettAdj",
  par = parameters(n=n))
round(adjp, 4)
```

Output

```
0.0118 0.0677 0.0283
```

Other options

StepDownDunnettAdj

Module F
Simultaneous confidence intervals

Module F outline

F1. Simultaneous confidence intervals

Univariate confidence intervals

Simultaneous confidence intervals for single-step procedures

Simultaneous confidence intervals for stepwise procedures

Simultaneous confidence intervals

EMA guidance (EMA, 2017)

"As can occur with multiple testing, multiple confidence intervals may also increase the chance of false decisions since the probability that a set of multiple non-adjusted confidence intervals cover correctly all parameters to be estimated would usually be less than the pre-specified nominal coverage probability related to the single confidence intervals"

Section F1
Simultaneous confidence intervals

Simultaneous confidence intervals

Univariate confidence intervals

Commonly used in univariate testing problems to help estimate the magnitude of treatment effect

Simultaneous confidence intervals

Used in multiple testing problems and ensure that overall coverage probability is kept at a pre-specified level, e.g., 95%

Play an important role in clinical trials as they facilitate risk/benefit assessments

Multiple testing problem

Parameters

θ_i, $i = 1, \ldots, m$, parameters of interest, e.g., mean difference (continuous endpoints), difference in proportions (binary endpoints) or log-hazard ratio (time-to-event endpoints)

$\widehat{\theta}_i$, $i = 1, \ldots, m$, parameter estimates assumed to be normal (θ_i, σ_i^2)

s_i, $i = 1, \ldots, m$, sample standard errors

Null hypotheses

$H_i: \ \theta_i \leq 0$, $i = 1, \ldots, m$, null hypotheses

α, Type I error rate (one-sided)

Univariate confidence intervals

One-sided confidence intervals

L_i, $i = 1, \ldots, m$, lower confidence limits for θ_i at level $1 - \alpha$

$$L_i = \widehat{\theta}_i - z_\alpha s_i$$

z_x, $(1 - x)$-quantile of the standard normal distribution

Univariate coverage probability

Univariate coverage probability is at least $1 - \alpha$

$$P(L_i \leq \theta_i) \geq 1 - \alpha, \quad i = 1, \ldots, m$$

Overall coverage probability is not controlled

Simultaneous confidence intervals

One-sided confidence intervals

$\tilde{L}_i$, $i = 1, \ldots, m$, lower limits of one-sided simultaneous confidence intervals for θ_i

Overall coverage probability is at least $1 - \alpha$

$$P(\tilde{L}_1 \leq \theta_1, \ldots, \tilde{L}_m \leq \theta_m) \geq 1 - \alpha$$

Consistency

Simultaneous confidence intervals are consistent with decision rules: $\tilde{L}_i \geq 0$ if and only if H_i is rejected, $i = 1, \ldots, m$

Single-step versus stepwise procedures

Single-step procedures

Simultaneous confidence intervals are easy to set up for single-step procedures (Bonferroni and Dunnett procedures)

Stepwise procedures

Constructing simultaneous confidence intervals for stepwise procedures is a challenging task

In general, the more powerful a procedure is, the less meaningful associated simultaneous confidence intervals are

Single-step versus stepwise procedures

FDA guidance (FDA, 2017)

"Multistep procedures are generally more efficient in that they better preserve the power of the tests, but do not readily provide adjusted confidence intervals"

Single-step and stepwise procedures

Single-step procedure

Null hypotheses are tested independently of each other

Simultaneous confidence intervals are defined independently of each other

Step-down procedures

Two-stage algorithm

- Test null hypotheses
- Define simultaneous confidence intervals

Other stepwise procedures

General nonparametric stepwise procedures

Simultaneous confidence intervals can be constructed for a broad family of nonparametric stepwise procedures, including Holm, fallback and chain procedures (Strassburger and Bretz, 2008; Guilbaud, 2008; Guilbaud, 2009)

Semiparametric stepwise procedures

Simultaneous confidence intervals for other stepwise procedures, e.g., Hommel and Hochberg procedures (Guilbaud and Karlsson, 2011; Guilbaud, 2012)

Nonparametric multiple testing procedures

Nonparametric procedures

Bonferroni procedure

Lower limits of one-sided simultaneous confidence intervals for θ_i at level $1-\alpha$

$$\widetilde{L}_i = \widehat{\theta}_i - z_{\alpha/m} s_i, \quad i = 1, \ldots, m$$

Nonparametric procedures

Holm procedure

Case 1: If H_i is rejected and some of the null hypotheses are accepted, $\widetilde{L}_i = 0$

Case 2: If all null hypotheses are rejected, $\widetilde{L}_i = \max(0, \widehat{\theta}_i - z_{\alpha/m} s_i)$

Case 3: If H_i is accepted, $\widetilde{L}_i = \widehat{\theta}_i - z_{\alpha/(m-r)} s_i$, where r is the number of rejected null hypotheses

Properties

In most cases lower confidence limits for rejected null hypotheses are set to 0

Example 4: Type 2 diabetes trial

Scenario 5

Comparison	Mean difference	P-value
Dose 1 vs Placebo (H_1)	0.63%	0.0045
Dose 2 vs Placebo (H_2)	0.46%	0.0276
Dose 3 vs Placebo (H_3)	0.55%	0.0110

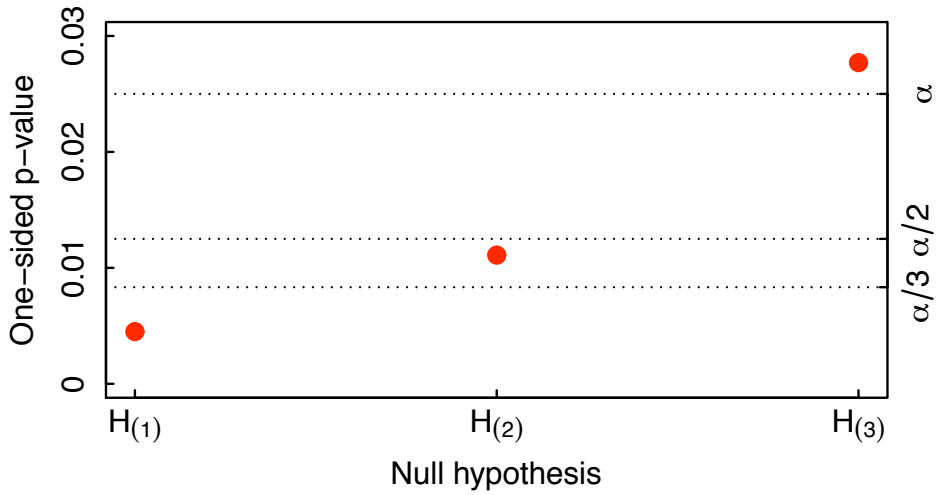
Sample size per group is 90 patients

Pooled standard deviation is 1.6

$H_{(1)} = H_1$, $H_{(2)} = H_3$ and $H_{(3)} = H_2$, ordered null hypotheses

Bonferroni and Holm procedures

Decision rules in Example 4 ($\alpha = 0.025$)

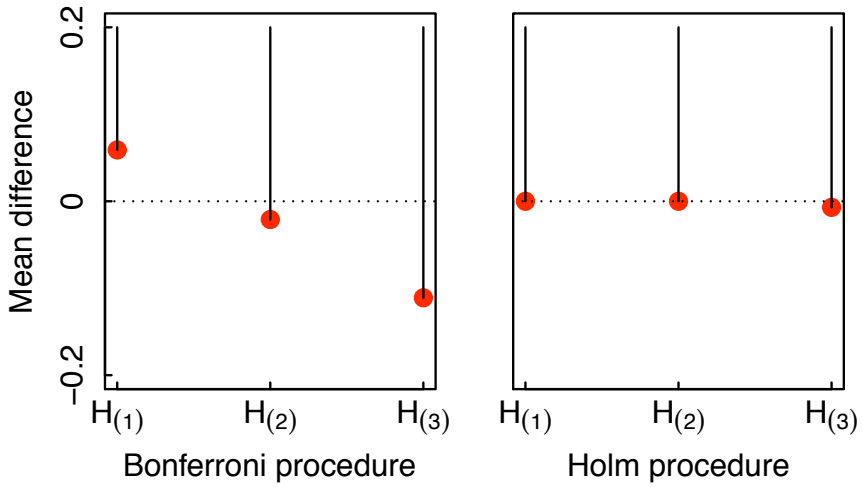


Bonferroni procedure rejects $H_{(1)} = H_1$

Holm procedure rejects $H_{(1)} = H_1$ and $H_{(2)} = H_3$

Bonferroni and Holm procedures

Simultaneous confidence intervals ($\alpha = 0.025$)



Holm procedure: Lower confidence limits for $\theta_{(1)}$ and $\theta_{(2)}$ are set at 0 since $H_{(1)}$ and $H_{(2)}$ are rejected

Simultaneous confidence intervals

EMA guidance (EMA, 2017)

"Informative confidence regions that correspond to multiplicity procedures may, however, not always be available or may be difficult to derive. If the confidence regions do not correspond to the hypothesis testing procedure, different conclusions are possible, e.g. a confidence interval excluding the null hypothesis combined with a non-significant testing result or vice versa."

Simultaneous confidence intervals

EMA guidance (EMA, 2017)

"The decision should, however, be based on the hypothesis test. In that case it is advised to use simple but conservative confidence interval methods, such as Bonferroni-corrected intervals, ensuring that the uncertainty about the beneficial effects is properly understood."

Parametric multiple testing procedures

Parametric testing problem

ANOVA model

Dose-finding trial with multiple dose-control comparisons

$y_{ij} = \mu_i + \varepsilon_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, n$

Parameters

$\theta_i = \mu_i - \mu_0$, $i = 1, \ldots, m$, mean treatment differences

$\widehat{\theta}_i$, $i = 1, \ldots, m$, sample means

s, pooled sample standard error

$\nu = (m + 1)(n - 1)$, degrees of freedom

Parametric stepwise procedures

Dunnett family of parametric procedures

Simultaneous confidence intervals can be constructed for step-down Dunnett procedure (Bofinger, 1987; Stefansson, Kim and Hsu, 1988)

Other procedures

Simultaneous confidence intervals for general stepwise procedures, including step-up Dunnett, parametric fallback and parametric chain procedures, have not been developed yet

Simultaneous confidence intervals

Single-step Dunnett procedure

Lower limit of one-sided simultaneous confidence intervals for θ_i at level $1-\alpha$

$$\widetilde{L}_i = \widehat{\theta}_i - d_\alpha(m,\nu)s, \quad i=1,\ldots,m$$

Notation

$d_x(m,\nu)$, $(1-x)$-quantile of the Dunnett distribution

$c_i = d_\alpha(m-i+1,\nu)$, $i=1,\ldots,m$, critical values of the step-down Dunnett procedure

Simultaneous confidence intervals

Step-down Dunnett procedure

Case 1: If H_i is rejected and some of the null hypotheses are accepted, $\widetilde{L}_i = 0$

Case 2: If all null hypotheses are rejected, $\widetilde{L}_i = \max(0, \widehat{\theta}_i - c_m s)$

Case 3: If H_i is accepted, $\widetilde{L}_i = \widehat{\theta}_i - c_{r+1}s$, where r is the number of rejected null hypotheses

Properties

In most cases lower confidence limits for rejected null hypotheses are set to 0

Example 4: Type 2 diabetes trial

Scenario 5

Comparison	Mean difference	Test statistic
Dose 1 vs Placebo (H_1)	0.63%	2.64
Dose 2 vs Placebo (H_2)	0.46%	1.93
Dose 3 vs Placebo (H_3)	0.55%	2.31

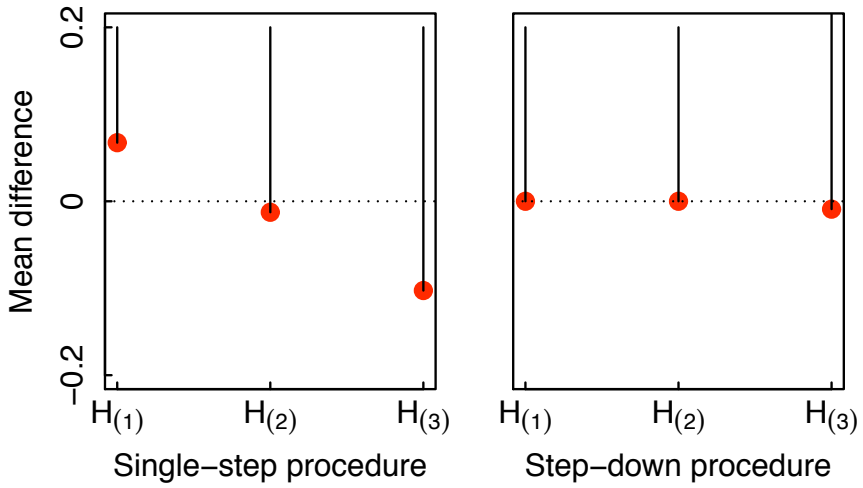
Sample size per group is 90 patients

Pooled standard deviation is 1.6

$H_{(1)} = H_1$, $H_{(2)} = H_3$ and $H_{(3)} = H_2$, ordered null hypotheses

Single-step and step-down procedures

Simultaneous confidence intervals ($\alpha = 0.025$)



Step-down Dunnett procedure: Lower confidence limits for $\theta_{(1)}$ and $\theta_{(2)}$ are set at 0 since $H_{(1)}$ and $H_{(2)}$ are rejected

Software implementation

Software implementation in SAS

Custom macros

PVALCI macro: Simultaneous confidence intervals for Bonferroni, Holm and fixed-sequence procedures

PARCI macro: Simultaneous confidence intervals for single-step and step-down Dunnett procedures

Software implementation in R

Mediana package

AdjustCIs function: Simultaneous confidence intervals for nonparametric procedures (Bonferroni, Holm and fixed-sequence) and parametric procedures (single-step and step-down Dunnett)

Web site

http://biopharmnet.com/mediana

Example 4: Type 2 diabetes trial

Scenario 5

Comparison	Mean difference	Test statistic
Dose 1 vs Placebo (H_1)	0.63%	2.64
Dose 2 vs Placebo (H_2)	0.46%	1.93
Dose 3 vs Placebo (H_3)	0.55%	2.31

Sample size per group is 90 patients

Pooled standard deviation is 1.6

Nonparametric procedures in SAS

Data set

```
data ex4;
   input raw_p est sd;
   weight=1/3;
   se=sd*sqrt(2/90);
   datalines;
   0.0045 0.63 1.6
   0.0276 0.46 1.6
   0.0110 0.55 1.6
run;
```


Nonparametric procedures in SAS

PVALCI macro

```
%pvalci(in=ex4,covprob=0.975,
        out=adjci);
proc print data=adjci noobs label;
   format univariate holm 6.3;
   var test univariate holm;
   run;
```

Other options

bonferroni fixedseq fallback

Nonparametric procedures in SAS

PVALCI macro: Output

```
Test   Univariate      Holm
  1        0.163        0.000
  2       -0.007       -0.007
  3        0.083        0.000
```


Parametric procedures in SAS

Data set

```
data ex4;
   input t est sd;
   se=sd*sqrt(2/90);
   datalines;
   2.64 0.63 1.6
   1.93 0.46 1.6
   2.31 0.55 1.6
run;
```

Parametric procedures in SAS

PARCI macro

```
%parci(in=ex4,n=90,covprob=0.975,
       out=adjci);
proc print data=adjci noobs label;
   format univariate dunnett 6.3;
   var test univariate dunnett;
   run;
```

Other options

stepdunnett

Parametric procedures in SAS

PARCI macro: Output

Test	Univariate	Single-step Dunnett
1	0.159	0.067
2	-0.011	-0.103
3	0.079	-0.013

Nonparametric procedures in R

Mediana package (AdjustCIs function)

```
est=c(0.63,0.46,0.55)
ci=AdjustCIs(est, proc="HolmAdj",
  par=parameters(sd=rep(1.6, 3), n=90,
  covprob=0.975))
round(ci, 3)
```

Output

```
0.000 -0.007 0.000
```

Other options

BonferroniAdj DunnettAdj StepDownDunnettAdj

Module G
Power calculations

Module G outline

G1. Power calculations in clinical trials with multiple objectives

Analytical and simulation-based power calculation approaches

Introduction to Clinical Scenario Evaluation

Mediana package: New R package for clinical trial simulations

Section G1
Power calculations in clinical trials with multiple objectives

Analytical power and sample size calculations

Key features

Analytical power and sample size calculations rely on closed-form expressions

Example

Sample size per treatment arm

$$n = \frac{2(z_\alpha + z_\beta)^2 \delta^2}{\sigma^2}$$

where δ is the true mean difference and σ is the true common standard deviation

Simple scenario

Assumptions

Two-arm design

Single normally distributed endpoint

Balanced design

No stratification factors

Realistic scenarios

New trends in Phase II and III trials

Multiple arms (multiple doses or active control arms)

Multiple clinical endpoints

Multiple patient populations (e.g., overall population and marker-positive subpopulation)

Interim looks and mid-course adaptations

Power and sample size calculations

Analytical approach

Closed-form expressions used in traditional sample size calculations often rely on simplifying/artificial assumptions

Are we cutting corners and looking for shortcuts in multi-million dollar clinical trials?

Simulation-based approach

Much more reliable approach to power and sample size calculations in trials with complex design and analysis strategies

Power calculations

EMA guidance (EMA, 2017)

"Sometimes a series of related objectives is pursued in the same trial, each with its own primary variable... In these situations planning of the sample size becomes more complex due to the different alternative hypotheses related to the different endpoints and due to the assumed correlation between endpoints."

Simulation-based approaches

FDA guidance (FDA, 2017)

"Determination of an appropriate study sample size to ensure that the study is appropriately powered can be difficult in these cases, and often will be dependent upon computer simulations rather than an analytic formula, which can be used for simpler situations"

Simulation-based approaches

Key features

Simulation-based approaches free clinical trial sponsors from artificial restrictions

Help provide answers to complex clinically important questions in trials with multiple objectives

Analogy

Probit regression versus logistic regression

Introduction to Clinical Scenario Evaluation

Clinical scenario evaluation

General framework

Clinical scenario evaluation (CSE) approach was developed in Benda et al. (2010), Friede et al. (2010) and other publications

Motivation

Clinical trial researchers have recognized the importance of employing quantitative, comprehensive and disciplined approaches to evaluating the design and analysis of clinical trials to enable better decision making

Clinical scenario evaluation

Goals

Critically evaluate the statistical assumptions, candidate trial designs and candidate analysis strategies

CSE framework supports evidence-based approaches to designing clinical trials and selecting analysis methods

Clinical scenario evaluation

Structured approach

Supports “workflows” that mimic the process of collecting, analyzing and evaluating clinical trial data

Break down a complex problem of clinical trial evaluation to key components (data model, analysis model and evaluation model)

Examine individual or synergistic effects of multiple design and analysis parameters

Key components

Data models (Assumptions)

Describe the data generation mechanism in a clinical trial

Analysis models (Options)

Define statistical tests, descriptive statistics and other analysis tools, e.g., multiplicity adjustments, computed from the trial data

Evaluation models (Metrics)

Specify measures for evaluating performance of the analysis strategies

Clinical trial optimization

Clinical scenario evaluation

An important application of general CSE approach is clinical trial optimization

General theme

Utilize CSE to transition from traditionally used approaches to optimal approaches to selecting trial designs and analysis strategies

Inform decision making at different stages of a drug development program to maximize the overall probability of success

Examples of clinical trial optimization

Phase II and Phase III trials

Optimal multiplicity adjustment in clinical trials with several objectives

Optimal decision rules at interim looks in clinical trials with adaptive designs

Optimal decision rules in clinical trials with several patient populations

Clinical trial optimization

Publications

Review of general approaches to clinical trial optimization (*Clinical Trial Optimization Using R* edited by Dmitrienko and Pulkstenis, 2017)

Optimal selection of multiplicity adjustments in Phase III trials (Dmitrienko, Paux and Brechenmacher, 2015)

Optimal selection of multiplicity adjustments and adaptive trial designs in Phase II and III trials (Dmitrienko, Paux, Pulkstenis and Zhang, 2016)

Mediana package

Mediana: New R package

Goals

Implement the Clinical Scenario Evaluation approach

Provide general framework for simulation-based power and sample size calculations typically performed in late-phase trials (Phase II and III trials)

Support clinical trial optimization aimed at identifying optimal trial designs and analysis strategies

Data model

Endpoint types

Single trial endpoint (univariate distributions for continuous, binary, time-to-event and count endpoints)

Multiple trial endpoints (multivariate distributions)

Trial design options

Patient enrollment and dropout modeling

Analysis model

Statistical tests

Commonly used statistical tests, including parametric tests, nonparametric tests and model-based analysis methods

Descriptive statistics

Commonly used descriptive statistics

Multiplicity adjustments

Popular multiplicity adjustments (traditional adjustments and advanced adjustments such as gatekeeping procedures)

Evaluation model

Evaluation criteria

Broadly used definitions of “probability of successful outcome”

Examples

Marginal power, disjunctive and conjunctive power

Metrics based on statistical and clinical significance

Trial designs

Fixed trial designs

Trial designs with a pre-defined sample size

Event-driven trial designs

Trial designs with time-to-event (e.g., survival-type) endpoints

Sample size is selected to achieve the target number of events

High-performance computing

Traditional approach: Sequential computations

Simulations are run sequentially on one processor (core)

Advanced approach: Parallel computations

Implemented in Mediana package and supports simulations that are distributed among multiple processors (cores)

Substantially reduces computation times

Mediana package

Release

First version (Version 1.0.1) was released in July 2015

CRAN web site

https://cran.r-project.org/web/packages/Mediana

Online manual

http://gpaux.github.io/Mediana/

Example 4

Clinical trial with multiple doses

Example 4: Type 2 diabetes trial

Characterize efficacy profile of three doses of an experimental treatment (Dose 1, Dose 2 and Dose 3) versus placebo

Three dose-placebo comparisons

Goal

Compute the number of patients to guarantee a sufficiently high level of overall success probability

Clinical trial with multiple doses

Naive approach

Bonferroni adjustment is applied to control overall Type I error rate

Sample size formula

Apply Bonferroni adjustment to α level: How meaningful is the result?

$$n = \frac{2(z_{\alpha/3} + z_{\beta})^2 \delta^2}{\sigma^2}$$

Clinical trial with multiple doses

Assumptions

$\alpha = 0.025$, One-sided Type I error rate

$\beta = 0.2$, One-sided Type II error rate (80% power)

$\delta/\sigma = 0.3$, Common effect size across the three dose-placebo comparisons

Clinical trial with multiple doses

Standard sample size formula

Sample size formula without adjustment: 175 patients per arm

Naive sample size formula

Sample size formula with the naive Bonferroni-type adjustment: 233 patients per arm (33% increase)

Clinical trial with multiple doses

Naive sample size formula

Sample size: 233 patients per arm

Marginal power

Marginal power of each dose-placebo comparison (probability each individual dose-placebo comparison is significant): 80%

Disjunctive power

Disjunctive power (probability that at least one dose-placebo comparison is significant): 95%

Clinical trial with multiple doses

Simulation-based evaluation

Sample size: 145 patients per arm

Marginal power

Marginal power of each dose-placebo comparison (probability each individual dose-placebo comparison is significant): 56%

Disjunctive power

Disjunctive power (probability that at least one dose-placebo comparison is significant): 80%

Clinical trial with multiple doses

Naive approach to sample size calculations

Sample size formula with the naive Bonferroni-type adjustment resulted in overestimating the probability of success in the trial (disjunctive power)

Recommended approach

Simulation-based approach to compute disjunctive power for a range of sample sizes

Simulation-based approach will be illustrated using the Mediana package

Data model

Key components

Outcome distribution

Samples (independent sets of patients, e.g., treatment arm in a trial)

Parameters of individual samples (outcome distribution parameters and sample sizes)

Trial design

Data model

Four samples

Sample 1: Placebo arm

Samples 2, 3 and 4: Dose 1 arm, Dose 2 arm and Dose 3 arm

Outcome distribution

Primary endpoint is normally distributed (*NormalDist*)

Trial design

Fixed design

Data model

Outcome distribution parameters

Sample	ID	Mean	SD
Sample 1	Placebo	0	1
Sample 2	Dose 1	0.3	1
Sample 3	Dose 2	0.3	1
Sample 4	Dose 3	0.3	1

Mediana code

Data model: Outcome distribution parameters

```
outcome.placebo = parameters(mean = 0, sd = 1)
outcome.dose1 = parameters(mean = 0.3, sd = 1)
outcome.dose2 = parameters(mean = 0.3, sd = 1)
outcome.dose3 = parameters(mean = 0.3, sd = 1)
```


Mediana code

Data model

```
ex4.data.model = DataModel() +
  OutcomeDist(outcome.dist = "NormalDist") +
  SampleSize(seq(130, 150, 5)) +
  Sample(id = "Placebo",
    outcome.par = parameters(outcome.placebo)) +
  Sample(id = "Dose 1",
    outcome.par = parameters(outcome.dose1)) +
  Sample(id = "Dose 2",
    outcome.par = parameters(outcome.dose2)) +
  Sample(id = "Dose 3",
    outcome.par = parameters(outcome.dose3))
```

Analysis and evaluation models

Key components of analysis model

Dose-placebo tests and their parameters, including a multiplicity adjustment

Key components of evaluation model

Success criteria and their parameters

Analysis and evaluation models

Analysis model

Treatment effect at each dose is assessed using the two-sample t test (*TTest*)

Multiplicity adjustment (*BonferroniAdj*)

Other adjustments could be used (*HochbergAdj* or *DunnettAdj*)

Evaluation model

Marginal power for each dose-placebo test (*MarginalPower*)

Disjunctive power (*DisjunctivePower*)

Mediana code

Analysis model

```
ex4.analysis.model = AnalysisModel() +
  MultAdjProc(proc = "BonferroniAdj") +
  Test(id = "Placebo vs Dose 1",
    samples = samples("Placebo", "Dose 1"),
    method = "TTest") +
  Test(id = "Placebo vs Dose 2",
    samples = samples ("Placebo", "Dose 2"),
    method = "TTest") +
  Test(id = "Placebo vs Dose 3",
    samples = samples("Placebo", "Dose 3"),
    method = "TTest")
```


Mediana code

Evaluation model

```
ex4.evaluation.model = EvaluationModel() +
  Criterion(id = "Marginal power",
    method = "MarginalPower",
    tests = tests("Placebo vs Dose 1",
                  "Placebo vs Dose 2",
                  "Placebo vs Dose 3"),
    labels = c("Placebo vs Dose 1",
               "Placebo vs Dose 2",
               "Placebo vs Dose 3"),
    par = parameters(alpha = 0.025)) +
```

Mediana code

Evaluation model (continued)

```
Criterion(id = "Disjunctive power",
  method = "DisjunctivePower",
  tests = tests("Placebo vs Dose 1",
                "Placebo vs Dose 2",
                "Placebo vs Dose 3"),
  labels = "Disjunctive power",
  par = parameters(alpha = 0.025))
```


Mediana code

Run simulations

```
ex4.sim.parameters =
    SimParameters(n.sims = 10000,
                  proc.load = "full",
                  seed = 42938001)

ex4.results = CSE(ex4.data.model,
                  ex4.analysis.model,
                  ex4.evaluation.model,
                  ex4.sim.parameters)

summary(ex4.results)
```

Example 4: Type 2 diabetes trial

Simulation results

Sample size per arm	Power	Value
140	Marginal	54.5%
	Disjunctive	78.7%
145	Marginal	56.2%
	Disjunctive	79.9%
150	Marginal	57.8%
	Disjunctive	81.3%

Example 5

Clinical trial with multiple patient populations

Non-small-cell lung cancer trial

Characterize efficacy profile of an experimental treatment versus placebo in two patient populations

Two tests (overall trial population and marker-positive subpopulation)

Goal

Compute the number of events to guarantee a sufficiently high level of overall success probability

Data model

Four samples

Sample 1: Placebo arm (marker-negative patients)

Sample 2: Placebo arm (marker-positive patients)

Sample 3: Treatment arm (marker-negative patients)

Sample 4: Treatment arm (marker-positive patients)

Prevalence of marker-positive patients

Prevalence: 60%

Data model

Outcome distribution

Primary endpoint follows an exponential distribution

Trial design

Event-driven design without censoring

Every patient is followed until disease progression (number of events is equal to number of patients)

Data model

Outcome distribution parameters

Sample	Median PFS	Rate parameter
Sample 1	11 months	$\log 2/11 = 0.063$
Sample 2	11 months	$\log 2/11 = 0.063$
Sample 3	12.5 months	$\log 2/12.5 = 0.055$
Sample 4	15 months	$\log 2/15 = 0.046$

Note: Rate parameter of the exponential distribution

Mediana code

Data model: Outcome distribution parameters

```
outcome.pn = parameters(rate = log(2) / 11)
outcome.pp = parameters(rate = log(2) / 11)
outcome.tn = parameters(rate = log(2) / 12.5)
outcome.tp = parameters(rate = log(2) / 15)
```


Mediana code

Data model: Sample sizes

```
sample.size.total = c(530)
sample.size.pn = as.list(0.2 * sample.size.total)
sample.size.pp = as.list(0.3 * sample.size.total)
sample.size.tn = as.list(0.2 * sample.size.total)
sample.size.tp = as.list(0.3 * sample.size.total)
```

Mediana code

Data model

```
ex5.data.model = DataModel() +
  OutcomeDist(outcome.dist = "ExpoDist") +
  Sample(id = "Placebo (marker-negative)",
         sample.size = sample.size.pn,
         outcome.par = parameters(outcome.pn)) +
  Sample(id = "Placebo (marker-positive)",
         sample.size = sample.size.pp,
         outcome.par = parameters(outcome.pp)) +
  Sample(id = "Treatment (marker-negative)",
         sample.size = sample.size.tn,
         outcome.par = parameters(outcome.tn)) +
  Sample(id = "Treatment (marker-positive)",
         sample.size = sample.size.tp,
         outcome.par = parameters(outcome.tp))
```


Analysis and evaluation models

Key components of analysis model

Treatment effect tests in two populations and their parameters, including a multiplicity adjustment

Key components of evaluation model

Success criteria and their parameters

Analysis and evaluation models

Analysis model

Treatment effect in each population is assessed using the log-rank test (*LogrankTest*)

Multiplicity adjustment (*HochbergAdj*)

Evaluation model

Marginal power for each population test (*MarginalPower*)

Disjunctive power (*DisjunctivePower*)

Mediana code

Analysis model

```
ex5.analysis.model = AnalysisModel() +
  MultAdjProc(proc = "HochbergAdj") +
  Test(id = "Overall population test",
   samples = samples(c("Placebo (marker-negative)",
                       "Placebo (marker-positive)"),
                     c("Treatment (marker-negative)",
                       "Treatment (marker-positive)")),
   method = "LogrankTest") +
  Test(id = "Marker-positive subpopulation test",
   samples = samples("Placebo (marker-positive)",
                     "Treatment (marker-positive)"),
   method = "LogrankTest")
```

Mediana code

Evaluation model

```
ex5.evaluation.model = EvaluationModel() +
  Criterion(id = "Marginal power",
    method = "MarginalPower",
    tests = tests("Overall population test",
                  "Marker-positive subpopulation
                     test"),
    labels = c("Overall population test",
               "Marker-positive subpopulation test"),
    par = parameters(alpha = 0.025)) +
```


Mediana code

Evaluation model (continued)

```
  Criterion(id = "Disjunctive power",
    method = "DisjunctivePower",
    tests = tests("Overall population test",
                  "Marker-positive subpopulation
                     test"),
    labels = "Disjunctive power",
    par = parameters(alpha = 0.025))
```

Mediana code

Run simulations

```
ex5.sim.parameters =
    SimParameters(n.sims = 10000,
                  proc.load = "full",
                  seed = 42938001)

ex5.results = CSE(ex5.data.model,
                  ex5.analysis.model,
                  ex5.evaluation.model,
                  ex5.sim.parameters)

summary(ex5.results)
```


Example 5: Non-small-cell lung cancer trial

Simulation results

Total number of events	Power	Value
530	Marginal in overall population	75.8%
	Marginal in subpopulation	75.6%
	Disjunctive	80.2%

Mediana package

Mediana package

Simulation report

Microsoft Word document that provides a detailed summary of assumptions and simulation results

Presentation model

User needs to define a presentation model to customize the report's structure (create sections and subsections, specify how the rows will be sorted within tables)

Mediana package

More information

Online manual at http://gpaux.github.io/Mediana

Case studies

Multiple case studies to illustrate CSE and clinical trial simulations in numerous settings at http://gpaux.github.io/Mediana/CaseStudies.html

Other Mediana code

Mediana code from *Clinical Trial Optimization Using R* (edited by Dmitrienko and Pulkstenis, 2017)

Key Multiplicity Issues in Clinical Trials (Part I)

Alex Dmitrienko (Mediana Inc)
admitrienko@medianainc.com