Multiple Testing Procedures in Clinical Trials

Alex Dmitrienko
alex.dmitrienko@quintiles.com

IBS workshop, Berlin, Sep 19-20, 2013

Multiplicity issues 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

CPMP: Points to consider on multiplicity issues in clinical trials (released on Sep 19, 2002)

U.S. FDA: Guidance document on multiplicity issues in clinical trials (expected to be released in late 2013)

http://www.multxpert.com/wiki/Regulatory_Guidance

Key concepts

Part I: Traditional multiplicity problems

Clinical trials with equally important objectives

Single source of multiplicity

Part II: Advanced multiplicity problems

Clinical trials with ordered objectives

Multiple sources of multiplicity

Part I outline

Traditional multiplicity problems

Module A: Problem formulation and clinical trial examples

Module B: Nonparametric procedures

Module C: Semiparametric procedures

Module D: Parametric procedures

Module E: Simultaneous confidence intervals

Module F: Sample size calculations

Part II outline

Advanced multiplicity problems

Module G: Introduction to gatekeeping procedures

Module H: Problems with serial and parallel gatekeepers

Module I: Problems with general gatekeepers

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

Web site

Multiplicity Expert web site

http://multxpert.com/wiki/Short_Courses

Supplementary materials

SAS and R code

References

Useful links

Online training

Instant Training web site

http://www.sprmm.com/

Online conferences and courses

Available 24 hours a day/7 days a week anywhere in the world

Full-day courses on multiple comparisons

Free conference on key multiplicity issues in confirmatory clinical trials

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

Part I
Traditional Multiplicity Problems

Module A
Problem Formulation
and Clinical Trial Examples

Outline

1. Clinical trial examples

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

2. Inferential goals

Different analyses are treated as independent entities or as components of a single overall analysis

At-least-one procedures (multiple testing procedures), all-or-none procedures and global procedures

Outline

3. Error rate definitions for at-least-one procedures

Familywise error rate

4. Closure principle

Method for constructing powerful multiple testing procedures

5. Selection of multiple testing procedures

Criteria used to select an optimal multiple testing (at-least-one) procedure

Classification of multiple testing procedures

1. Clinical trial examples

Clinical trial examples

Multiple endpoints

Example 1: Osteoporosis/breast cancer trial

Example 2: Alzheimer’s disease trial

Example 3: Fracture healing trial

Multiple doses

Example 4: Major depressive disorder trial

Multiple populations

Example 5: Schizophrenia trial

Clinical trial with multiple endpoints

Example 1: Osteoporosis/breast cancer trial

Objective

Evaluate the effects of a treatment on the risk of new vertebral fractures and incidence of invasive breast cancer in postmenopausal women with osteoporosis

Design

Treatment versus placebo

Clinical trial with multiple endpoints

Example 1: Osteoporosis/breast cancer trial

Two endpoints

Endpoint 1: Incidence of vertebral fractures

Endpoint 2: Incidence of breast cancer

Overall analysis

Treatment effect on at least one endpoint must be significant

Clinical trial with multiple endpoints

Example 2: Alzheimer’s disease trial

Objective

Evaluate the effects of a treatment on cognition and global changes in patients with mild to moderate Alzheimer’s disease

Design

Treatment versus placebo

Clinical trial with multiple endpoints

Example 2: Alzheimer's disease trial

Two endpoints

Endpoint 1: Cognition endpoint (ADAS-Cog)

Endpoint 2: Clinical global scale (CIBIC plus)

Overall analysis

Treatment effect on both endpoints must be significant

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

Clinical trial with multiple doses

Example 4: Major depressive disorder trial

Objective

Evaluate the effects of a treatment on depressive symptoms in patients with major depressive disorder

Primary endpoint

Montgomery-Asberg Depression Rating Scale (MADRS) total score

Design

Three dose groups versus placebo

Clinical trial with multiple patient populations

Example 5: Schizophrenia trial

Objective

Evaluate the efficacy of a treatment in patients diagnosed with schizophrenia

Primary endpoint

Positive and Negative Symptoms Scale (PANSS) total score

Design

Treatment versus placebo

Clinical trial with multiple patient populations

Example 5: Schizophrenia trial

Tailored therapy approach is implemented in this trial

Three patient populations

General population

Subpopulation 1: Females

Subpopulation 2: Based on a genotypic classifier

2. Inferential goals

Inferential goals

Multiple testing problem

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

Notation

δ_i, $i = 1, \ldots, m$, measures of treatment effect

λ, pre-specified clinically relevant threshold

$H_i: \ \delta_i \leq \lambda$, null hypothesis of no effect (non-inferiority or superiority)

$K_i: \ \delta_i > \lambda$, alternative hypothesis of therapeutic benefit

Inferential goals

Three inferential goals

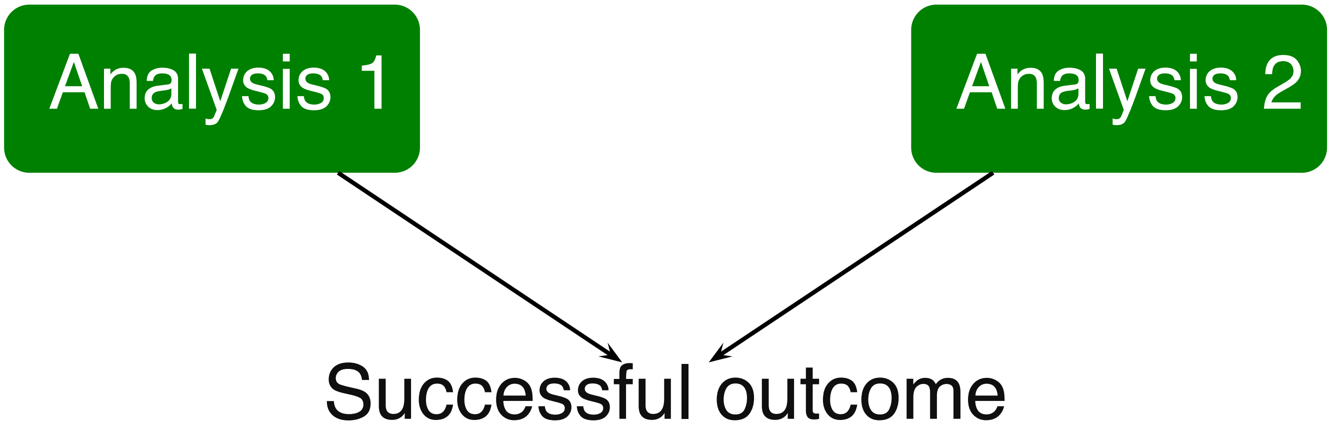
Individual analyses separately lead to a successful outcome (at-least-one procedures, also known as multiple testing procedures)

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



At-least-one procedures

Each analysis is independently clinically relevant

Each endpoint, dose or population analysis independently provides a proof of efficacy

The trial’s outcome is declared positive if at least one analysis produces a significant result

Examples

Example 1: Osteoporosis/breast trial

Example 4: Major depressive disorder trial

Example 5: Schizophrenia trial

At-least-one procedures

Inferential goal

Global null hypothesis

$$H_I = \bigcap_{i=1}^{m} \{\delta_i \leq 0\}$$

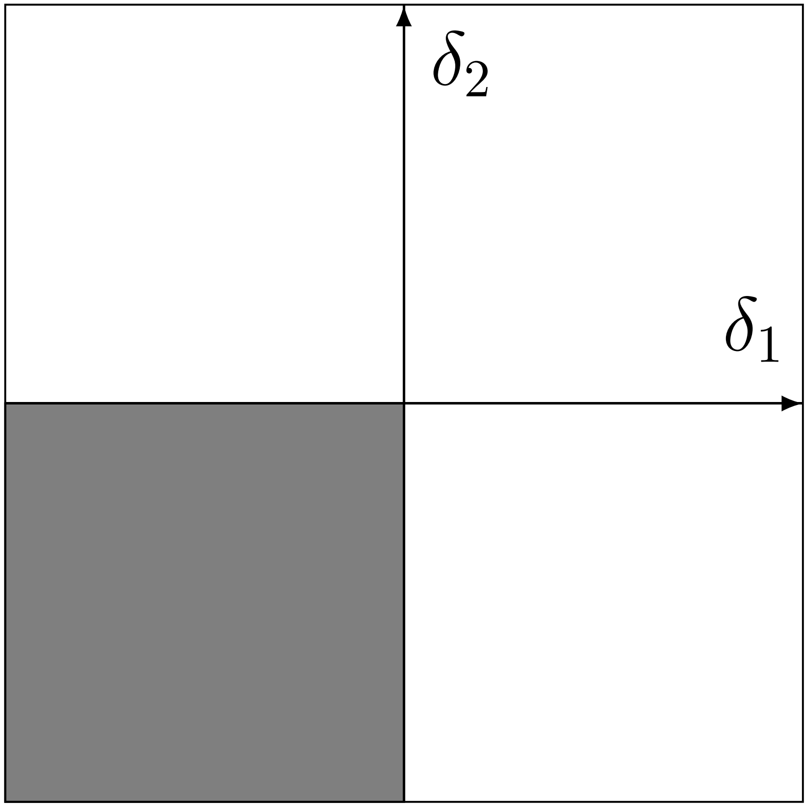
is rejected if one or more null hypotheses are shown to be false

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

At-least-one procedures (multiple testing procedures) are discussed in Modules B, C and D

Example 1: Osteoporosis/breast cancer trial

Global null hypothesis (shaded region)

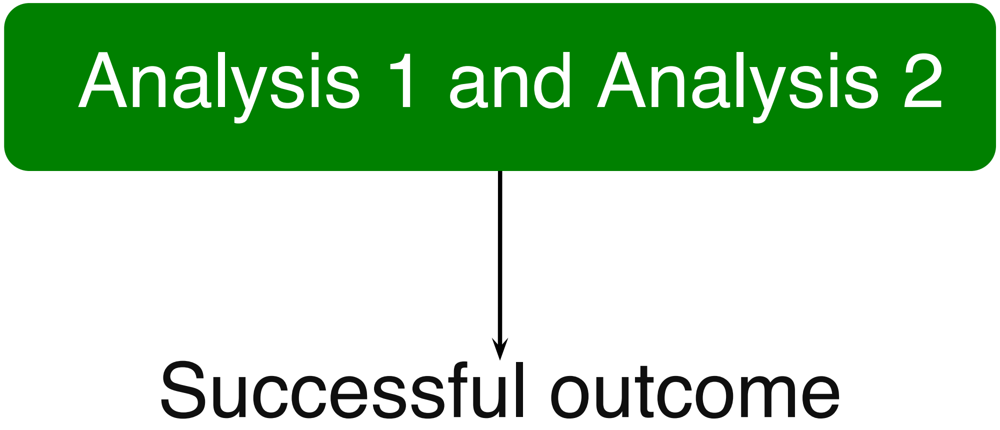


δ_1, treatment difference in incidence of vertebral fractures

δ_2, treatment difference in incidence of breast cancer

All-or-none procedures

All analyses must show benefit



Analysis 1 and Analysis 2

Successful outcome

All-or-none procedures

All analyses must show benefit

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

Example

Example 2: Alzheimer's disease trial

All-or-none procedures

Inferential goal

Global null hypothesis

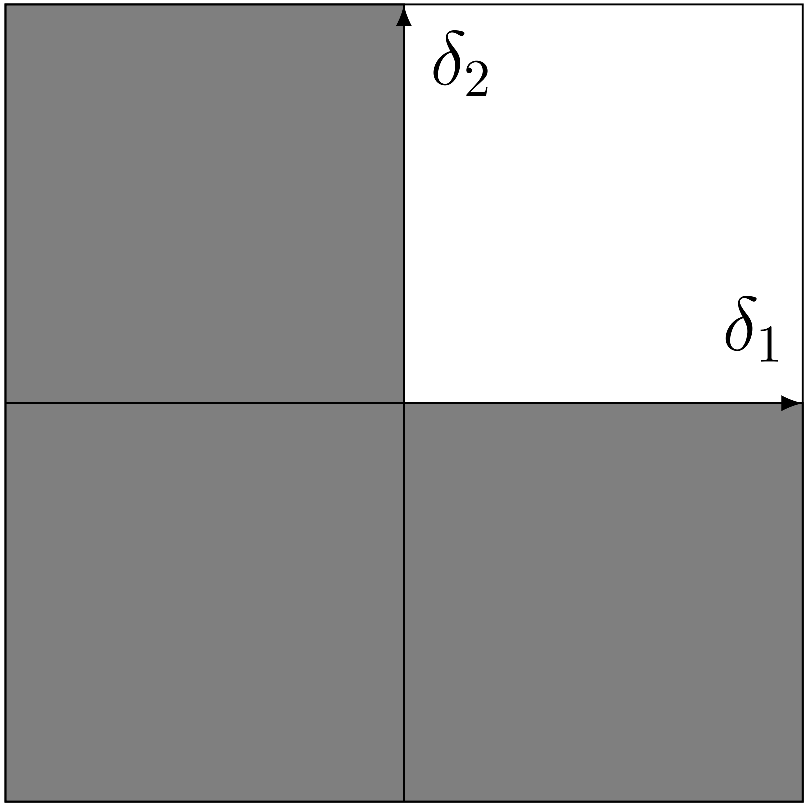
$$H_U = \bigcup_{i=1}^{m} \{\delta_i \le 0\}$$

is rejected if all null hypotheses are shown to be false

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

Example 2: Alzheimer's disease trial

Global null hypothesis (shaded region)



δ_1, treatment difference for cognition endpoint (ADAS-Cog)

δ_2, treatment difference for clinical global scale (CIBIC plus)

All-or-none procedures

Notation

$p_1, \ldots, p_m$, p-values for null hypotheses $H_1, \ldots, H_m$

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

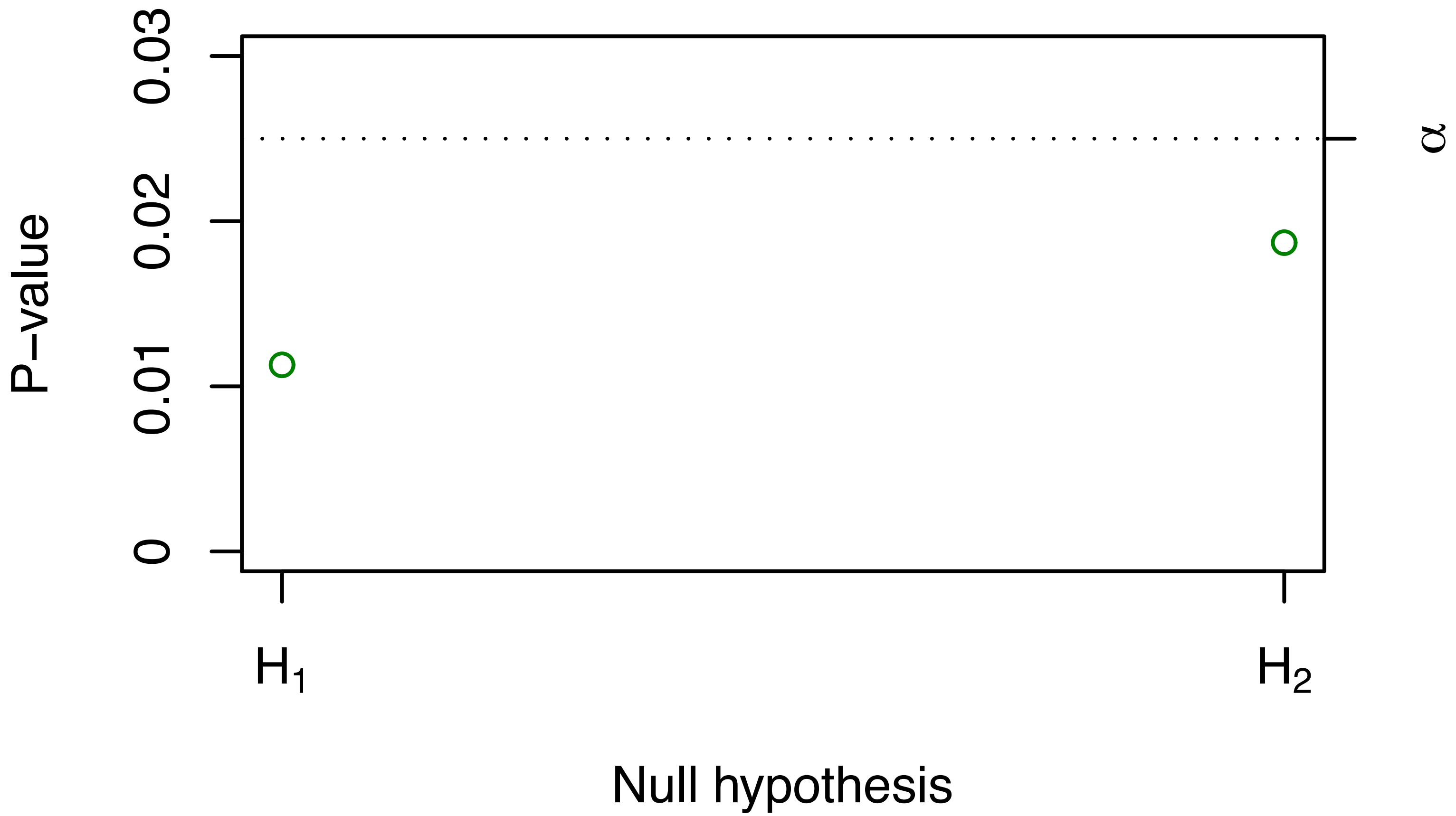
Intersection-union problem

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

This problem will not be discussed further in this course

Example 2: Alzheimer's disease trial

Decision rule for $\alpha = 0.025$

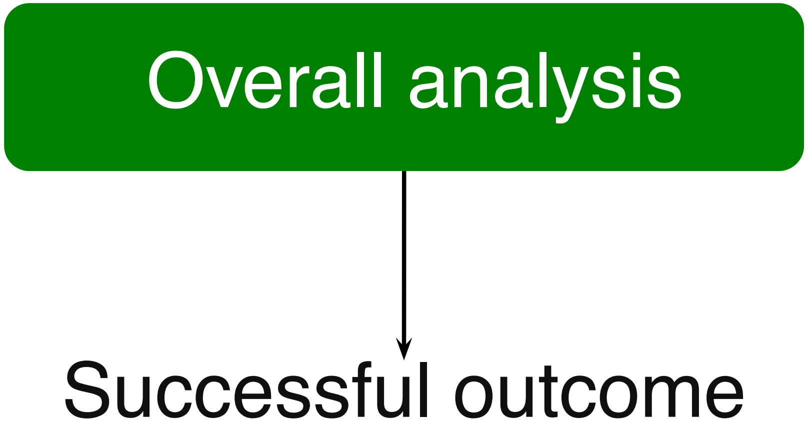


$p_1 = 0.0113$ and $p_2 = 0.0187$

All-or-none procedure rejects H_1 and H_2

Global procedures

Individual analyses are components of an overall analysis



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

Example

Example 3: Fracture healing trial

Global procedures

Inferential goal

Null hypothesis

$$H = \{\delta \leq 0\}$$

is rejected if there is evidence of an overall treatment effect (δ is a measure of treatment effect across the analyses)

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

3. Error rate definitions

Error rate definitions

At-least-one procedures (multiple testing procedures)

To choose an appropriate multiple 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 clinical trials

Marginal and simultaneous tests

Marginal tests

Consider the null hypotheses $H_1, \ldots, H_m$

Each null hypothesis is tested at α level

Probability of an incorrect decision (incorrect rejection of a single null hypothesis) is α

Simultaneous testing procedure

How to define the probability of an incorrect decision when the null hypotheses are tested simultaneously?

Familywise error rate

Key assumption

Null hypotheses may be true (no treatment effect) or may be false (treatment effect is present)

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: Major depressive disorder trial

Three null hypotheses H_1, H_2 and H_3

Consider all combinations of true null hypotheses and show that FWER $\leq \alpha$ for any combination

For instance, 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 (strong sense)

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 Modules B, C and D provide FWER control in the strong sense

Regulatory position

CPMP guidance document

Points to consider on multiplicity issues in clinical trials (CPMP/EWP/908/99, Sep 19, 2002)

“A clinical study that requires no adjustment of the Type I error is one that consists of two treatment groups, that uses a single primary variable, and has a confirmatory statistical strategy that prespecifies just one single null hypothesis relating to the primary variable and no interim analysis”

4. Closure principle

Closure principle

Powerful tool for building multiple testing procedures

Closure principle (Marcus, Peritz and Gabriel, 1976) provides a foundation for virtually all multiple testing methods used in clinical trial applications

Key method for constructing powerful stepwise procedures, e.g., Holm, Hochberg, Hommel, fixed-sequence, fallback, chain and other procedures

Closure principle

Example 1: Osteoporosis/breast cancer trial

$H_1: \ \delta_1 \leq 0$ (δ_1 is the treatment difference in the incidence of vertebral fractures)

$H_2: \ \delta_2 \leq 0$ (δ_2 is the treatment difference in the incidence of breast cancer)

Closed testing procedure

Set up a procedure for testing H_1 and H_2 which controls FWER at α level, e.g., $\alpha = 0.025$

Closure principle

Available options

Bonferroni procedure: Widely used but conservative

More powerful procedure: Can be constructed using the closure principle

Bonferroni procedure

Decision rule

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

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

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

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

Bonferroni procedure

Decision rules in Example 1: Osteoporosis/breast cancer trial

Reject H_1 if $p_1 \leq \alpha/2$

Reject H_2 if $p_2 \leq \alpha/2$

More powerful procedure

Closed testing procedure

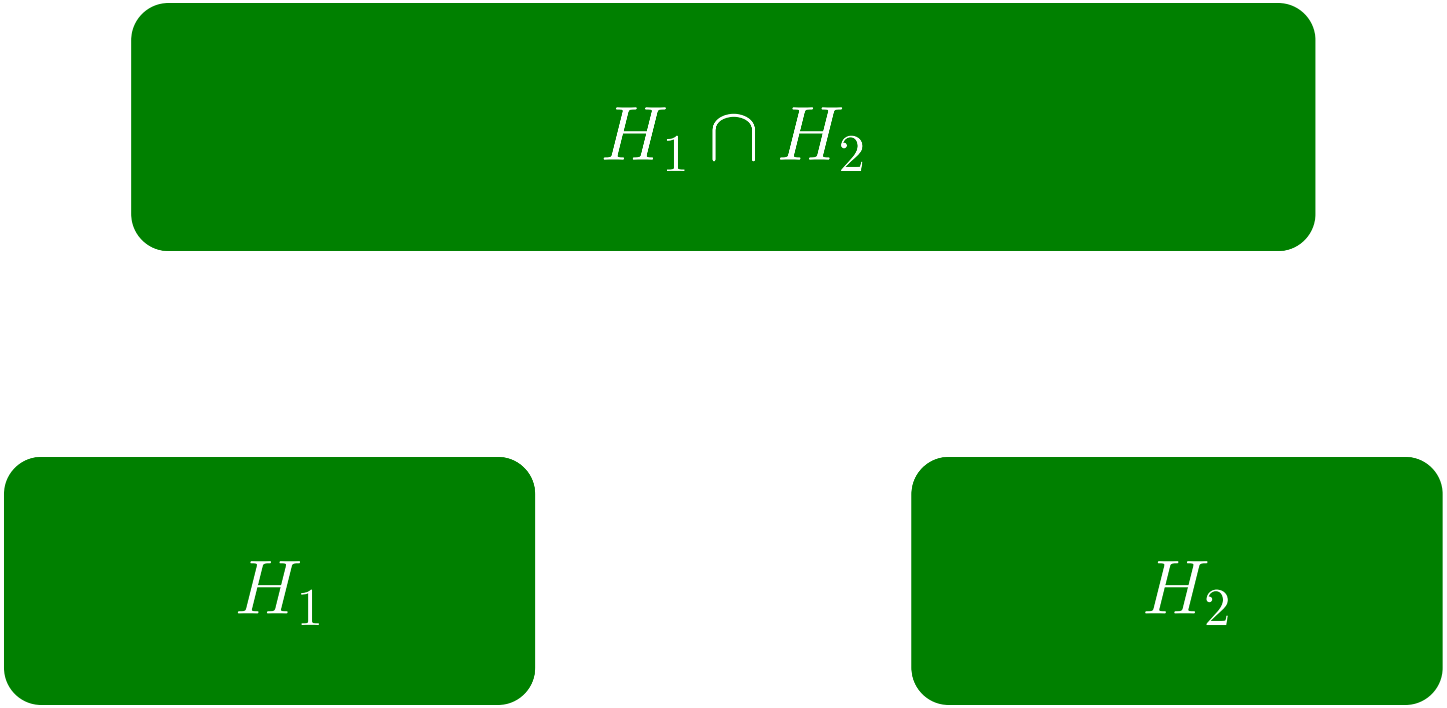
Step 1: Define closed family which includes all possible intersections of H_1 and H_2 (three intersection hypotheses)

Step 2: Define α-level local tests for all intersection hypotheses

Step 3: Define decision rules: Reject a null hypothesis if all intersection hypotheses containing this null hypothesis are rejected by local tests

Example 1: Osteoporosis/breast cancer trial

Step 1: Define closed family



$H_1 \cap H_2$

H_1

H_2

All possible intersections of null hypotheses H_1 and H_2

Example 1: Osteoporosis/breast cancer trial

Step 2: Define local tests

Reject $H_1 \cap H_2$
if $p_1 \leq \alpha/2$ or $p_2 \leq \alpha/2$

Reject H_1
if $p_1 \leq \alpha$

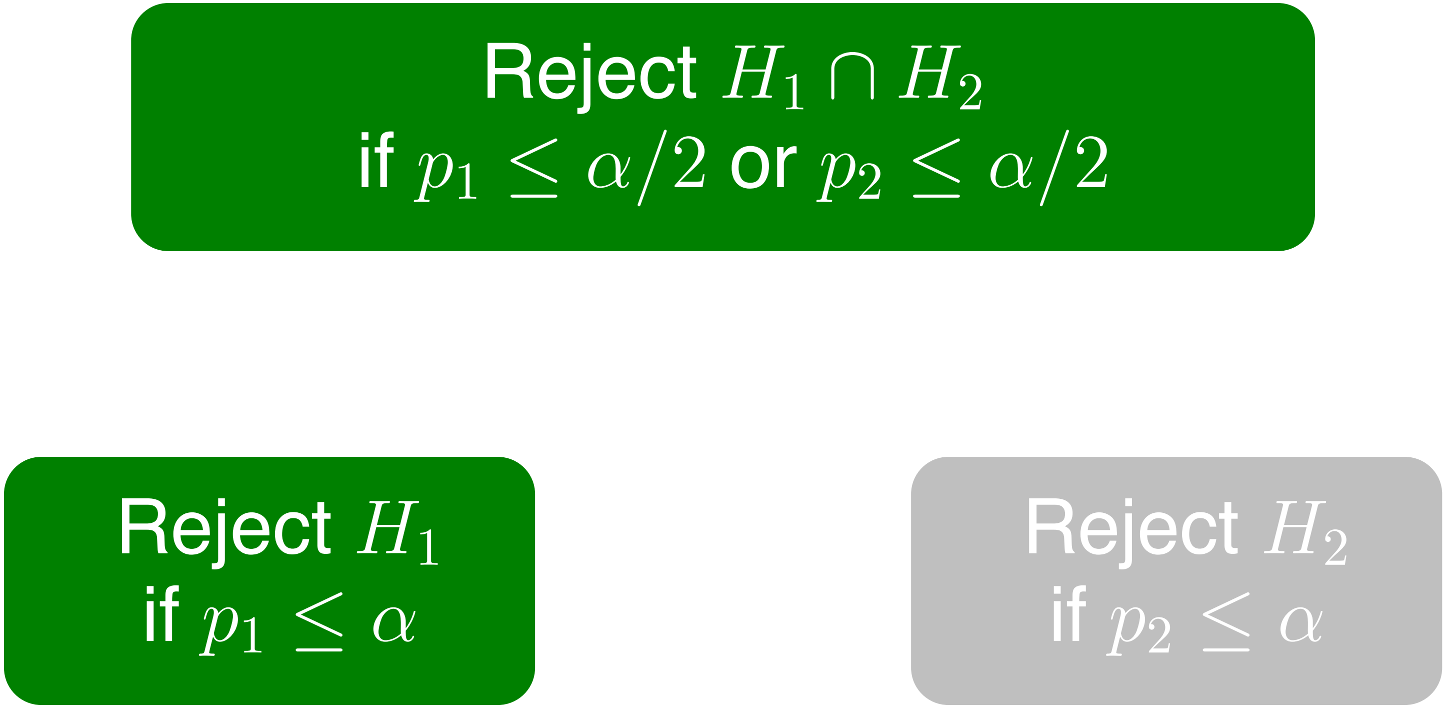
Reject H_2
if $p_2 \leq \alpha$

Bonferroni test is used for testing $H_1 \cap H_2$ and univariate tests for testing H_1 and H_2

Each local test is an α-level test

Example 1: Osteoporosis/breast cancer trial

Step 3: Define decision rules



Reject $H_1 \cap H_2$
if $p_1 \leq \alpha/2$ or $p_2 \leq \alpha/2$

Reject H_1
if $p_1 \leq \alpha$

Reject H_2
if $p_2 \leq \alpha$

Treatment effect for Endpoint 1 is significant if all intersection hypotheses containing H_1 are rejected

Example 1: Osteoporosis/breast cancer trial

Step 3: Define decision rules

Reject $H_1 \cap H_2$
if $p_1 \leq \alpha/2$ or $p_2 \leq \alpha/2$

Reject H_1
if $p_1 \leq \alpha$

Reject H_2
if $p_2 \leq \alpha$

Treatment effect for Endpoint 2 is significant if all intersection hypotheses containing H_2 are rejected

Familywise error rate control

Definition

FWER is controlled at α level if the probability of incorrectly rejecting at least one true null hypothesis is $\leq \alpha$ for all possible combinations of true and false null hypotheses

Closure principle

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

Closed testing procedure

Holm procedure

Bonferroni-based closed testing procedure

Decision rules

Reject H_1 if (1) $p_1 \leq \alpha/2$ or $p_2 \leq \alpha/2$ and (2) $p_1 \leq \alpha$

Reject H_2 if (1) $p_1 \leq \alpha/2$ or $p_2 \leq \alpha/2$ and (2) $p_2 \leq \alpha$

Closed testing procedure

Ordered p-values

$p_{(1)} < p_{(2)}$, ordered p-values

Example: If $p_1 > p_2$, then $p_{(1)} = p_2$ and $p_{(2)} = p_1$

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

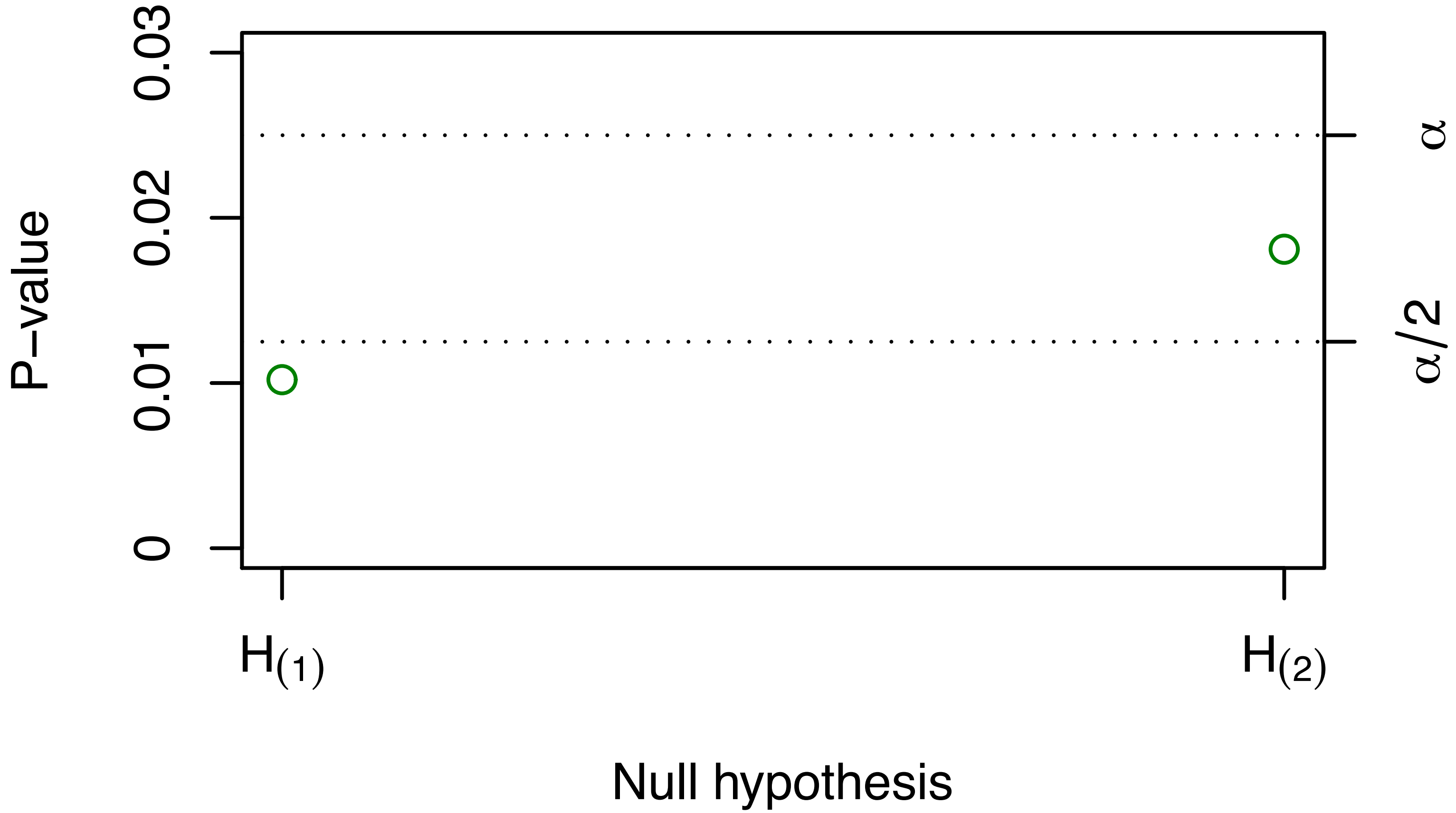
Alternative decision rules (more convenient)

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

Closure principle

Familywise error rate control

Closed testing procedures control FWER in the strong sense

Power

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

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

5. Selection of multiple testing procedures

Selection of multiple testing procedures

1. Define hypothesis testing problem

2. Define relationships among null hypotheses

Logical relationships

Are null hypotheses ordered?

Pre-specified versus data-driven ordering

Distributional relationships

Is joint distribution of hypothesis test statistics known?

Nonparametric, semiparametric and parametric methods

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

Hypothesis testing problem

Notation

$H_1, \ldots, H_m$, null hypotheses

α, familywise error rate

Assumptions

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

Logical relationships

Pre-specified hypothesis ordering

$H_1, \ldots, H_m$ are ordered to reflect clinical importance or probability of success for associated objectives

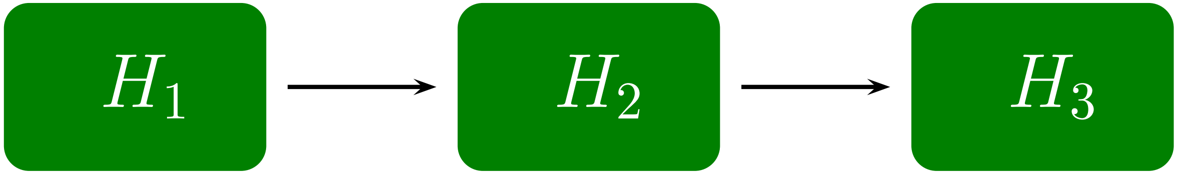
Example 4: Major depressive disorder trial

Strong evidence of a positive dose-response relationship (Dose H $>$ Dose M $>$ Dose L)

Logical relationships

Pre-specified hypothesis ordering

Null hypotheses H_1 (Dose H versus Placebo), H_2 (Dose M versus Placebo) and H_3 (Dose L versus Placebo) are tested sequentially



H_1 → H_2 → H_3

Pre-specified testing sequence in Example 4

Fixed-sequence, fallback and chain procedures

Logical relationships

Data-driven hypothesis ordering

$H_1, \ldots, H_m$ are not ordered

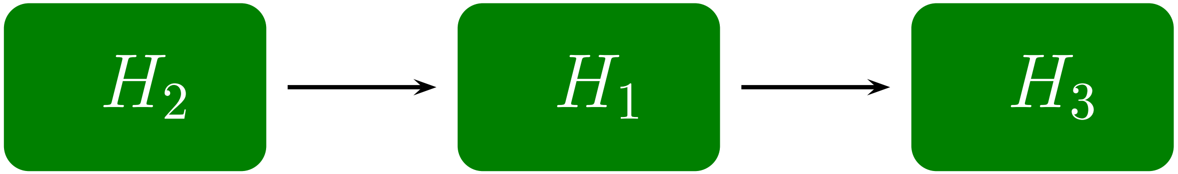
Example 4: Major depressive disorder trial

Difficult to assume a positive dose-response relationship

Logical relationships

Data-driven hypothesis ordering

Null hypotheses H_1 (Dose H versus Placebo), H_2 (Dose M versus Placebo) and H_3 (Dose L versus Placebo) are tested in the order determined by significance of hypothesis test statistics



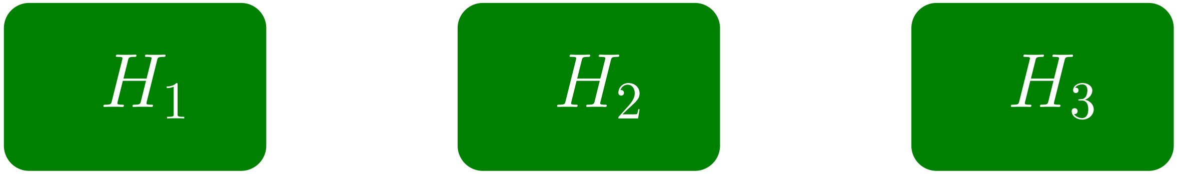
Data-driven testing sequence in Example 4

Holm, Hommel, Hochberg and step-down Dunnett procedures

Logical relationships

Single-step procedures

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



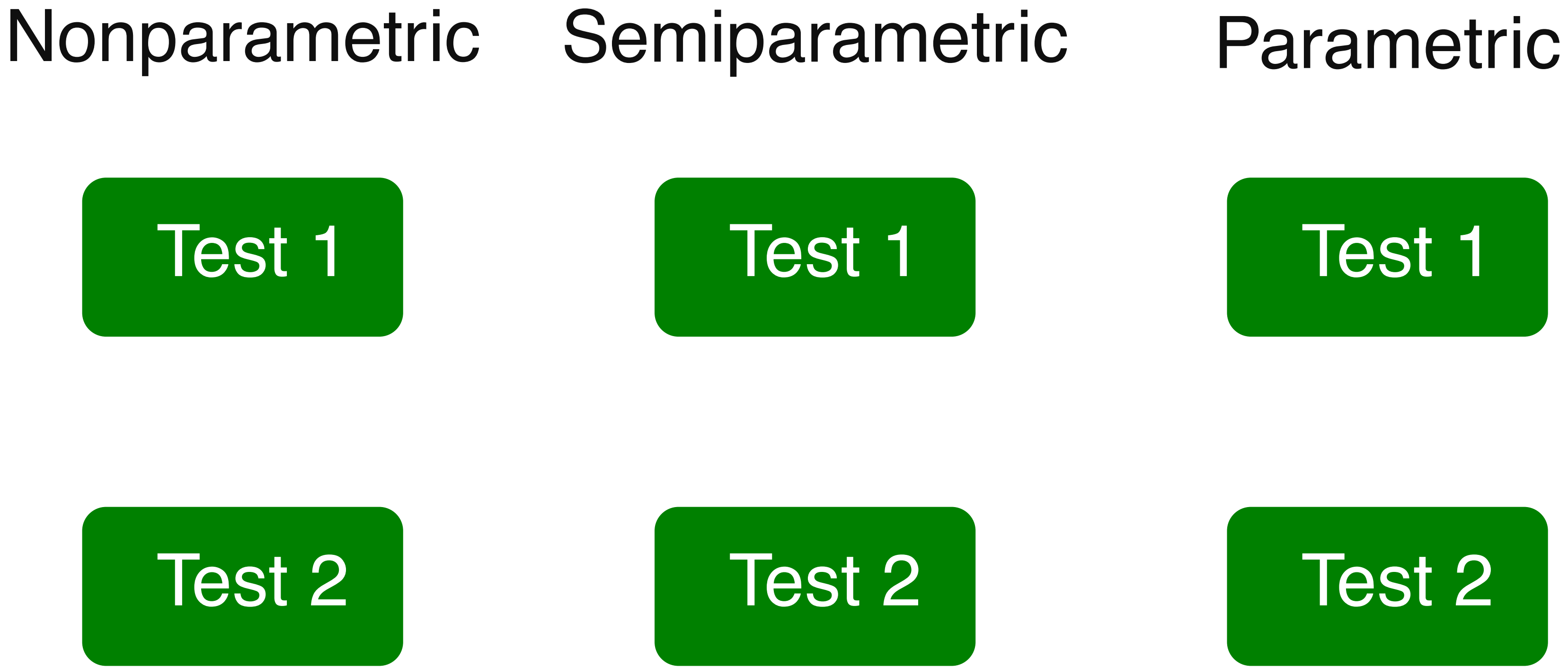
H_1 H_2 H_3

No testing sequence in Example 4

Bonferroni and Dunnett procedures

Distributional relationships

Three classes of procedures



Nonparametric

Test 1

Test 2

Semiparametric

Test 1

Test 2

Parametric

Test 1

Test 2

Distributional relationships

Nonparametric procedures

Based on univariate p-values

No distributional assumptions

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

Properties

Popular due to their simplicity but perform poorly with too many null hypotheses or strongly correlated test statistics

Distributional relationships

Semiparametric procedures

Based on univariate p-values

Some distributional assumptions

Examples: Hochberg and Hommel procedures

Properties

Perform better in problems with strongly correlated test statistics

More powerful than nonparametric procedures

Distributional relationships

Parametric procedures

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

Specific distributional assumptions

Examples: Dunnet procedures

Properties

Perform well in multiplicity problems with a known joint distribution of test statistics

More powerful than nonparametric procedures

Distributional relationships

Resampling-based procedures

Based on multivariate p-values computed from an approximate joint distribution of hypothesis test statistics

No distributional assumptions

Not widely used in clinical trial applications

Exercise

Example 5: Schizophrenia trial

Three tests (General population, Subpopulation 1 and Subpopulation 2)

Three test statistics are positively correlated and their joint distribution is known

Can a parametric procedure be used?

Example 1: Osteoporosis/breast cancer trial

Two endpoint tests (Vertebral fractures, Breast cancer) are positively correlated

Can a parametric procedure be used?

Distributional relationships

FWER control

Correlation information can be taken into account only if it is known at design stage

Correlations can be estimated but use of sample correlations in multiple testing procedures can potentially inflate FWER

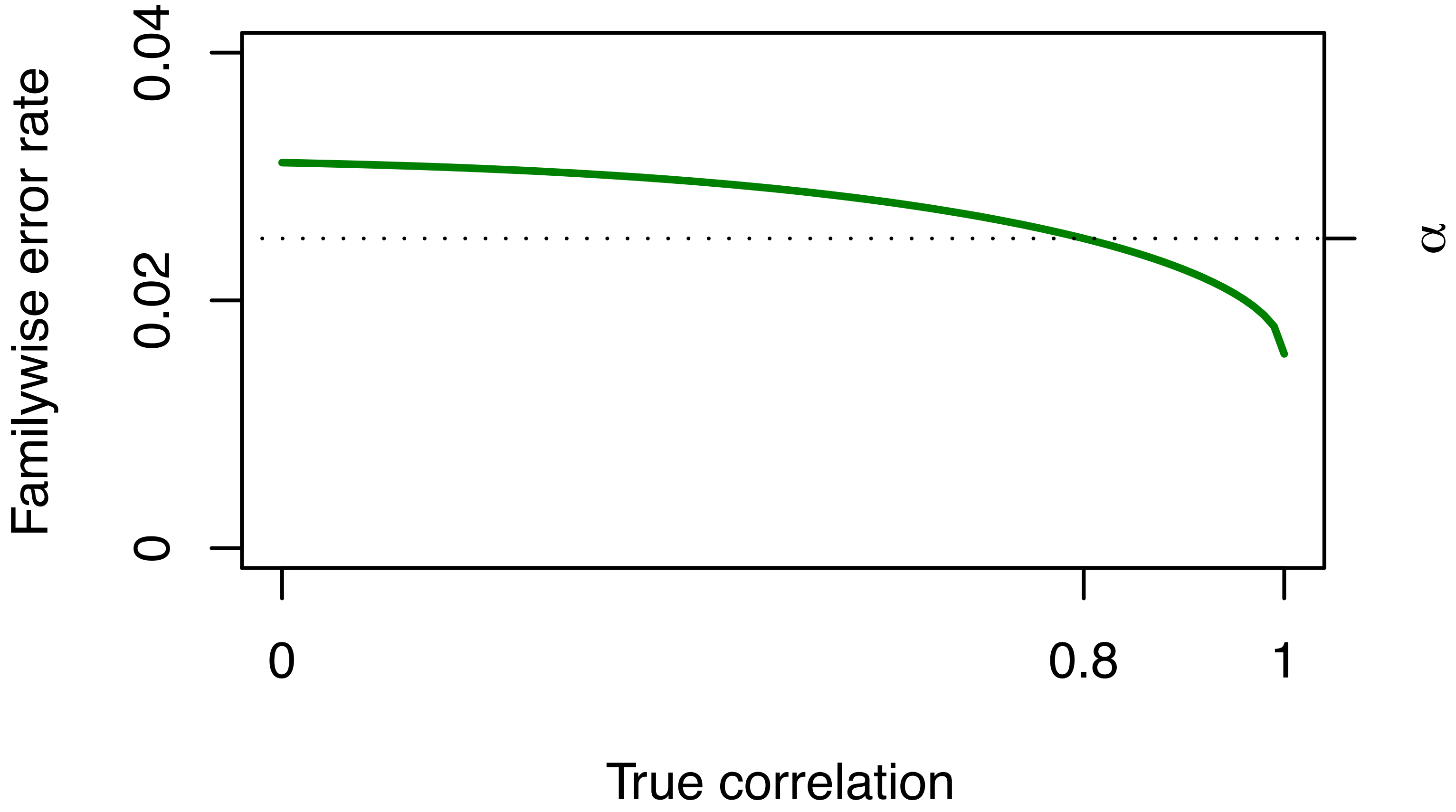
Clinical trial examples

Example 5: Parametric procedure

Example 1: Nonparametric procedure

Example 1: Osteoporosis/breast cancer trial

Parametric procedure ($\alpha = 0.025$)



Based on a bivariate normal distribution with sample correlation $\rho = 0.8$

Summary

Summary

Selection of multiple testing procedures

Critical to account for relevant information

Clinical information: Logical relationships among null hypotheses

Statistical information: Distributional relationships among null hypotheses

Aim for choosing a multiple testing procedure which is aligned with clinical objectives of a trial and maximizes power

Classification of multiple testing procedures

Distributional relationships	Logical relationships		
	Single-step	Data-driven hypothesis ordering	Pre-specified hypothesis ordering
Nonparametric	Bonferroni	Holm	Fixed-sequence Fallback Chain
Semiparametric	Simes	Hochberg Hommel	
Parametric	Dunnett	Step-down Dunnett Step-up Dunnett	Parametric fallback Parametric chain Feedback

Module A

Further reading

Multiple Testing Problems in Pharmaceutical Statistics (edited by Alex Dmitrienko, Ajit Tamhane and Frank Bretz)

Sections

2 (Inferential goals): Chapter 4

3 (Error rate definitions), 4 (Closure principle) and 5 (Selection of at-least-one procedures): Chapter 2

Module B
Nonparametric Procedures

Outline

1. Simple procedures with a pre-specified hypothesis ordering

Fixed-sequence procedure (Maurer et al., 1995) and fallback procedures (Wiens, 2003; Wiens and Dmitrienko, 2005)

2. Advanced procedures with a pre-specified hypothesis ordering

Chain procedures (Bretz et al., 2009; Millen and Dmitrienko, 2011)

Outline

3. Multiplicity-adjusted p-values

Adjusted significance levels and p-values

4. Simulation study

Comparison of fixed-sequence and fallback procedures

Setting

Distributional relationships

No distributional assumptions (any joint distribution of hypothesis test statistics)

Logical relationships

Multiple testing procedures with a pre-specified hypothesis ordering

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

1. Simple procedures with a pre-specified hypothesis ordering

Simple procedures

Fixed-sequence procedure

Multiple testing procedure based on sequentially rejective method

α allocation and propagation

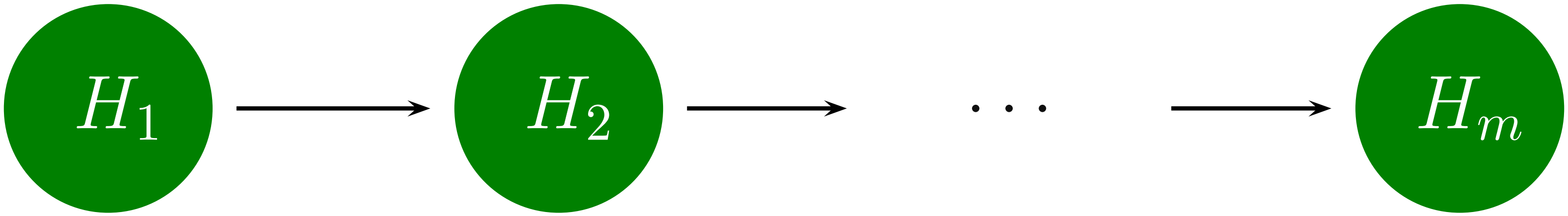
General rules for defining stepwise multiple testing procedures

Fallback procedures

Class of more flexible multiple testing procedures, which includes fixed-sequence procedure

Fixed-sequence procedure

Sequentially rejective method



— null hypothesis is rejected

Each null hypothesis is tested at α level

Single-strike rule: 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

Example 4: Major depressive disorder trial

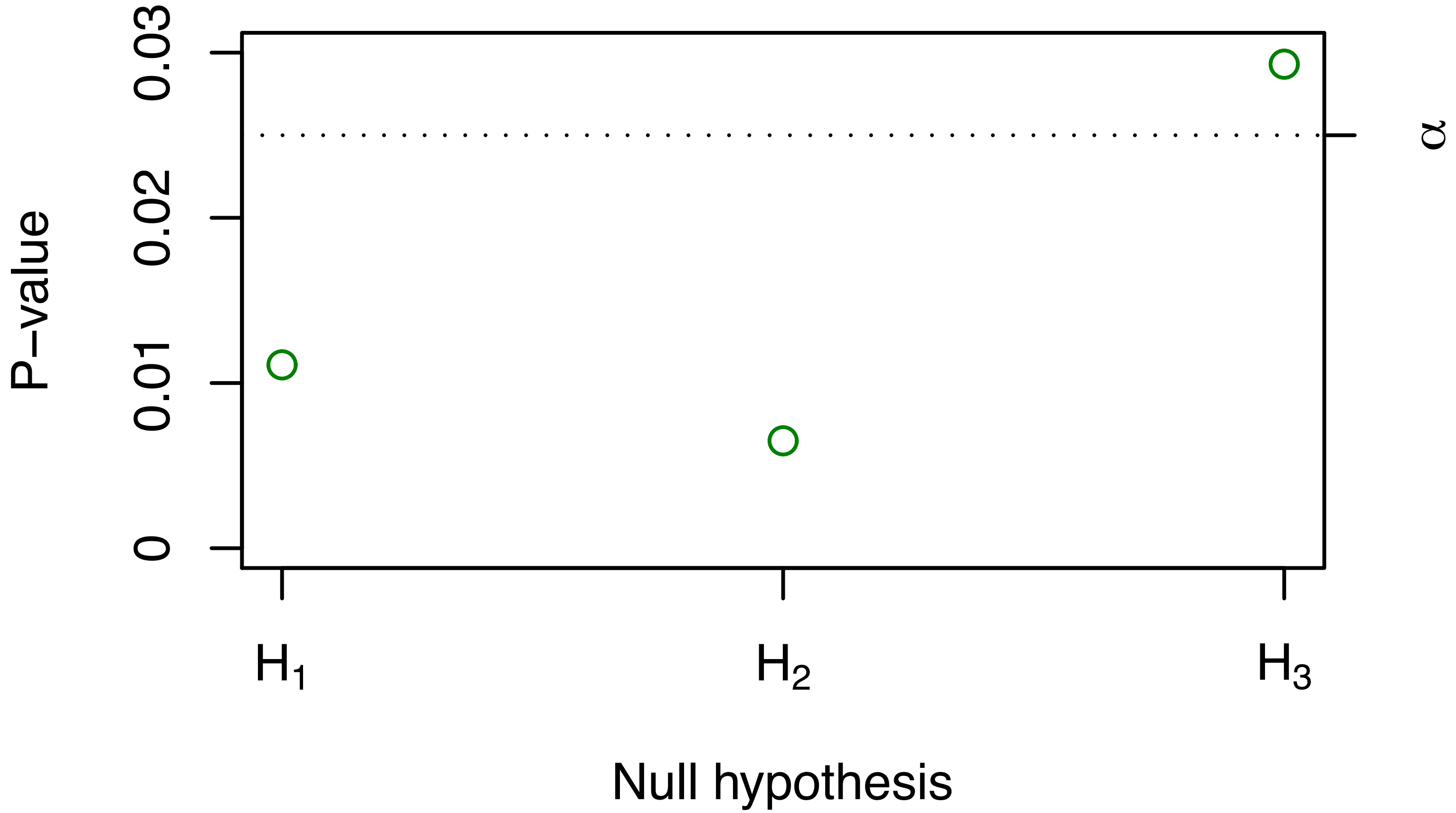
Scenario 1

Comparison	P-value
Dose H vs Placebo (H_1)	0.0111
Dose M vs Placebo (H_2)	0.0065
Dose L vs Placebo (H_3)	0.0293

Evidence of treatment effect at Doses 1 and 2

Fixed-sequence procedure

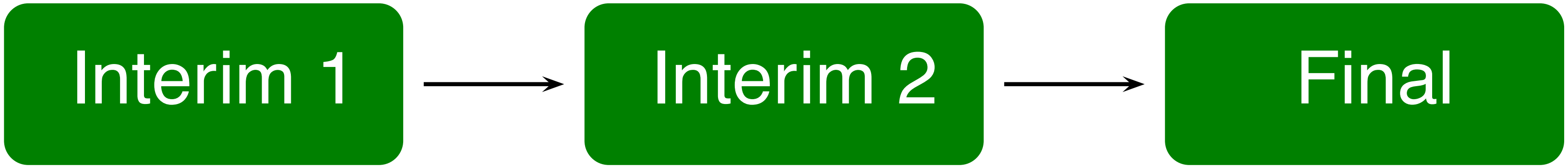
Decision rules in Example 4 ($\alpha = 0.025$)



Fixed-sequence procedure rejects H_1 (Dose H vs Placebo) and H_2 (Dose M vs Placebo)

Exercise

Clinical trial with two interim analyses and final analysis



Interim 1

Interim 2

Final

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

Fixed-sequence procedure

Type I error rate

Fixed-sequence procedure 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: Major depressive disorder trial

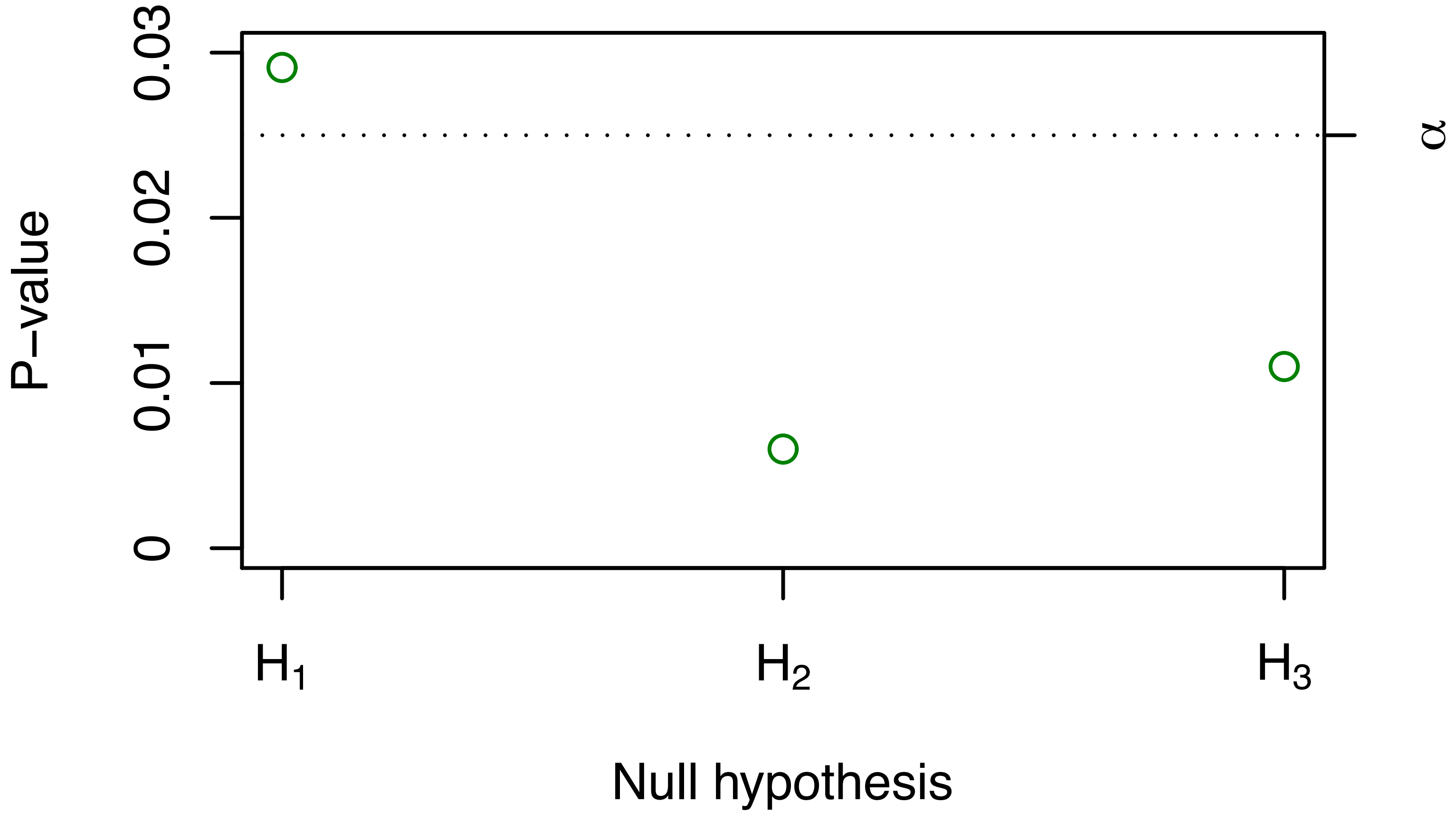
Scenario 2

Comparison	P-value
Dose H vs Placebo (H_1)	0.0291
Dose M vs Placebo (H_2)	0.0060
Dose L vs Placebo (H_3)	0.0110

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

Fixed-sequence procedure

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



Fixed-sequence procedure rejects no null hypotheses

Power evaluation

Clinical trial

Two arms: Experimental treatment versus placebo

Sample size: 98 patients per arm

Endpoints

Three continuous endpoints (E1, E2 and E3)

Endpoint test statistics are equicorrelated and follow a multivariate normal distribution

Common correlation is 0 or 0.5

Power evaluation

Pre-specified hypotheses ordering

H_1, H_2 and H_3, null hypotheses corresponding to Endpoints E1, E2 and E3

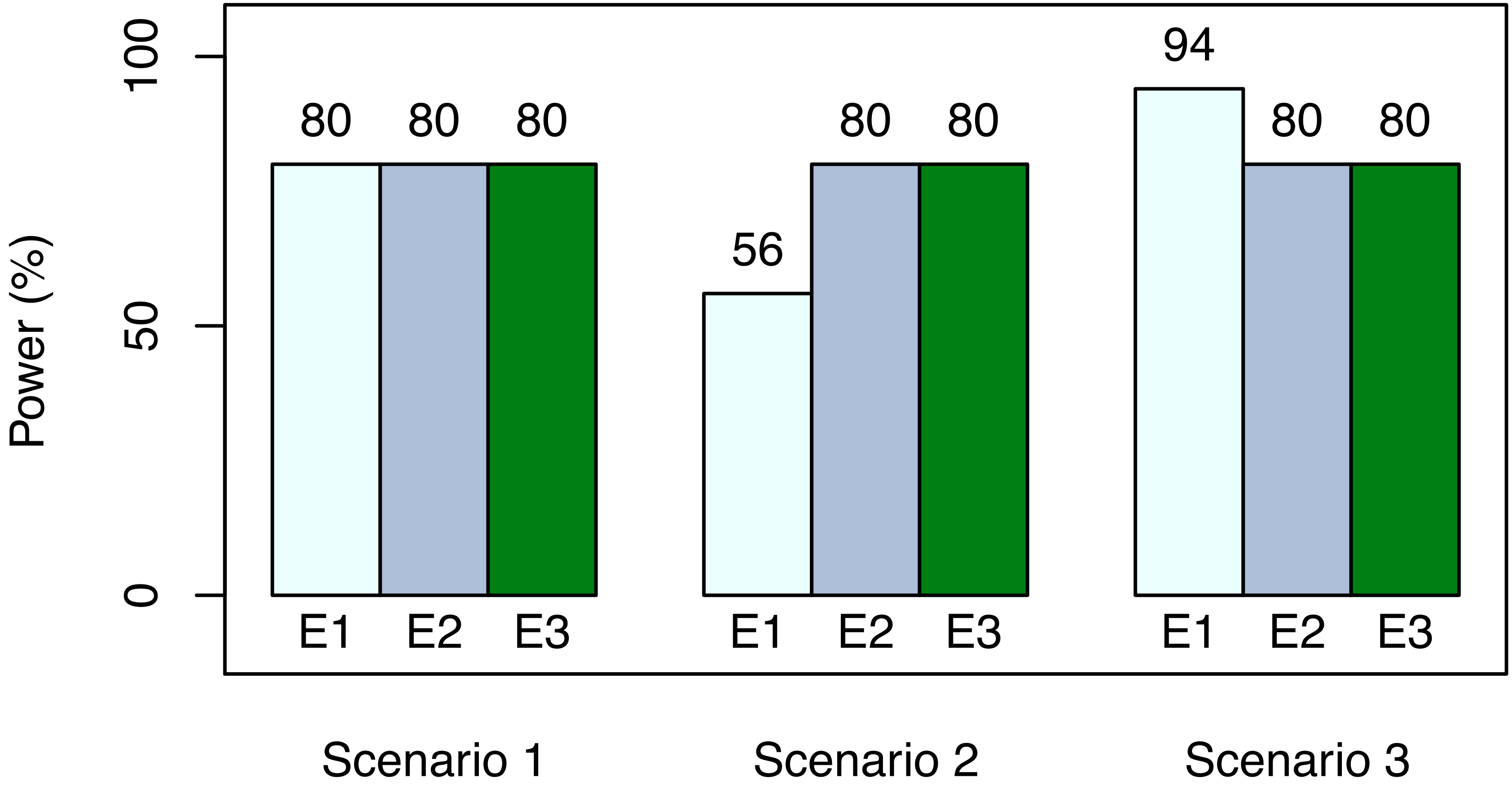
H_1, H_2 and H_3 are tested sequentially

Multiple testing procedure

Fixed-sequence procedure

Evaluation scenarios

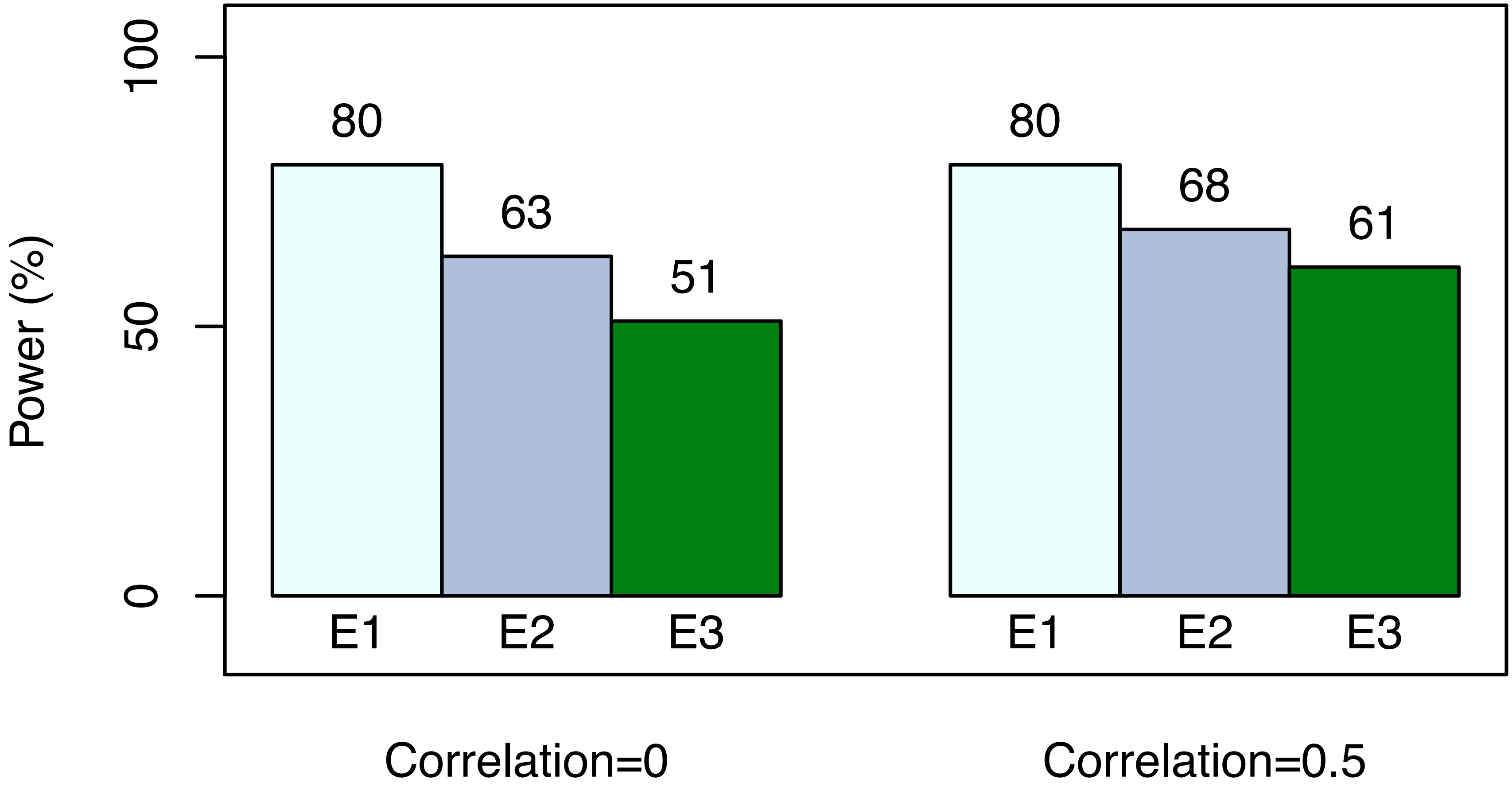
Marginal power for endpoint tests



Scenario 1: E1 is adequately powered, Scenario 2: E1 is underpowered, Scenario 3: E1 is overpowered

Scenario 1

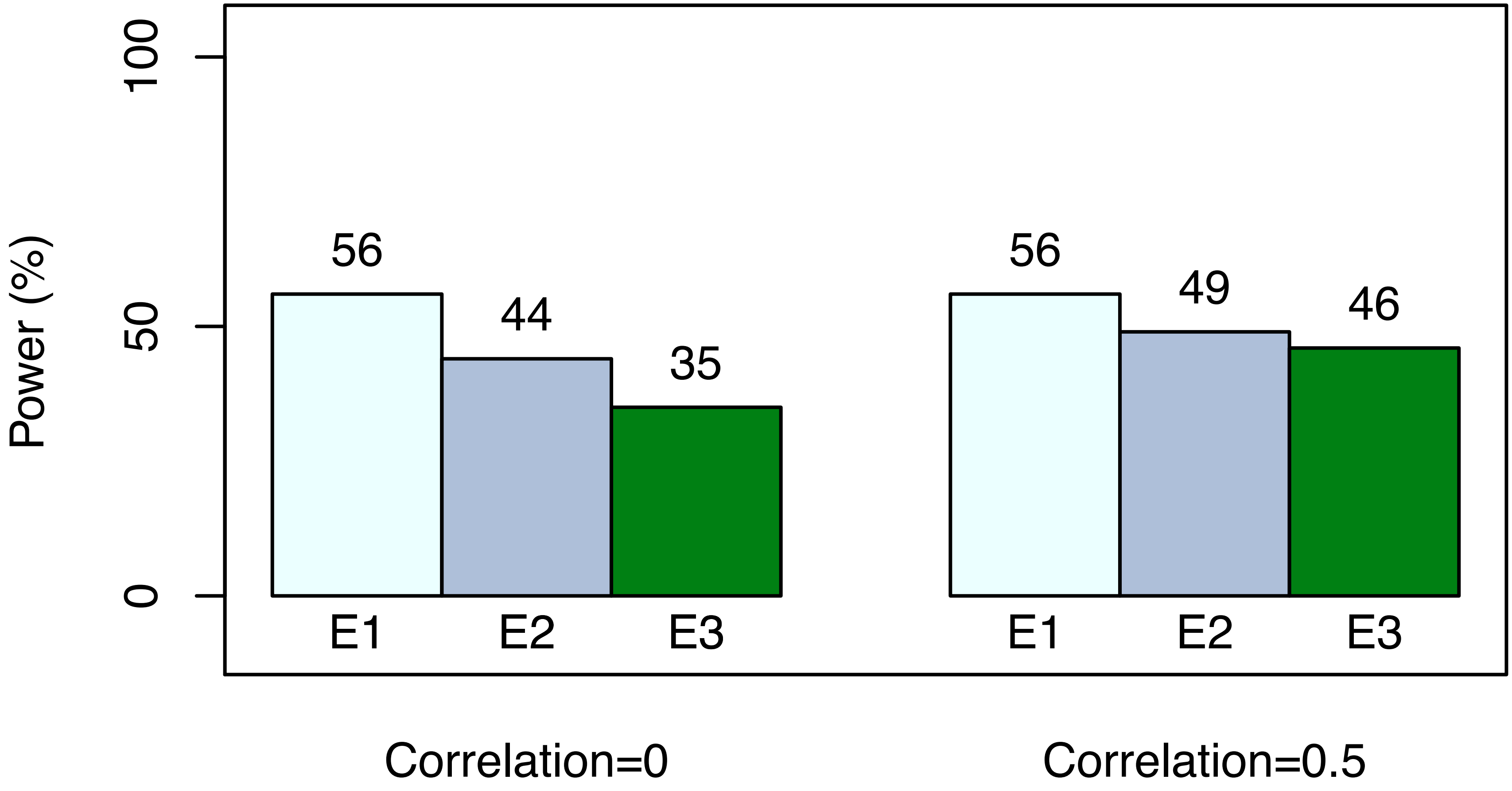
E1 is adequately powered



Considerable loss of power toward the end of the testing sequence (domino effect)

Scenario 2

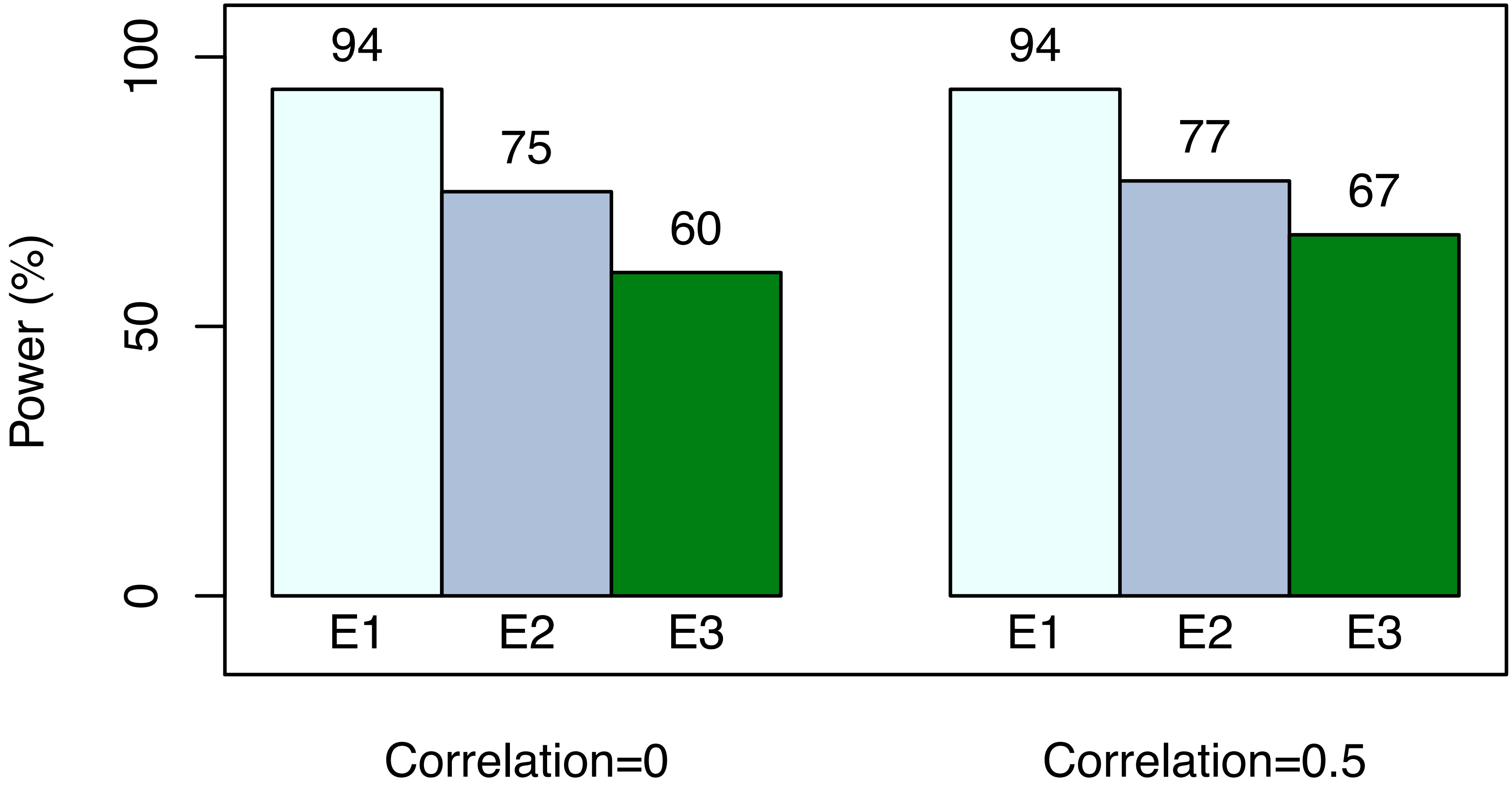
E1 is underpowered



Considerable loss of power toward the end of the testing sequence (domino effect)

Scenario 3

E1 is overpowered



Considerable loss of power toward the end of the testing sequence (domino effect)

Stepwise multiple testing procedures

Single-step procedures

Each hypothesis is tested at a fixed significance level

General rules for defining stepwise 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

Stepwise multiple testing procedures

α allocation rule

Initial distribution of the error rate is specified using hypothesis weights

Hypothesis weights

w_i, Weight assigned to H_i

Properties

$$w_i \geq 0,\ i = 1, \ldots, m$$

$$w_1 + \ldots + w_m = 1$$

Stepwise multiple testing procedures

α propagation rule

Error rate transfer is specified using transition parameters

Transition parameters

g_{ij}, Proportion of the error rate transferred from H_i to H_j after H_i is rejected

Properties

$g_{ij} \geq 0$, $i, j = 1, \ldots, m$

$g_{ii} = 0$, $i = 1, \ldots, m$

$g_{i1} + \ldots + g_{im} \leq 1$, $i = 1, \ldots, m$

α propagation

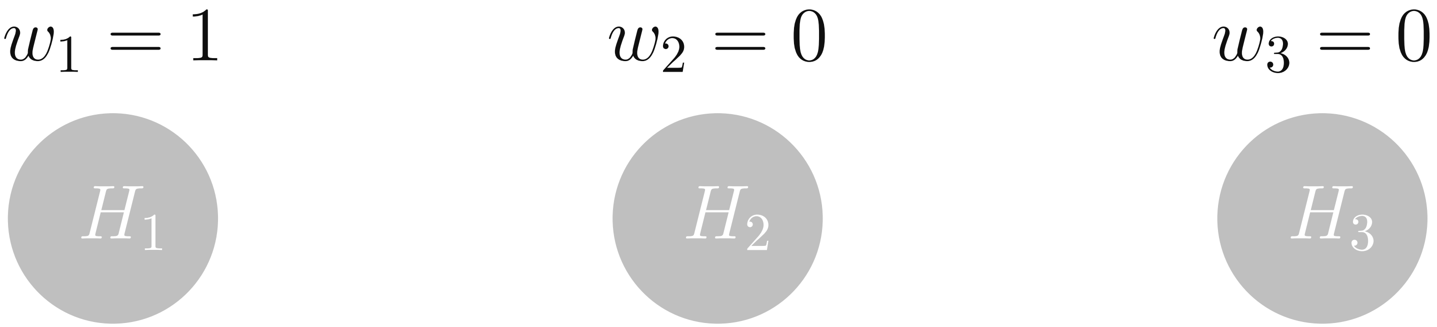
“Use it or lose it” principle

After a hypothesis is rejected, the significance level used in the test can be applied to remaining non-rejected hypotheses (Dmitrienko, Tamhane and Wiens, 2008)

“Use it or lose it” principle follows from the closure principle and does not result in FWER inflation

Fixed-sequence procedure

α allocation rule in Example 4

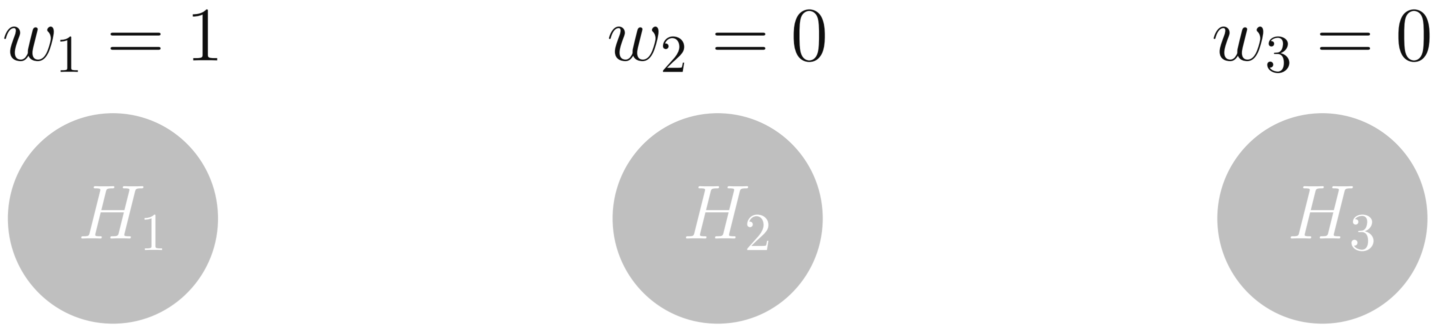


α allocation rule defines the initial hypothesis weights w_1, w_2 and w_3

All of initial weight is allocated to the first hypothesis in the testing sequence (H_1)

Fixed-sequence procedure

α allocation rule in Example 4



$w_1 = 1$ $w_2 = 0$ $w_3 = 0$

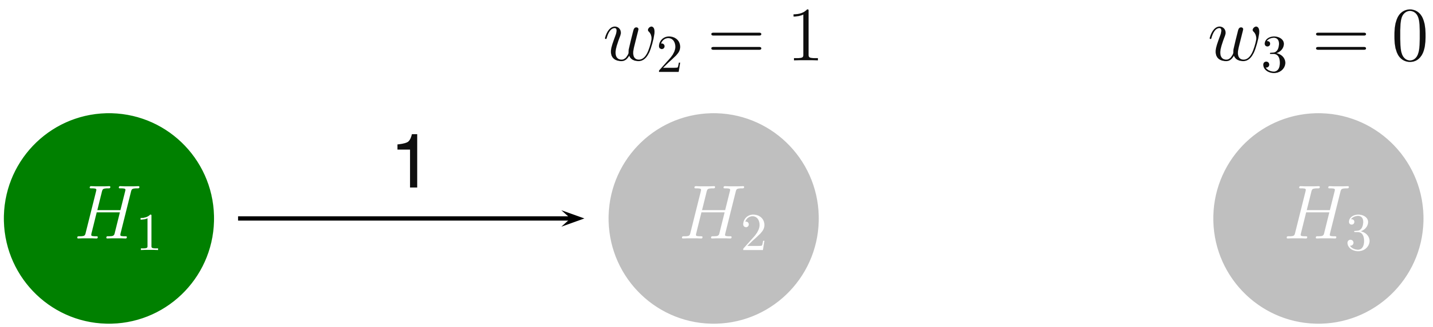
H_1 H_2 H_3

First hypothesis in the sequence (H_1) is tested at $\alpha w_1 = \alpha$

H_1 is rejected if $p_1 \leq \alpha$

Fixed-sequence procedure

α propagation rule in Example 4



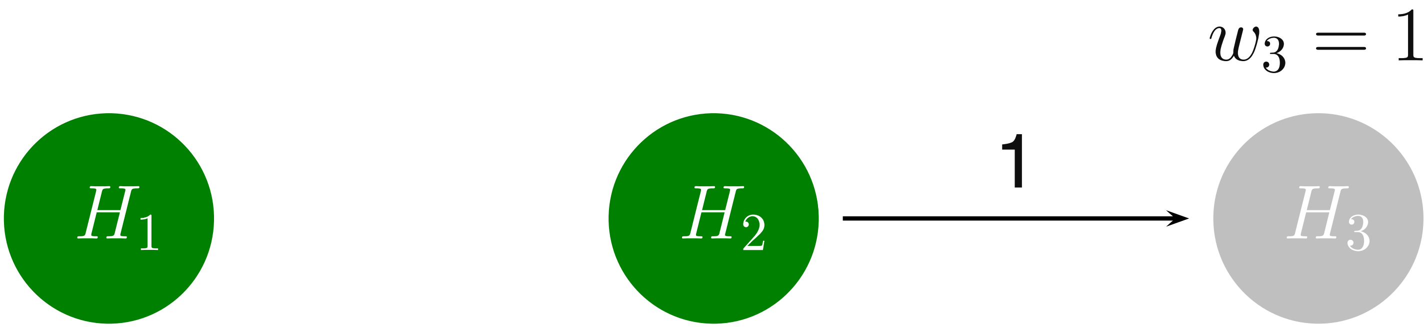
Hypotheses are tested sequentially and α propagation rule defines the process of updating hypothesis weights after each hypothesis is rejected

Entire weight allocated to H_1 is transferred to H_2 after H_1 is rejected

H_2 is rejected if $p_1 \leq \alpha$ and $p_2 \leq \alpha w_2 = \alpha$

Fixed-sequence procedure

α propagation rule in Example 4

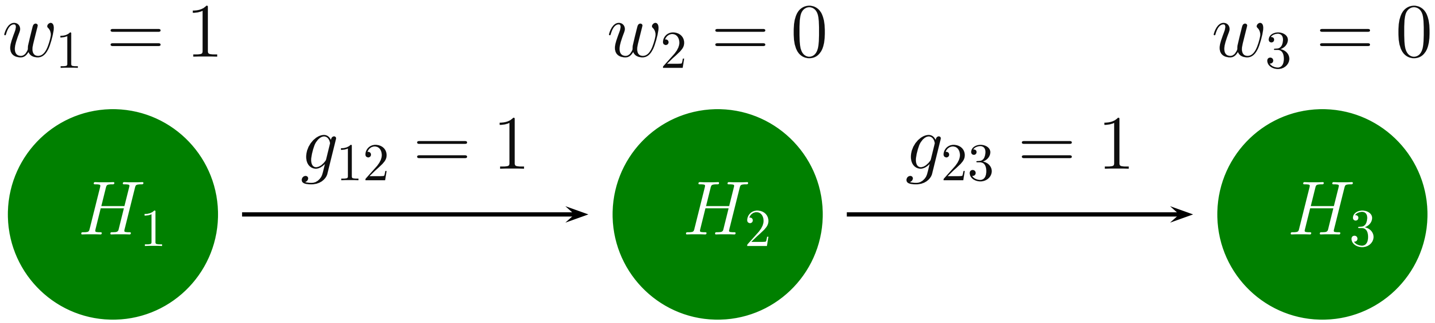


Entire weight allocated to H_2 is transferred to H_3 after H_2 is rejected

H_3 is rejected if $p_1 \leq \alpha$ and $p_2 \leq \alpha$ and $p_3 \leq \alpha w_3 = \alpha$

Fixed-sequence procedure in Example 4

Graphical representation



Fixed-sequence procedure in Example 4

α allocation rule

Hypothesis weights

$$W = (1, 0, 0)$$

α propagation rule

Transition parameters

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

Fallback procedures

Class of flexible stepwise procedures

Fallback procedures serve as an extension of the fixed-sequence procedure

Fixed-sequence procedure is a special case of fallback procedures

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

Fallback procedures

α allocation rule

$w_1, \ldots, w_m$, Non-negative hypothesis weights

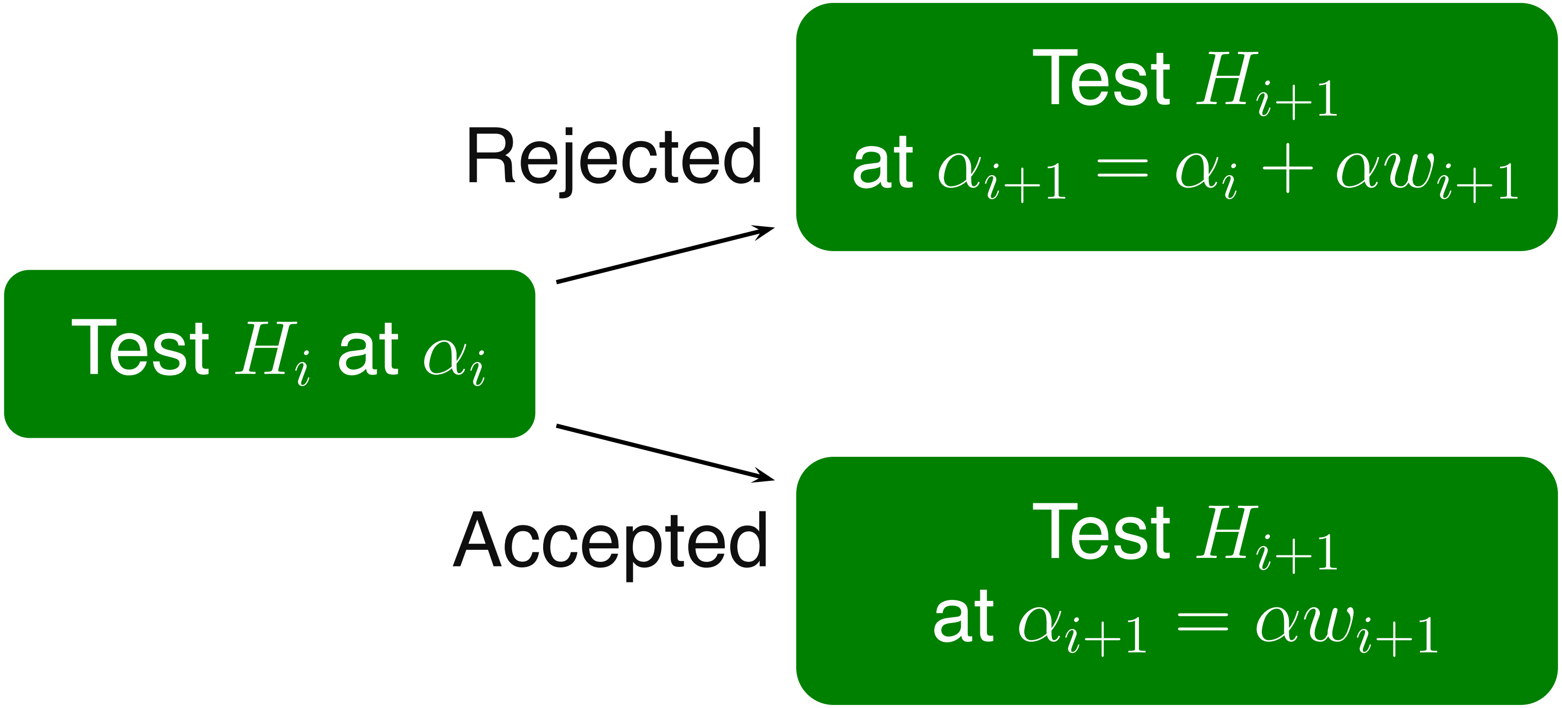
Fixed-sequence procedure is a special case with $w_1 = 1$, $w_2 = \ldots = w_m = 0$

α propagation rule

Error rate is transferred to the next hypothesis in the testing sequence

Fallback procedures

Testing method with a fallback option



Test H_i at α_i

Rejected

Test H_{i+1} at $\alpha_{i+1} = \alpha_i + \alpha w_{i+1}$

Accepted

Test H_{i+1} at $\alpha_{i+1} = \alpha w_{i+1}$

No single-strike rule: Next hypothesis (H_{i+1}) can be tested even if current hypothesis (H_i) is not rejected

Example 4: Major depressive disorder trial

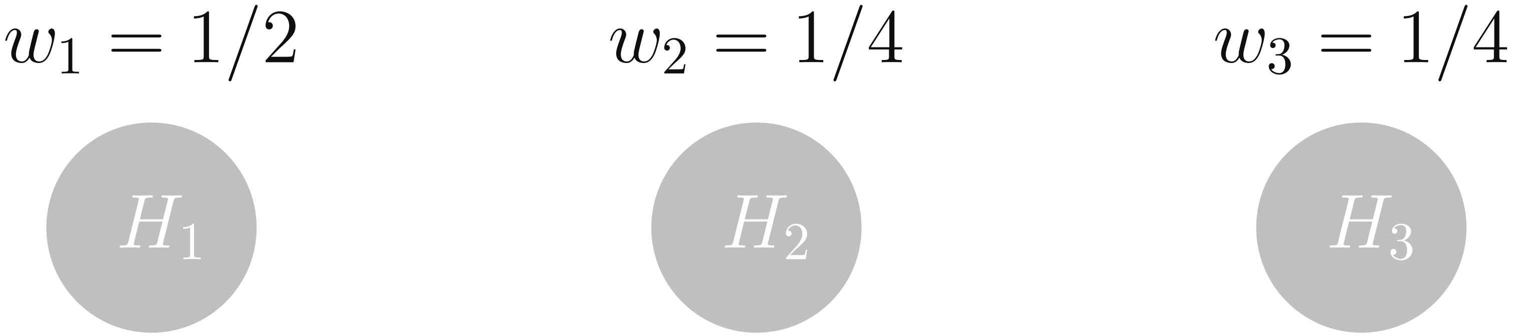
Scenario 2

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

Greater weight is assigned to Dose H since it is expected to be more efficacious than Dose M or Dose L

Fallback procedure

α allocation rule in Example 4



$w_1 = 1/2$ $w_2 = 1/4$ $w_3 = 1/4$

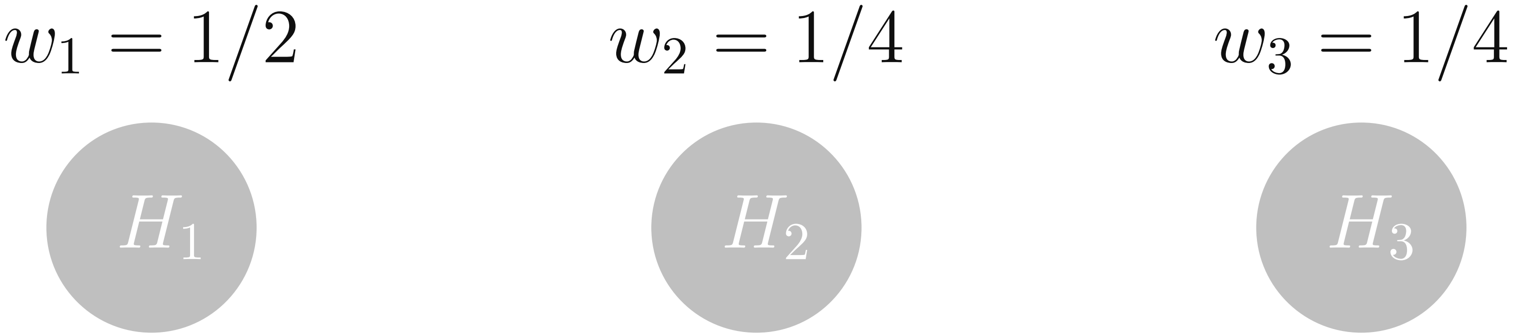
H_1 H_2 H_3

Positive weights are allocated to all hypotheses

H_2 and H_3 can be tested even if H_1 is not rejected

Fallback procedure

α allocation rule in Example 4



$w_1 = 1/2$

H_1

$w_2 = 1/4$

H_2

$w_3 = 1/4$

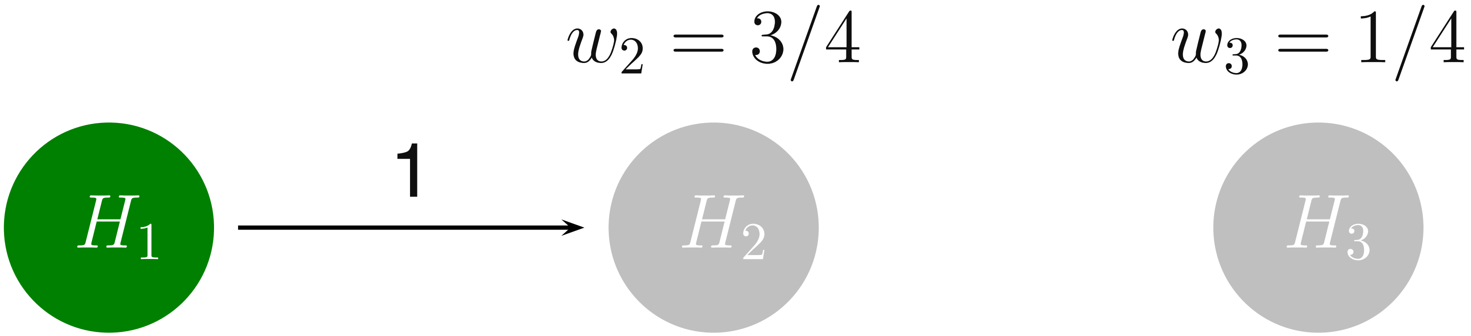
H_3

First hypothesis in the sequence (H_1) is tested at $\alpha w_1 = \alpha/2$

H_1 is rejected if $p_1 \leq \alpha/2$

Fallback procedure

α propagation rule in Example 4

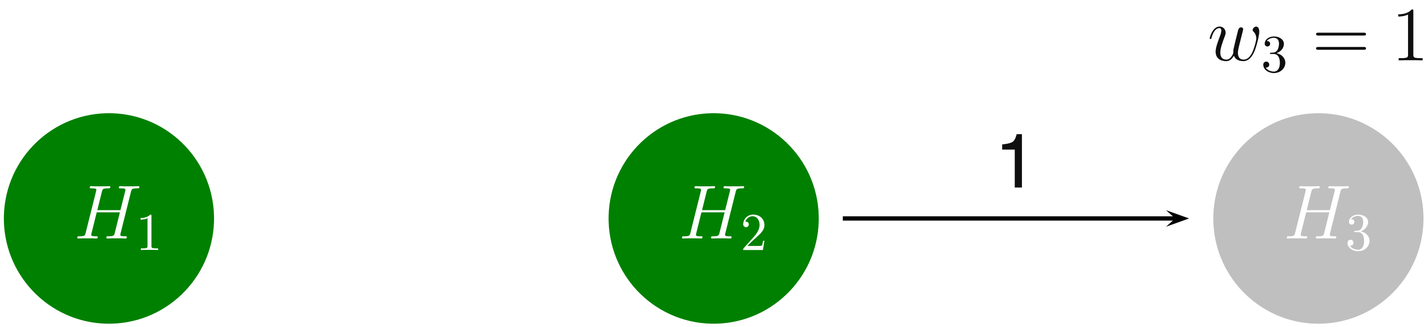


Entire weight allocated to H_1 is transferred to H_2 if H_1 is rejected

H_2 is rejected if $p_1 \leq \alpha/2$ and $p_2 \leq \alpha w_2 = 3\alpha/4$

Fallback procedure

α propagation rule in Example 4

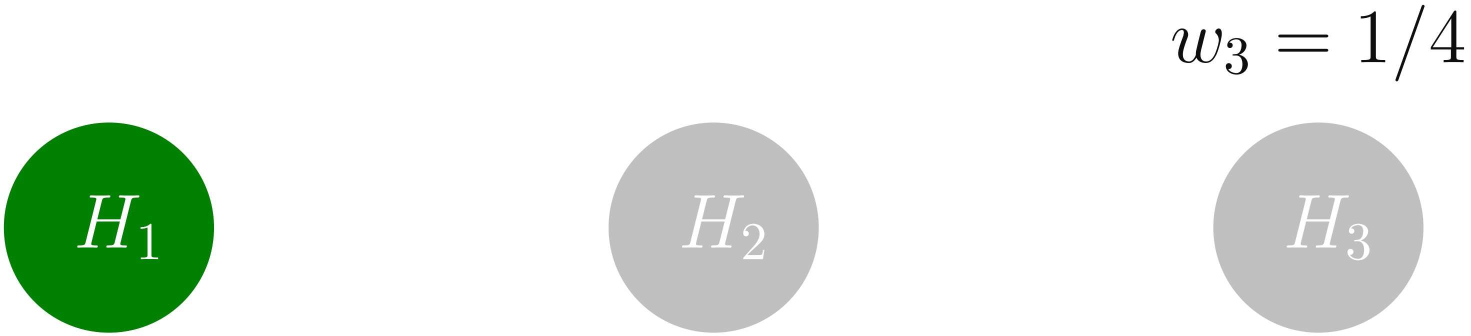


Entire weight allocated to H_2 is transferred to H_3 if H_2 is rejected

H_3 is rejected if $p_1 \leq \alpha/2$, $p_2 \leq 3\alpha/4$ and $p_3 \leq \alpha w_3 = \alpha$

Fallback procedure

α propagation rule in Example 4

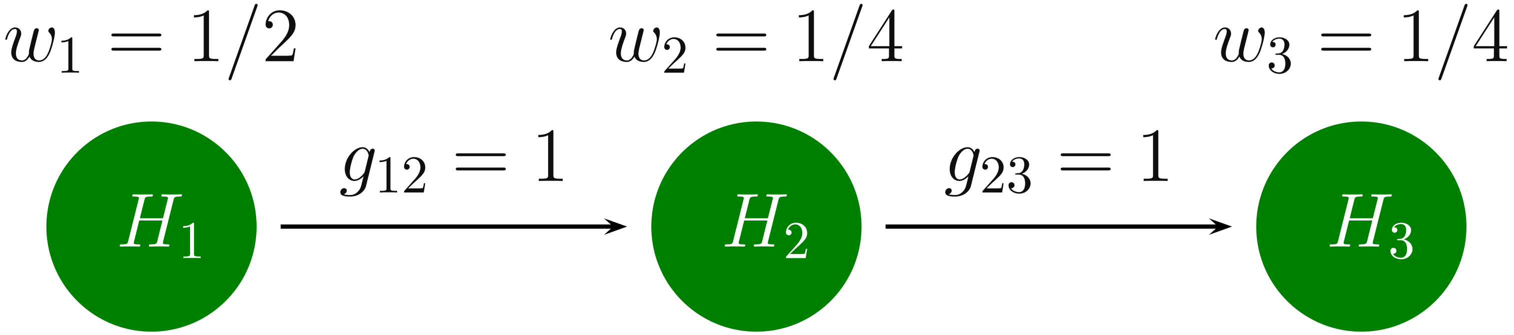


No weight is transferred from H_2 to H_3 if H_2 is accepted

H_3 is rejected if $p_2 > 3\alpha/4$ and $p_3 \leq \alpha w_3 = \alpha/4$

Fallback procedure in Example 4

Graphical representation



Fallback procedure in Example 4

α allocation rule

Hypothesis weights

$$W = (1/2, 1/4, 1/4)$$

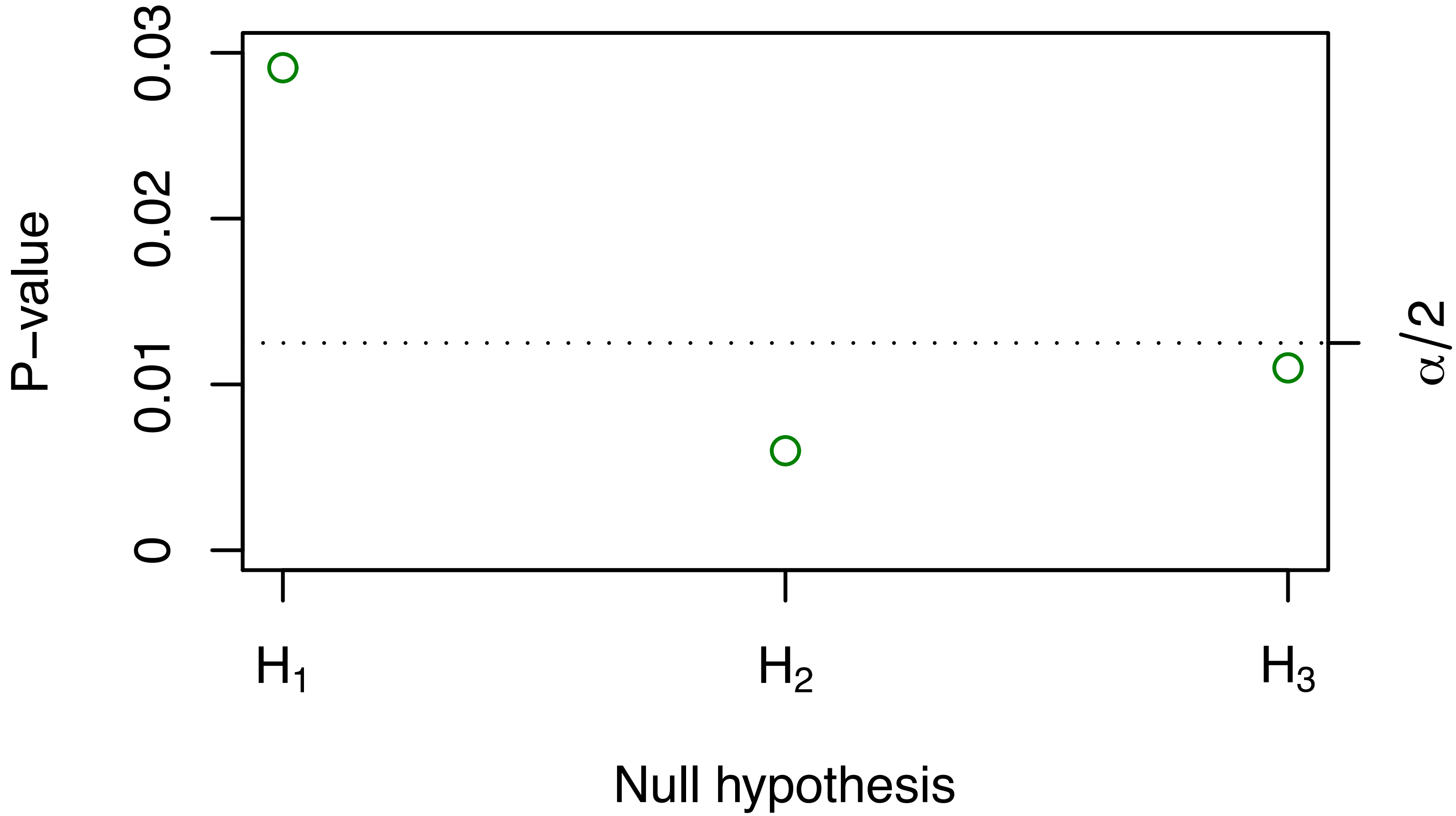
α propagation rule

Transition parameters

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

Fallback procedure

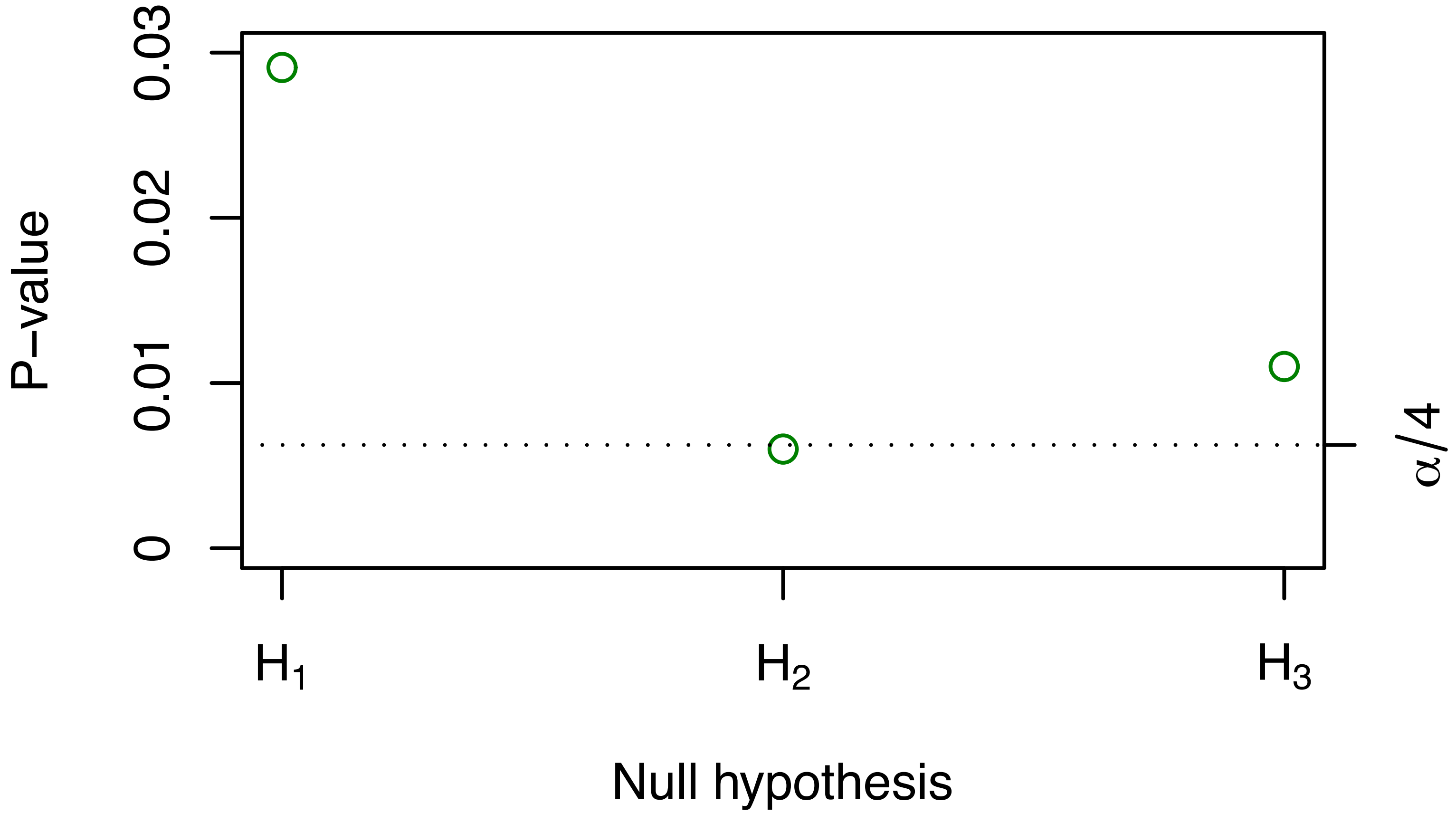
Decision rules in Example 4 ($\alpha = 0.025$)



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

Fallback procedure

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

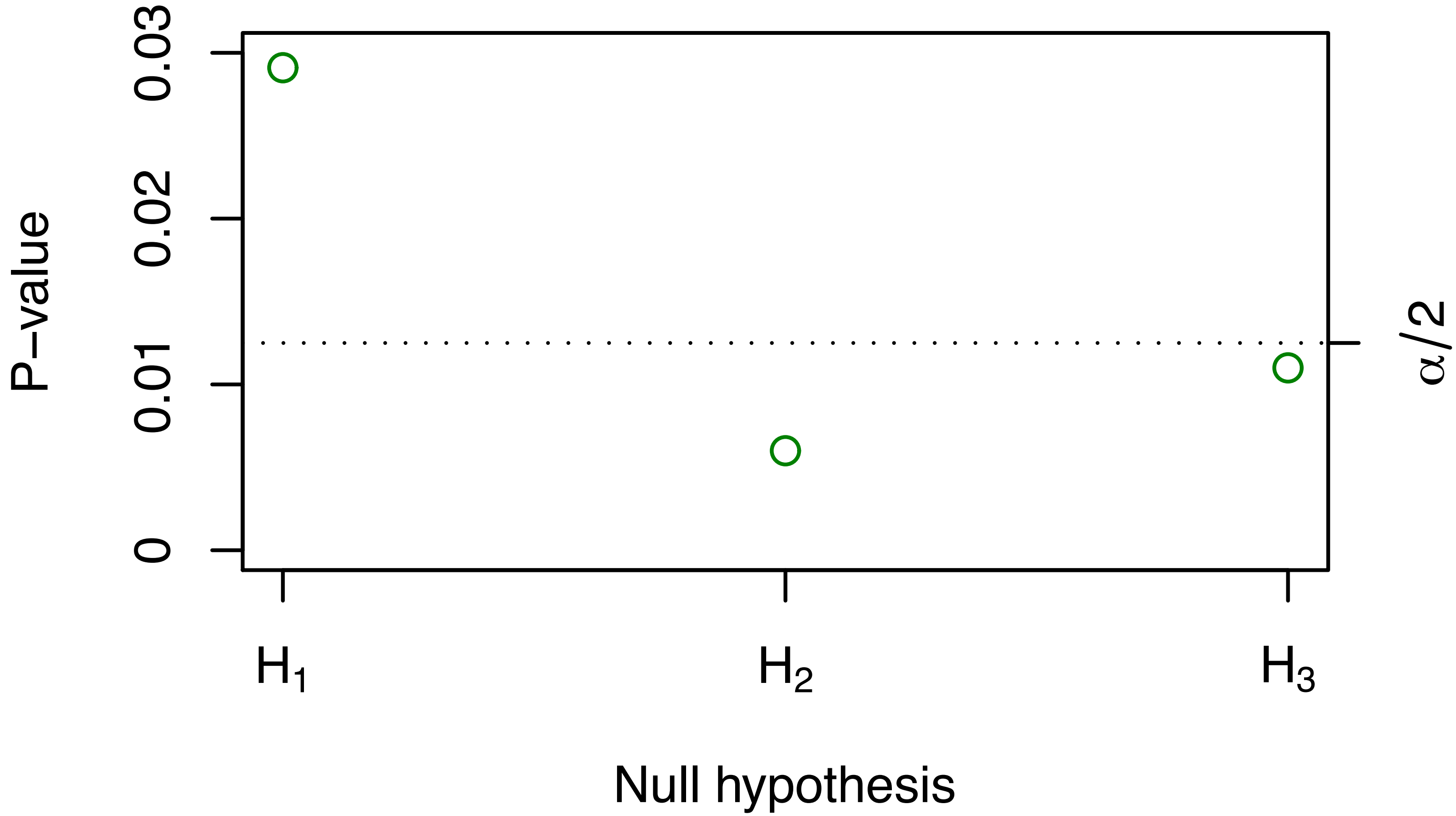


Step 2: H_2 is rejected since $w_2 = 1/4$ and $p_2 < \alpha w_2 = \alpha/4$

Error rate is transferred to H_3 and $w_3 = w_3 + w_2 = 1/2$

Fallback procedure

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



Step 3: H_3 is rejected since $w_3 = 1/2$ and $p_3 < \alpha w_3 = \alpha/2$

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

Fallback procedure

Type I error rate

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

Power

Fallback procedure is uniformly more powerful than weighted Bonferroni procedure with the same set of weights

Power evaluation

Clinical trial

Two arms: Experimental treatment versus placebo

Sample size: 98 patients per arm

Endpoints

Three continuous endpoints (E1, E2 and E3)

Endpoint test statistics are equicorrelated and follow a multivariate normal distribution

Common correlation is 0.5

Power evaluation

Pre-specified hypotheses ordering

H_1, H_2 and H_3, null hypotheses corresponding to Endpoints E1, E2 and E3

H_1, H_2 and H_3 are tested sequentially

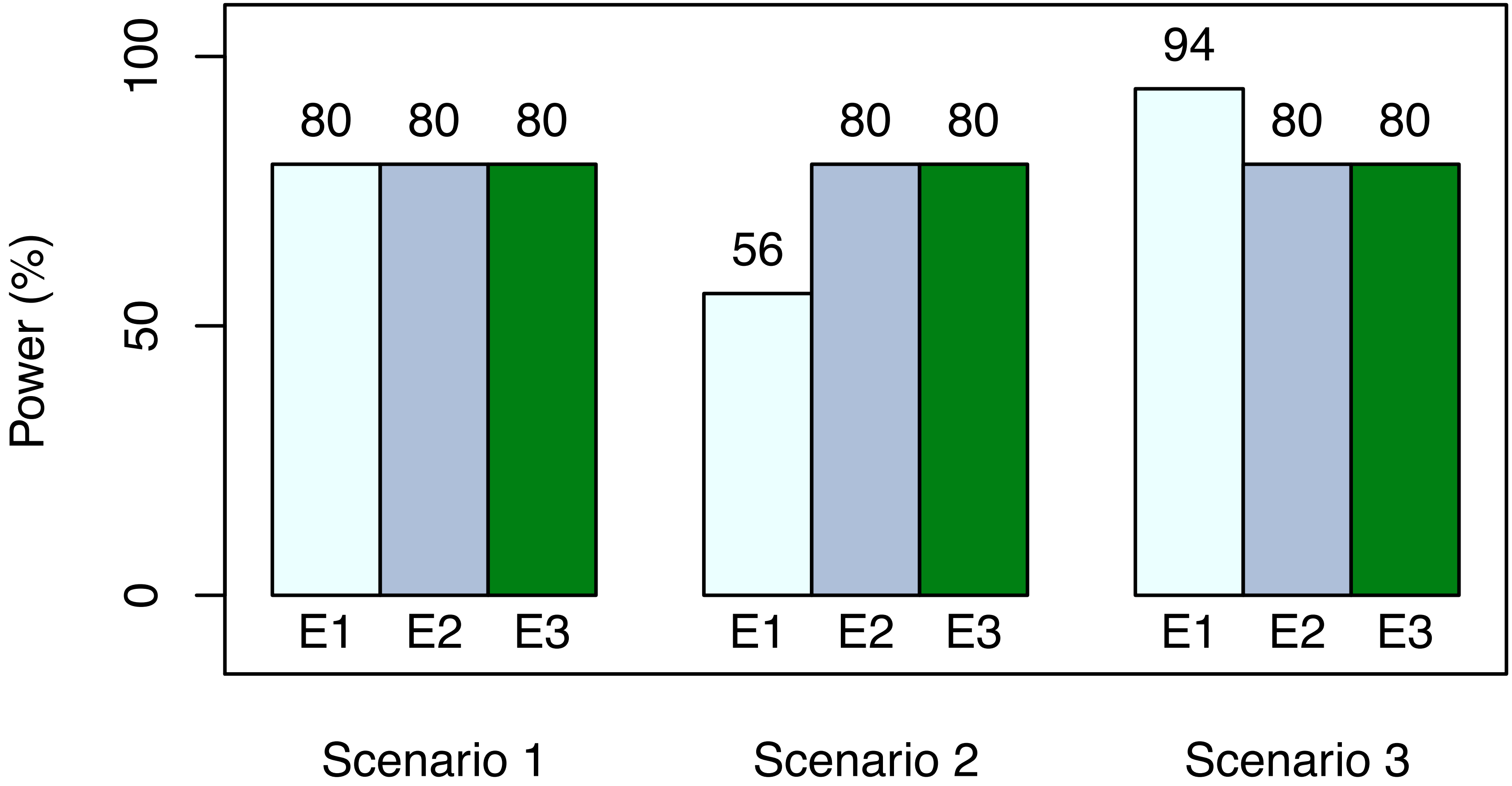
Multiple testing procedures

Fixed-sequence procedure

Fallback procedure with $w_1 = 1/2$, $w_2 = 1/4$, $w_3 = 1/4$

Evaluation scenarios

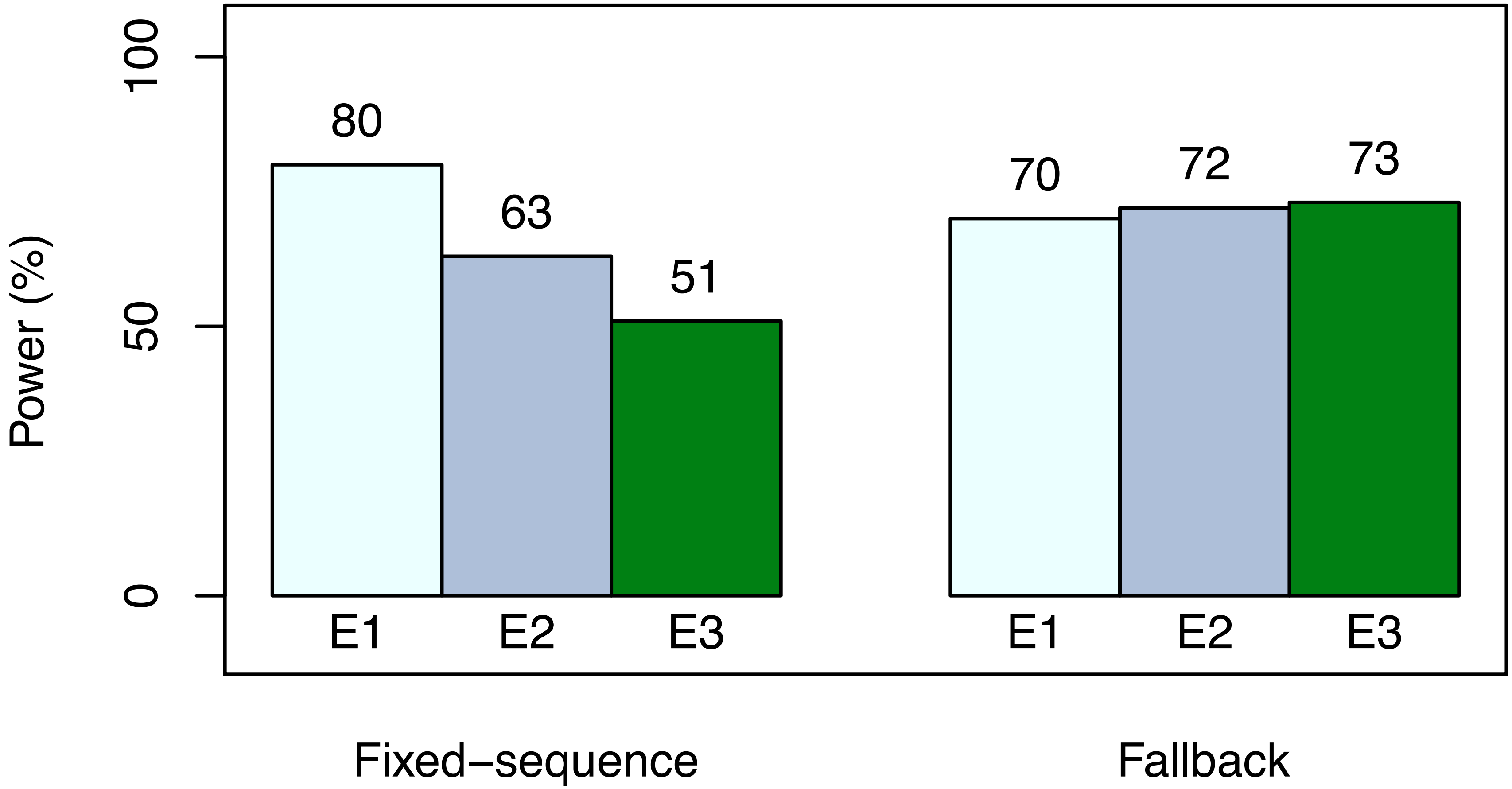
Marginal power for endpoint tests



Scenario 1: E1 is adequately powered, Scenario 2: E1 is underpowered, Scenario 3: E1 is overpowered

Scenario 1

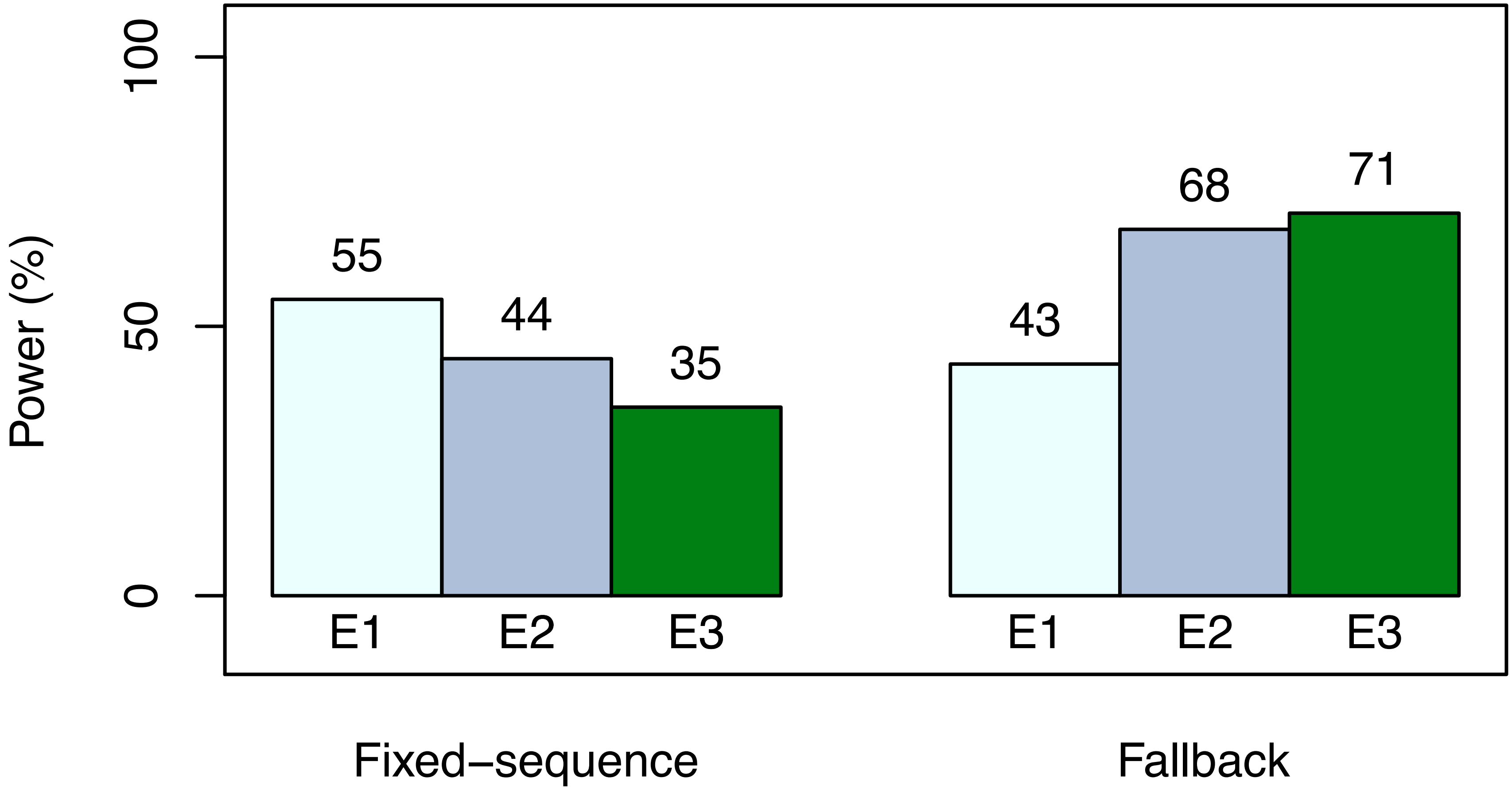
E1 is adequately powered



Fallback procedure does not lead to the domino effect (power does not decrease toward the end of the sequence)

Scenario 2

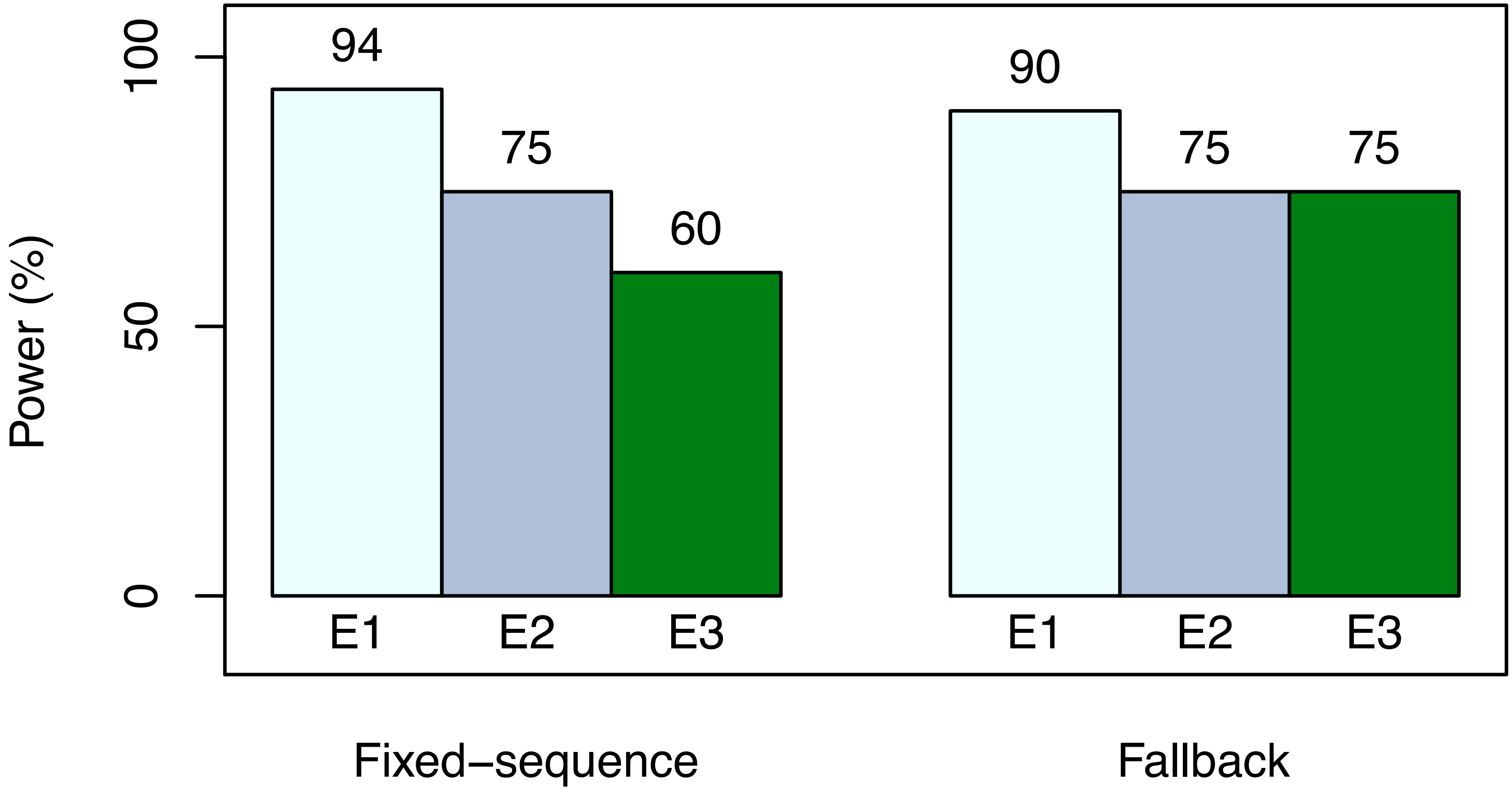
E1 is underpowered



Fallback procedure does not lead to the domino effect (power does not decrease toward the end of the sequence)

Scenario 3

E1 is overpowered



Fallback procedure does not lead to the domino effect (power does not decrease toward the end of the sequence)

2. Advanced procedures with a pre-specified hypothesis ordering

Advanced procedures

Bonferroni-based chain procedures

Family of Bonferroni-based procedures that support general logical relationships among null hypotheses

Known as chain procedures since testing algorithm is similar to a chain (discrete random process), e.g., a Markov chain

Also known as procedures that utilize graphical methods (Bretz et al., 2009)

Example 4: Major depressive disorder trial

Phase II data

Phase II data suggest that Dose H is effective but Doses L and M are likely to be less effective

Testing strategy

H_1 (Dose H versus Placebo) is tested first

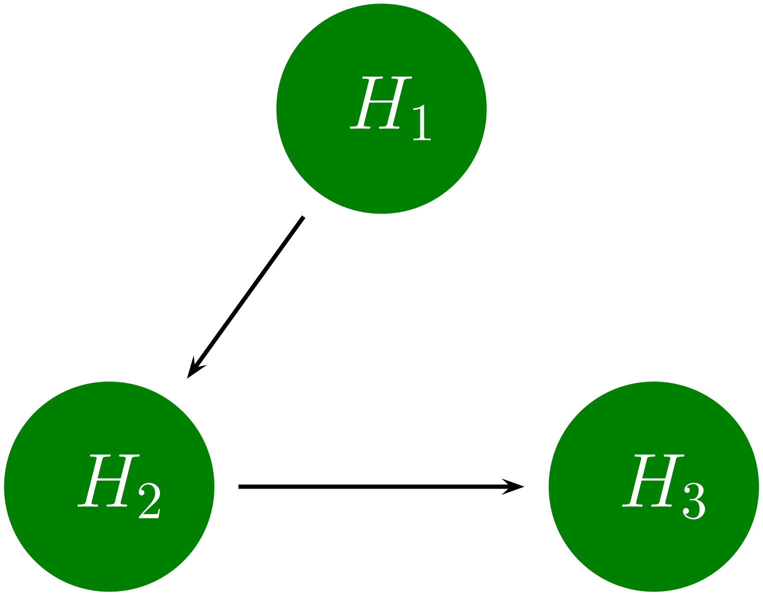
H_2 (Dose M versus Placebo) and H_3 (Dose L versus Placebo) are tested after H_1 is rejected

If H_2 is rejected, the error rate is transferred to H_3

If H_3 is rejected, the error rate is transferred to H_2

Fallback procedure

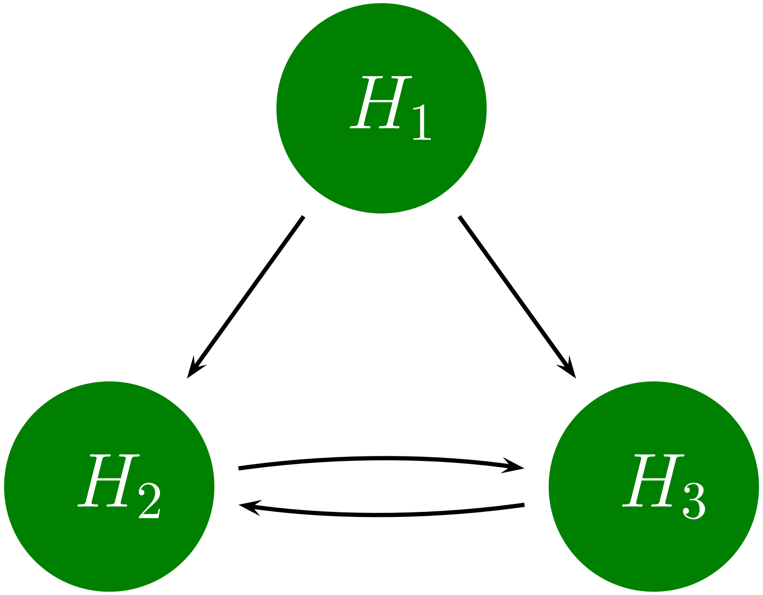
Logical relationships in Example 4



Error rate can transferred only along a single sequence

Logical relationships among three null hypotheses are not taken into account

Logical relationships in Example 4

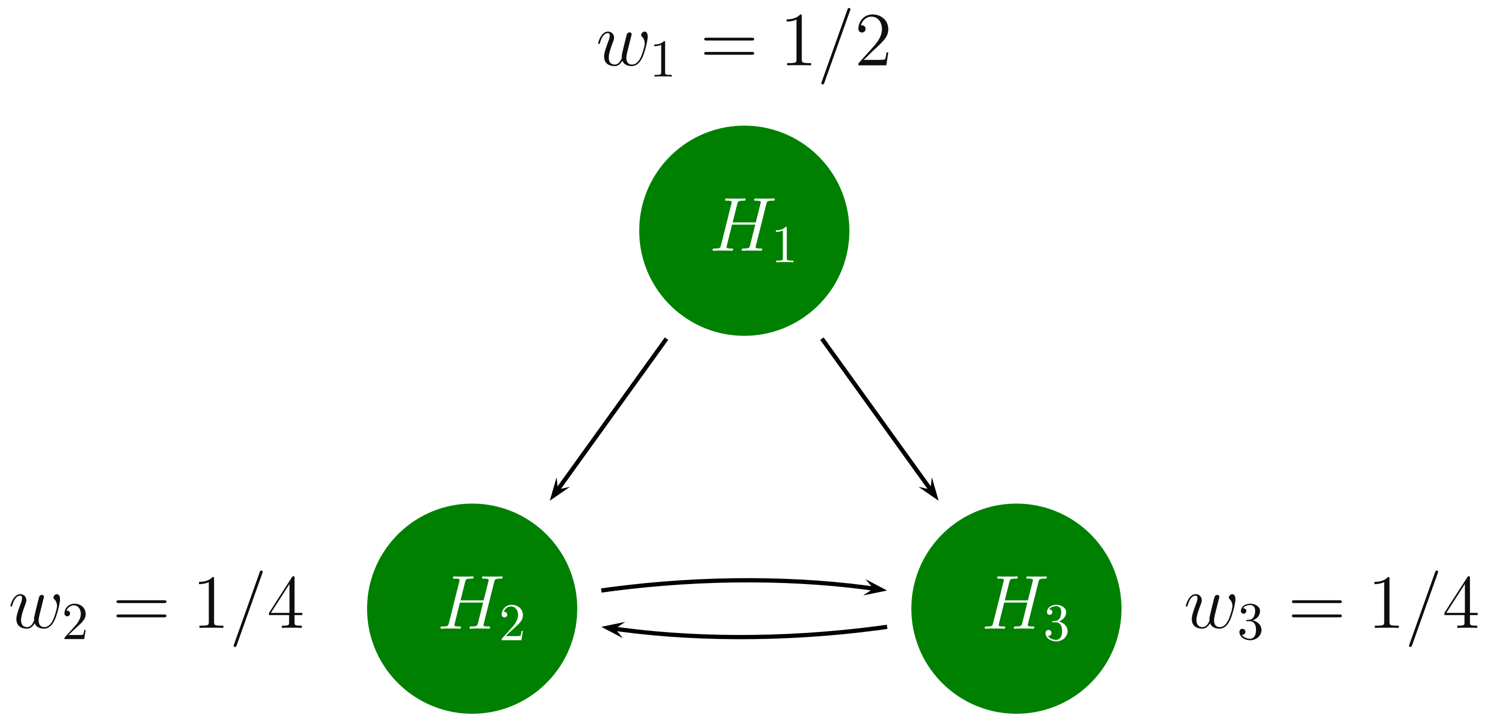


H_1 is tested first followed by H_2 and H_3

Logical relationships among three null hypotheses are incorporated into decision rules

Chain procedure

α allocation rule in Example 4

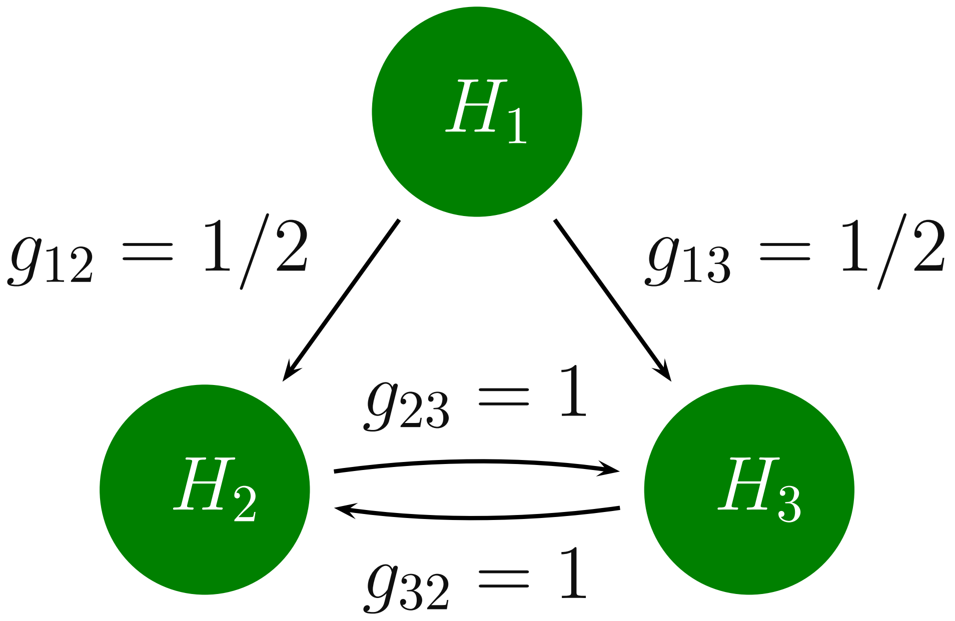


Positive weights are allocated to all hypotheses

H_2 and H_3 can be tested even if H_1 is not rejected

Chain procedure

α propagation rule in Example 4



Error rate is split between H_2 and H_3 after H_1 is rejected

Error rate is transferred between H_2 and H_3

Chain procedure in Example 4

α **allocation rule**

Hypothesis weights

$$W = (1/2, 1/4, 1/4)$$

α **propagation rule**

Transition parameters

$$G = \begin{bmatrix} 0 & 1/2 & 1/2 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

Example 4: Major depressive disorder trial

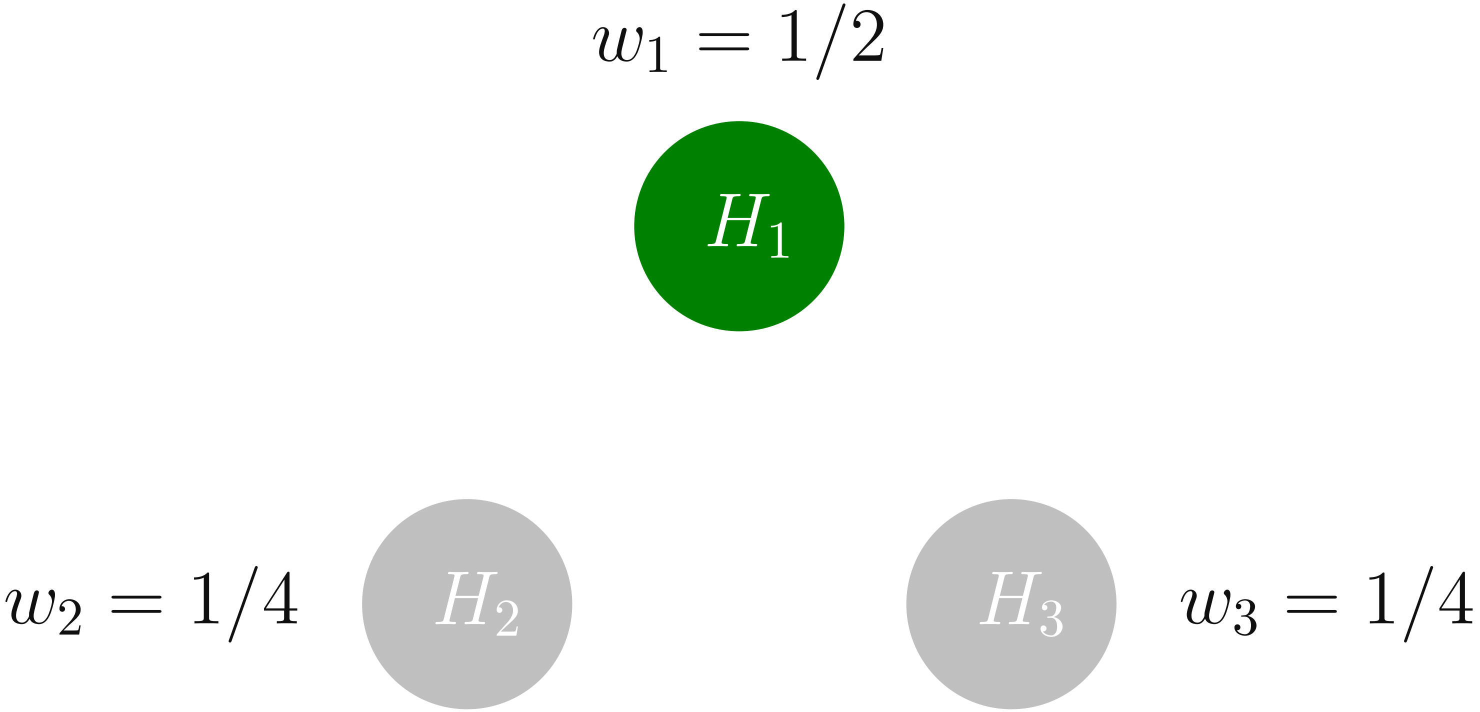
Scenario 4

Comparison	P-value
Dose H vs Placebo (H_1)	0.0098
Dose M vs Placebo (H_2)	0.0114
Dose L vs Placebo (H_3)	0.0211

Evidence of treatment effect at Doses M and H

Chain procedure

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



$w_1 = 1/2$

H_1

$w_2 = 1/4$

H_2

H_3

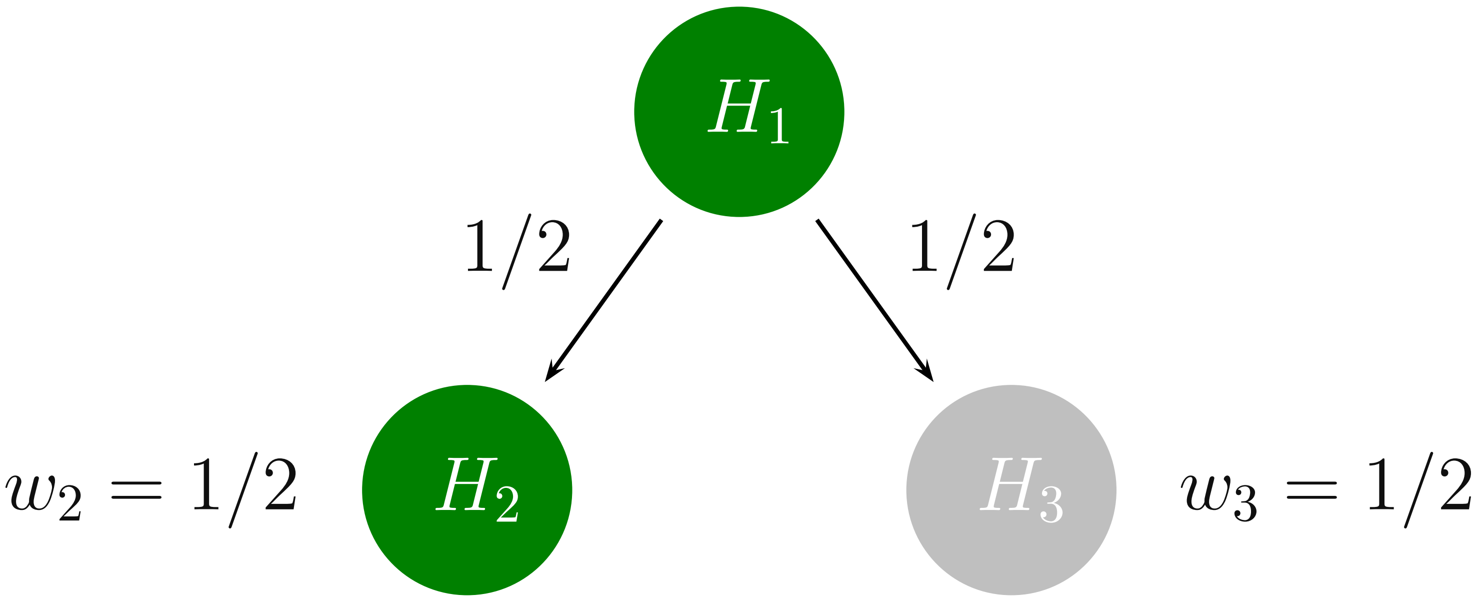
$w_3 = 1/4$

H_1 is rejected since $p_1 = 0.0098 \leq \alpha w_1 = \alpha/2 = 0.0125$

Error rate is transferred from H_1 to H_2 and H_3

Chain procedure

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



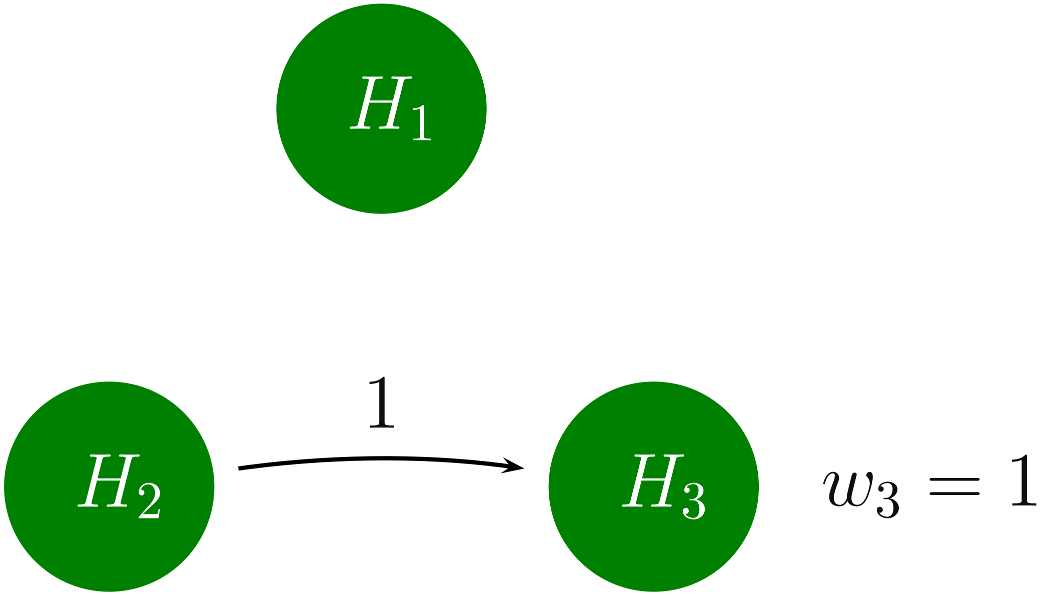
H_2 is rejected since $p_2 = 0.0114 \leq \alpha w_2 = \alpha/2 = 0.0125$

H_3 is not rejected since $p_3 = 0.0211 > \alpha w_3 = \alpha/2 = 0.0125$

Error rate is transferred from H_2 to H_3

Chain procedure

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



H_3 is rejected since $p_3 = 0.0211 \leq \alpha w_3 = \alpha = 0.025$

General chain procedures in Example 4

α allocation rule

Hypothesis weights

$$W = (w_1, w_2, w_3)$$

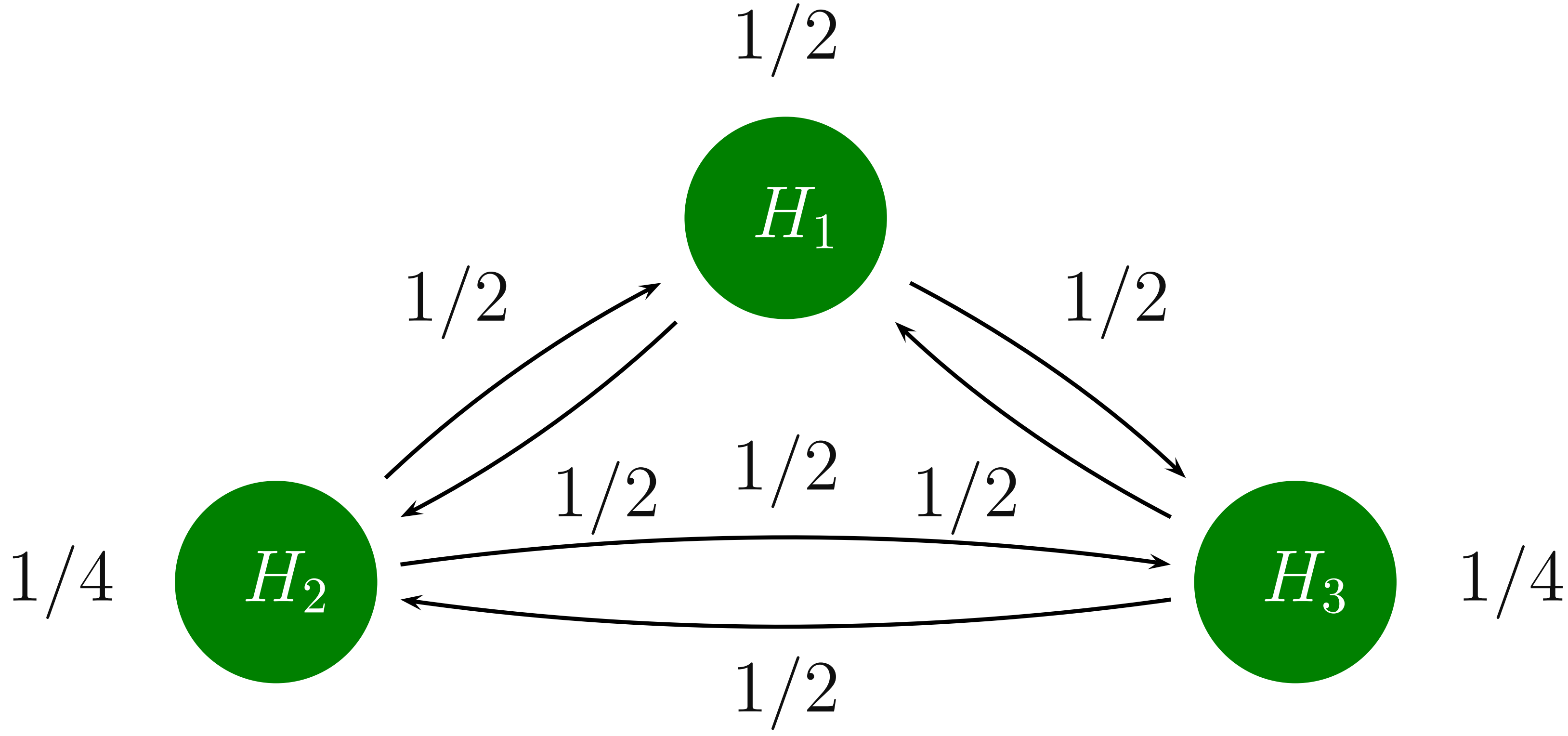
α propagation rule

Transition parameters

$$G = \begin{bmatrix} 0 & g_{12} & g_{13} \\ g_{21} & 0 & g_{23} \\ g_{31} & g_{32} & 0 \end{bmatrix}$$

Chain procedure

α allocation and propagation rules



Error rate is split between H_2 and H_3 after H_1 is rejected

Error rate is transferred between H_2 and H_3

Error rate is transferred from H_2 and H_3 back to H_1

Chain procedures

Re-testing loop

A hypothesis can be re-tested at a higher significance level

Example

H_1 is not rejected in the first step if $p_1 > \alpha w_1$

After H_2 and H_3 are rejected, w_1 is set to 1 and H_1 is rejected if $p_1 \leq \alpha$

Familywise error rate control

FWER is controlled in the strong sense since chain procedures are based on the closure principle

Chain procedures

General testing problem

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

g_{ij}, $i, j = 1, \ldots, m$, Transition parameters

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, $i, j = 1, \ldots, m$

Chain procedures

Notation

M, Index set of non-rejected hypotheses ($M = \{1, \ldots, m\}$ in the beginning)

General decision rules

Step 1: If $p_1 \leq \alpha w_1$, 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

Chain procedures

Notation

M, Index set of non-rejected hypotheses ($M = \{1, \ldots, m\}$ in the beginning)

General decision rules

Step 1: If $p_1 \leq \alpha w_1$, reject H_1, remove the index from M, update the hypothesis weights and transition parameters:

$$w_j = w_j + w_1 g_{1j}, \;\; j \in M$$

$$g_{jk} = \frac{g_{jk} + g_{j1} g_{1k}}{1 - g_{j1} g_{1j}}, \;\; j, k \in M; \;\;\; g_{jk} = 0, \text{otherwise}$$

and go to Step 2

Chain procedures

General decision rules

Steps $i = 2, \ldots, m-1$: If $p_i \leq \alpha w_i$, reject H_i, remove the index from M, update the hypothesis weights and transition parameters and go to Step $i+1$

Step m: If $p_m \leq \alpha w_m$, reject H_m, remove the index from M, update the hypothesis weights and transition parameters and return to Step 1

Continue until the algorithm reaches a stable point (no more hypotheses are rejected)

3. Multiplicity-adjusted p-values

Multiplicity-adjusted levels and p-values

Multiplicity adjustments

Multiplicity adjustments are applied by adjusting significance levels downward or adjusting p-values upward

Adjusted significance levels and p-values

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

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

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

Adjusted p-values

General definition

Overall error rate α is treated as a variable between 0 and 1

Adjusted p-value for H_i is the lowest overall error rate α at which the hypothesis is rejected

Bonferroni procedure

Hypothesis H_i is rejected if $p_i \leq \alpha/m$

Adjusted p-value for H_i is the lowest value of α, ie, $\widetilde{p}_i = mp_i$

Example 4: Major depressive disorder trial

Scenario 1

Comparison	P-value
Dose H vs Placebo (H_1)	$p_1 = 0.0111$
Dose M vs Placebo (H_2)	$p_2 = 0.0065$
Dose L vs Placebo (H_3)	$p_3 = 0.0293$

Example 4: Major depressive disorder 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 at $\alpha = 0.025$

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

4. Simulation study

Simulation study

Selection of hypothesis weights

Hypothesis weights in fallback and general chain procedures must be pre-specified

Hypothesis weight versus importance

Hypothesis importance is an inherent property which reflects its therapeutic value

Weights are assigned by procedures to help quantify hypothesis importance

Selection of hypothesis weights

Clinical trial

Two arms: Experimental treatment versus placebo

Endpoints

Two endpoints (E1 and E2)

Endpoint test statistics follow a bivariate normal distribution

Selection of hypothesis weights

Endpoint roles

Formally, E1 is primary endpoint and E2 is secondary endpoint

Each endpoint supports an independent regulatory claim

Sufficient to “win” on only one endpoint

Examples

Cardiovascular trials: Exercise capability/Quality of life (primary) and Mortality (secondary)

Oncology trials: Overall survival (primary) and Progression-free survival (secondary)

Selection of hypothesis weights

Multiple testing procedure

Family of fallback procedures:

Procedure A: $w_1 = 1$, $w_2 = 0$ (fixed-sequence procedure)

Procedure B: $w_2 = 0.9$, $w_2 = 0.1$

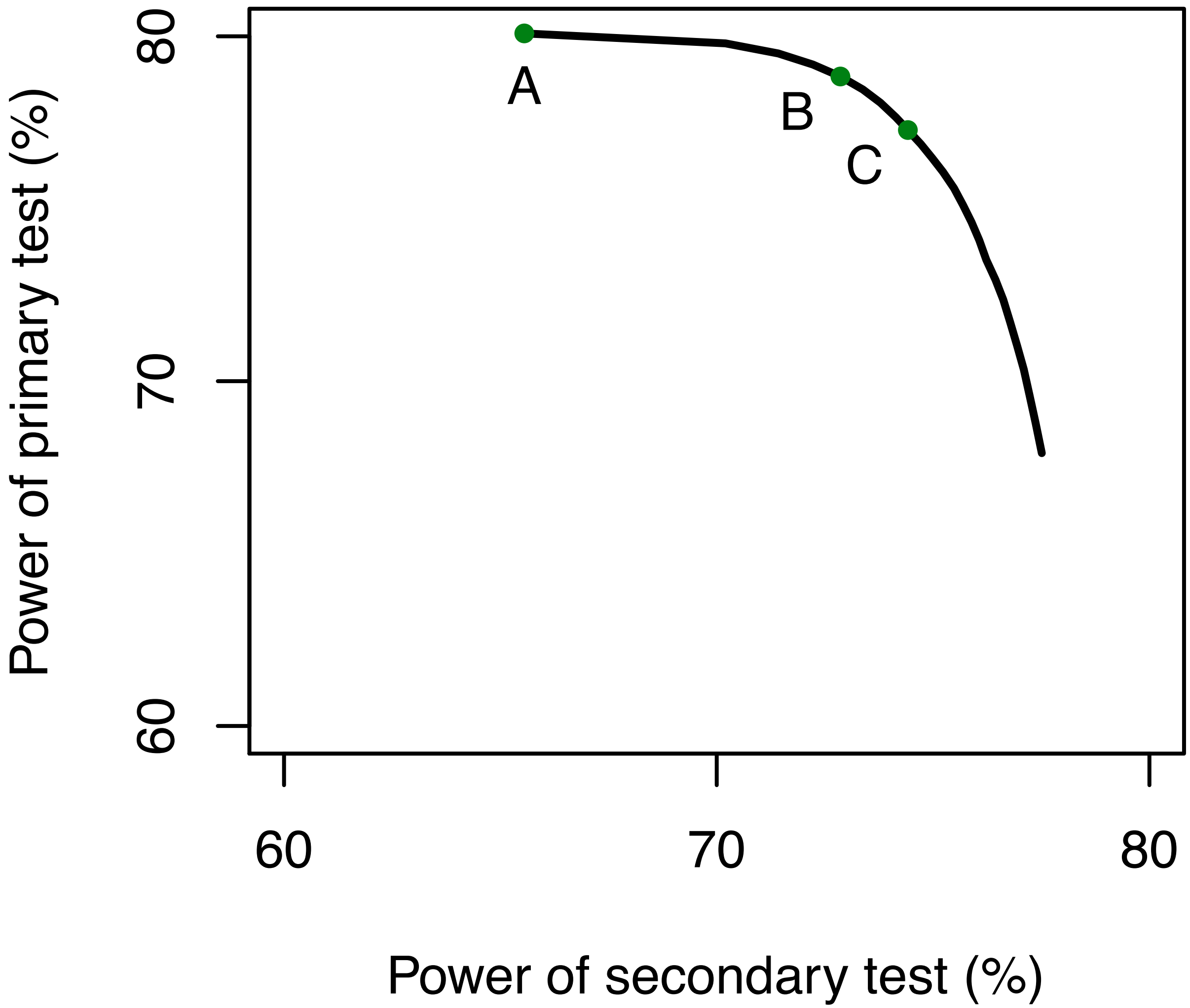
Procedure C: $w_3 = 0.7$, $w_2 = 0.3$

Correlation

Weakly correlated endpoints (correlation=0.2)

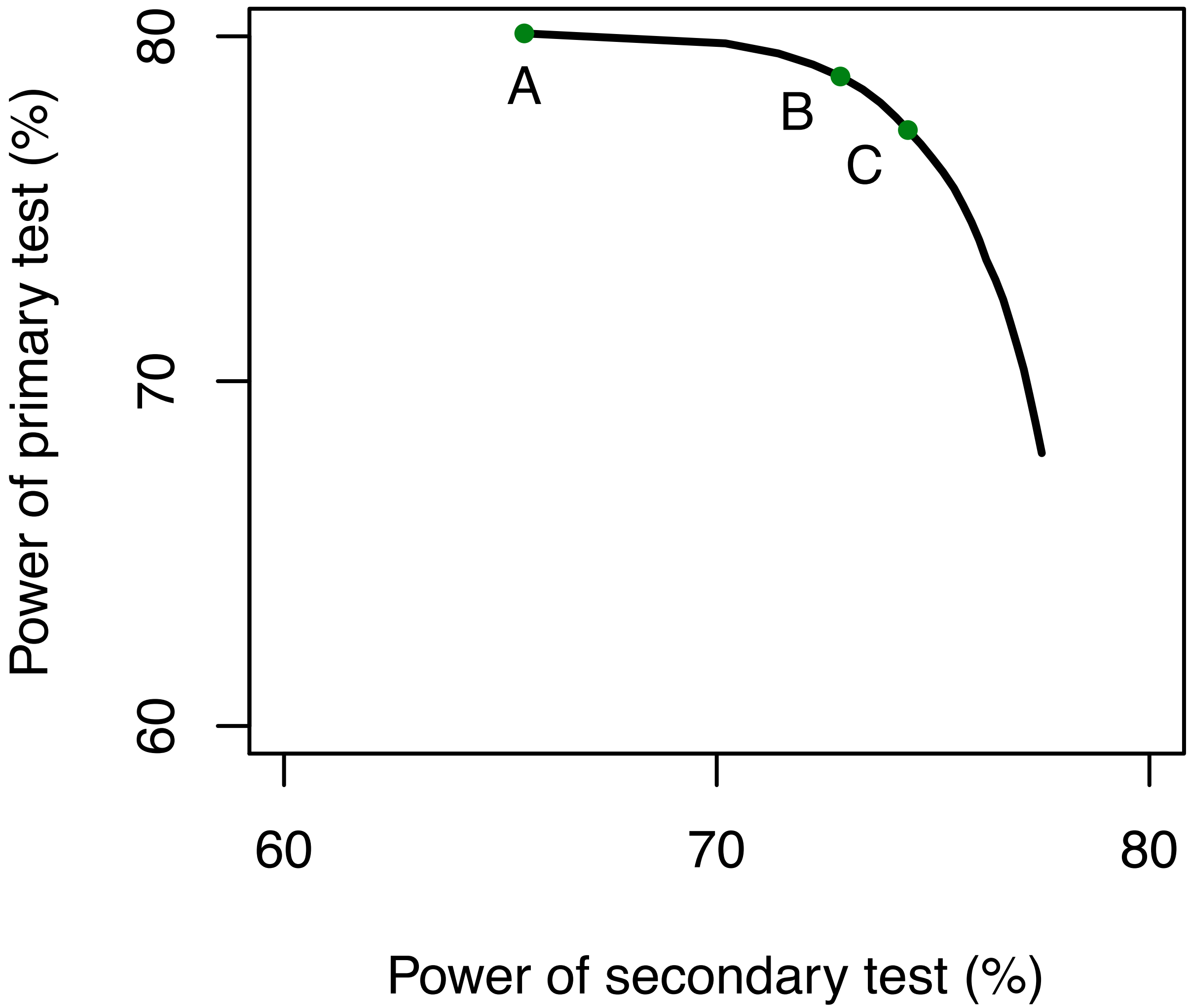
Strongly correlated endpoints (correlation=0.5)

Weakly correlated endpoints



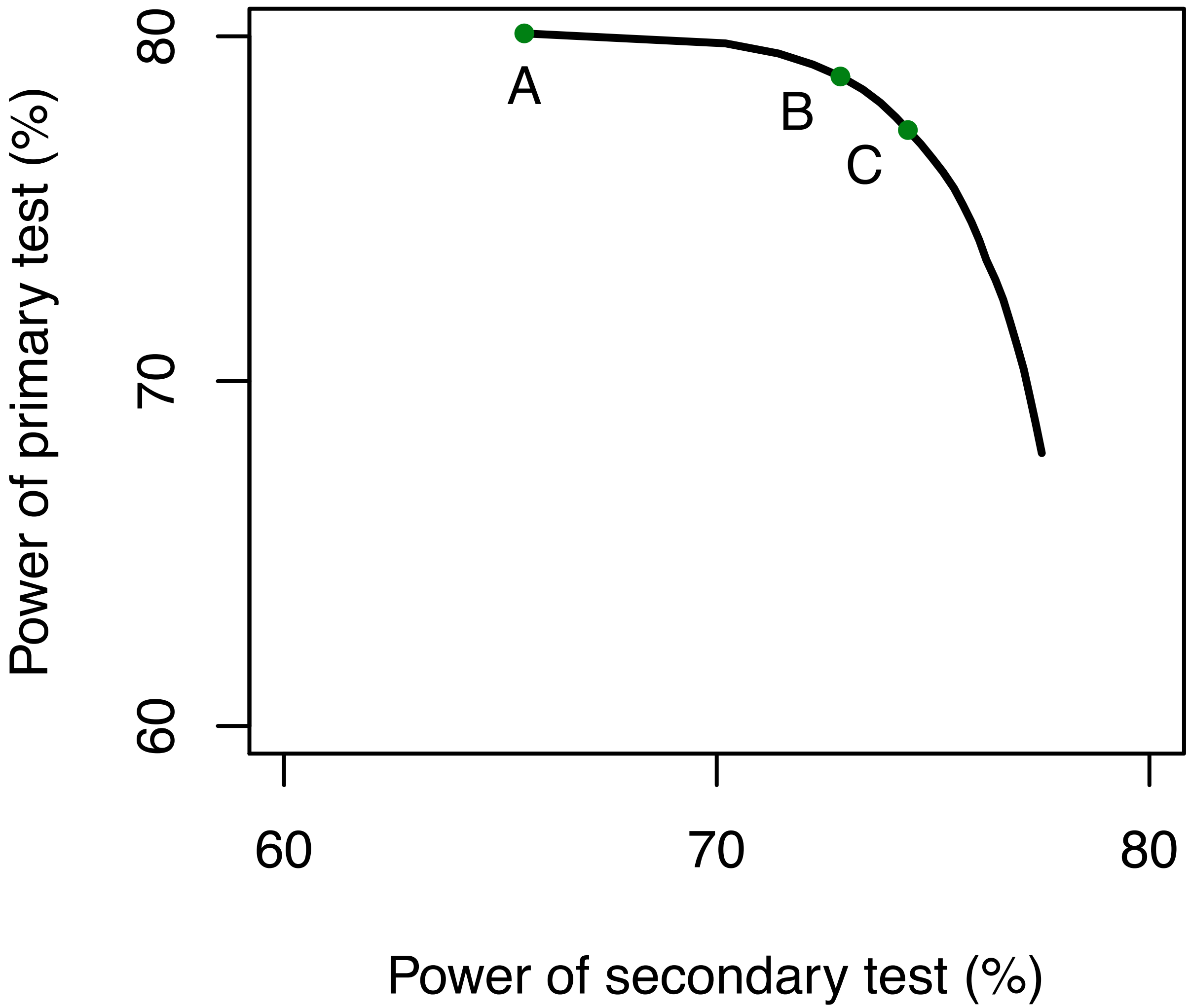
Procedure A: $w_1 = 1$, $w_2 = 0$; Procedure B: $w_1 = 0.9$, $w_2 = 0.1$; Procedure C: $w_1 = 0.7$, $w_2 = 0.3$

Weakly correlated endpoints



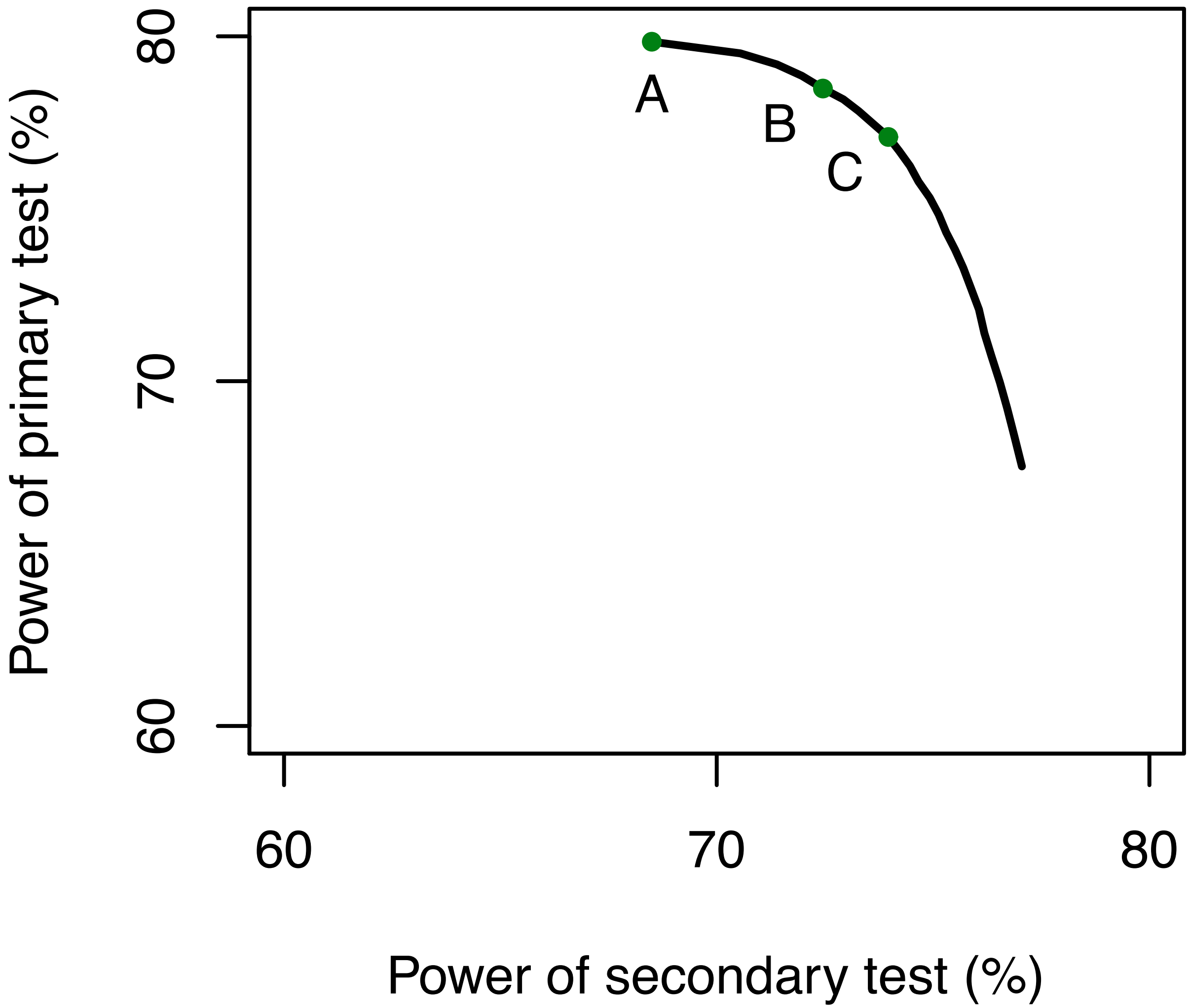
Procedure A: Primary test 80%, Secondary test 66%.
Procedure B: Primary test 79%, Secondary test 73%.

Weakly correlated endpoints



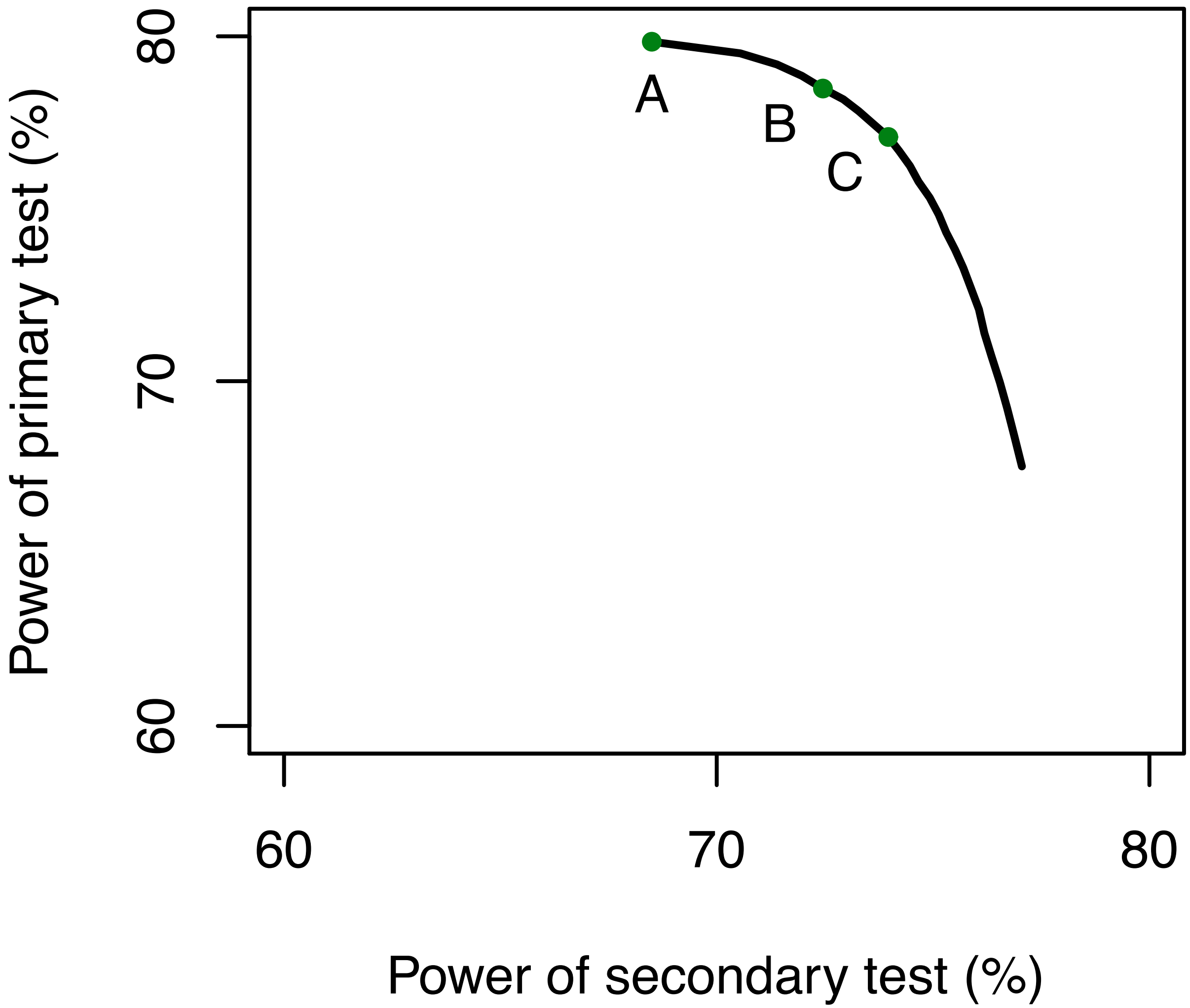
Procedure B: Primary test 79%, Secondary test 73%.
Procedure C: Primary test 77%, Secondary test 75%.

Strongly correlated endpoints



Procedure A: $w_1 = 1$, $w_2 = 0$; Procedure B: $w_1 = 0.9$, $w_2 = 0.1$; Procedure C: $w_1 = 0.7$, $w_2 = 0.3$

Strongly correlated endpoints



Procedure A: Primary test 80%, Secondary test 68%.
Procedure B: Primary test 78%, Secondary test 72%.

Optimality criteria

Formal optimality criteria

Facilitate the process of specifying hypothesis weights

Commonly used criteria

Simple and generalized disjunctive power

Weighted power

More criteria will be defined in Module F (Sample size calculations)

Optimality criteria

Simple disjunctive power

Probability of rejecting at least one of the two hypotheses

$\psi = P\{\text{Reject } H_1 \text{ or } H_2\}$

Properties

Does not differentiate between clinically distinct outcomes, e.g.

- H_1 is rejected but H_2 is not rejected
- H_1 and H_2 are both rejected

Optimality criteria

Weighted power

Weighted sum of marginal power functions

$\psi = v_1 P\{\text{Reject } H_1\} + v_2 P\{\text{Reject } H_2\}$

v_1 and v_2, Hypothesis importance ($v_1 > 0$, $v_2 > 0$ and $v_1 + v_2 = 1$)

Properties

More sensitive than simple disjunctive power

Selection of hypothesis weights

Endpoint importance

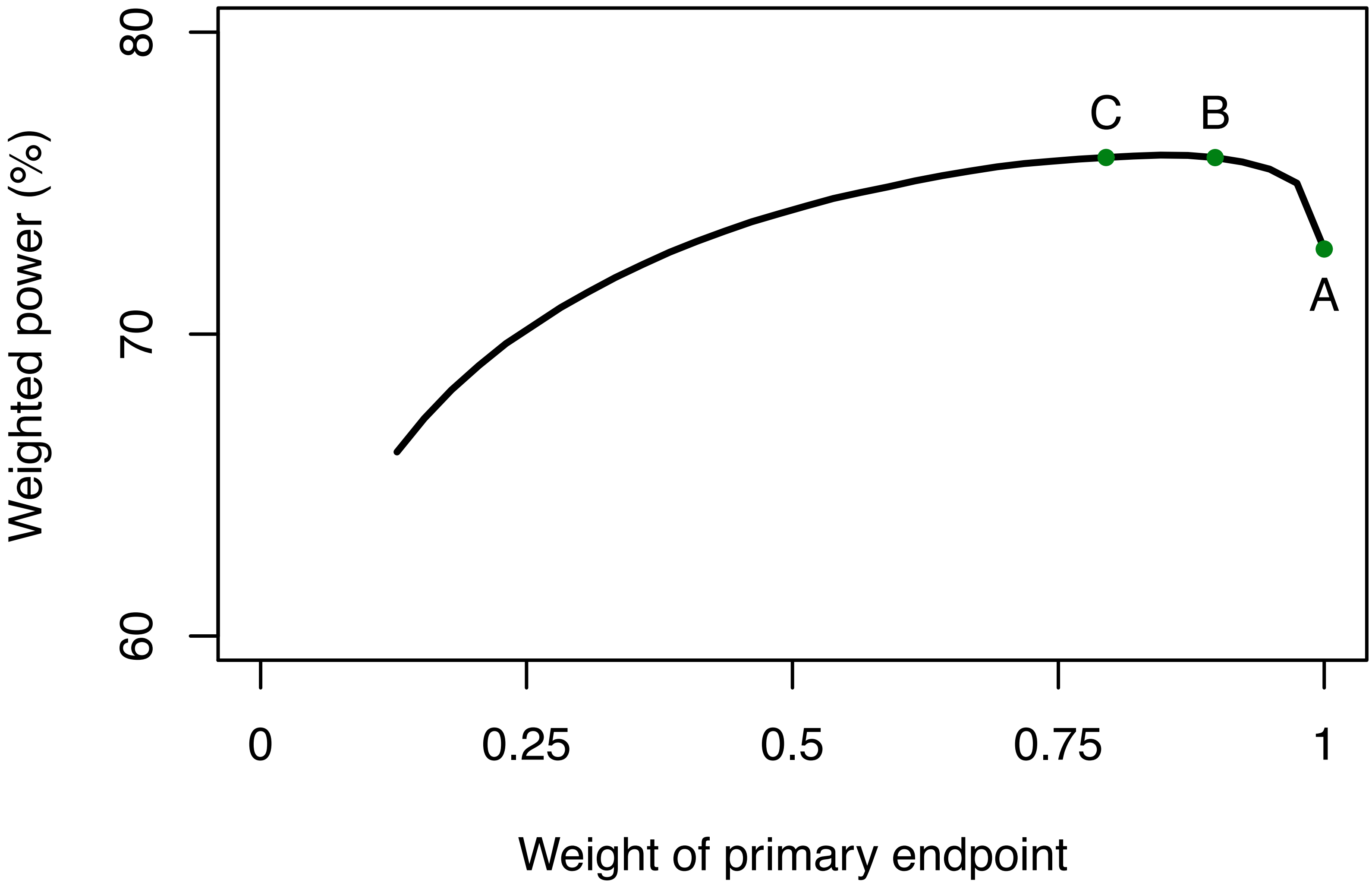
Equally important endpoints ($v_1 = 0.5$, $v_2 = 0.5$)

Unequally important endpoints ($v_1 = 0.7$, $v_2 = 0.3$)

Correlation

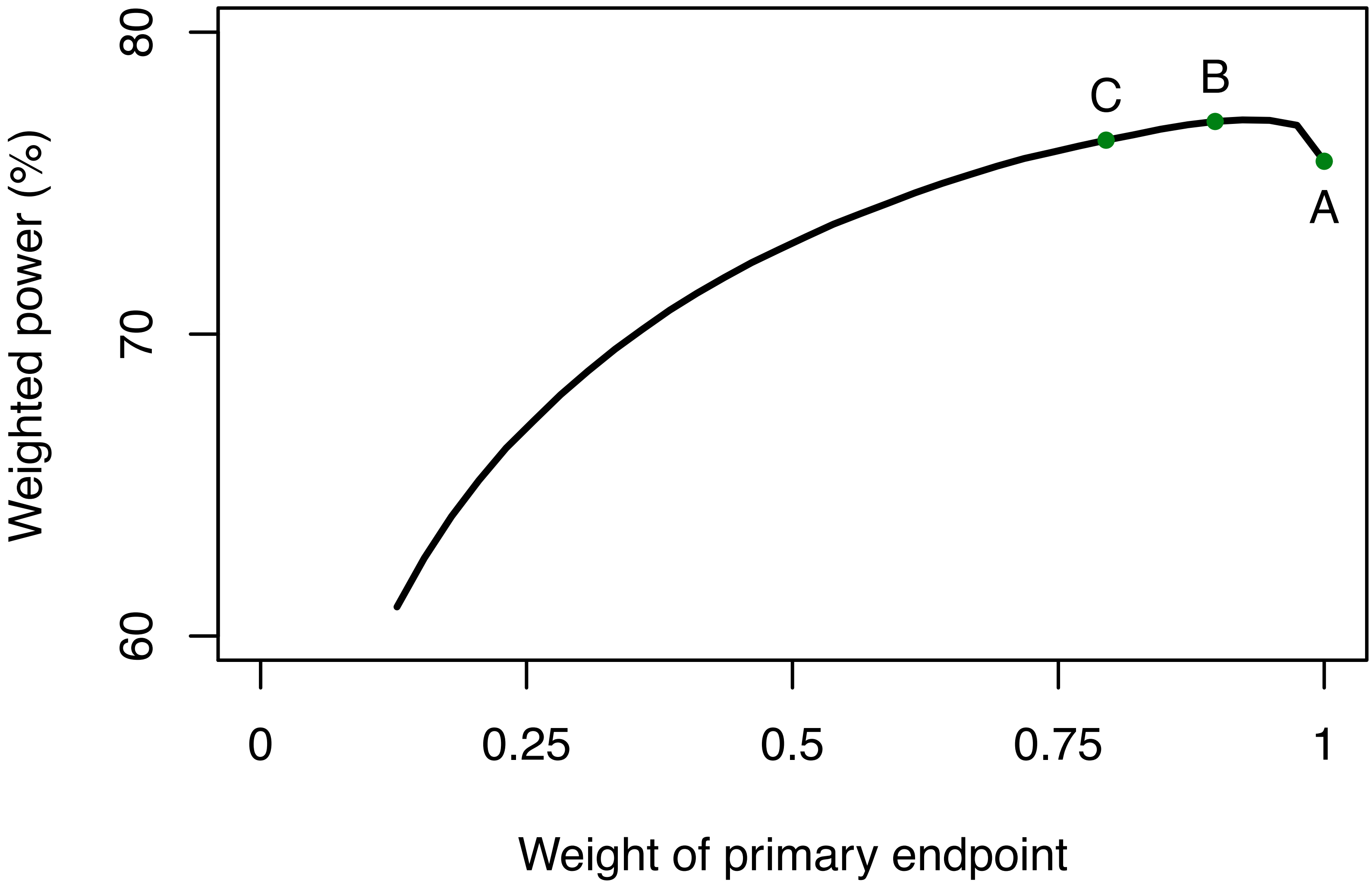
Weakly correlated endpoints (correlation=0.2)

Equally important endpoints



Procedure A: $w_1 = 1$, $w_2 = 0$; Procedure B: $w_1 = 0.9$, $w_2 = 0.1$; Procedure C: $w_1 = 0.7$, $w_2 = 0.3$

Unequally important endpoints



Procedure A: $w_1 = 1$, $w_2 = 0$; Procedure B: $w_1 = 0.9$, $w_2 = 0.1$; Procedure C: $w_1 = 0.7$, $w_2 = 0.3$

Summary

Nonparametric procedures

Pre-specified hypothesis ordering

Fixed-sequence, fallback and chain procedures

Chain procedures

Broad class of multiple testing procedures which includes fixed-sequence and fallback procedures

Flexible procedures that enable trial sponsors to incorporate relevant clinical information (logical restrictions)

Quantitative assessment

Quantitative evaluation and comparison

Comprehensive simulation-based quantitative evaluation of candidate multiple testing procedures to maximize power and select an optimal set of procedure parameters

Performed at the design stage of all confirmatory Phase III trials to ensure robust performance of chosen multiple testing procedure under a wide range of plausible scenarios

Module B

Further reading

Multiple Testing Problems in Pharmaceutical Statistics (edited by Alex Dmitrienko, Ajit Tamhane and Frank Bretz)

Sections

1 (Simple procedures with a pre-specified hypothesis ordering) and 2 (Advanced procedures with a pre-specified hypothesis ordering): Chapter 2

Module C
Semiparametric Procedures

Outline

1. Basic procedures

Bonferroni and Simes procedures (Simes, 1986)

2. Procedures with a data-driven hypothesis ordering

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

Outline

3. Simulation study

Comparison of Bonferroni, fixed-sequence and Hochberg procedures

4. Software implementation

Software implementation of nonparametric and semiparametric procedures in SAS and R

1. Basic procedures

Basic single-step and global procedures

Bonferroni procedure

Single-step procedure

Simes procedure

Global procedure

Objective

Bonferroni and Simes procedures will be introduced to provide a foundation for more powerful stepwise multiple testing procedures

Bonferroni and Simes procedures

Bonferroni procedure

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

Simes global procedure

Tests the global null hypothesis

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

Simes procedure 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: Major depressive disorder trial

Three dose-placebo comparisons

Comparison	P-value
Dose H vs Placebo (H_1)	$p_1 = 0.0111$
Dose M vs Placebo (H_2)	$p_2 = 0.0065$
Dose L vs Placebo (H_3)	$p_3 = 0.0293$

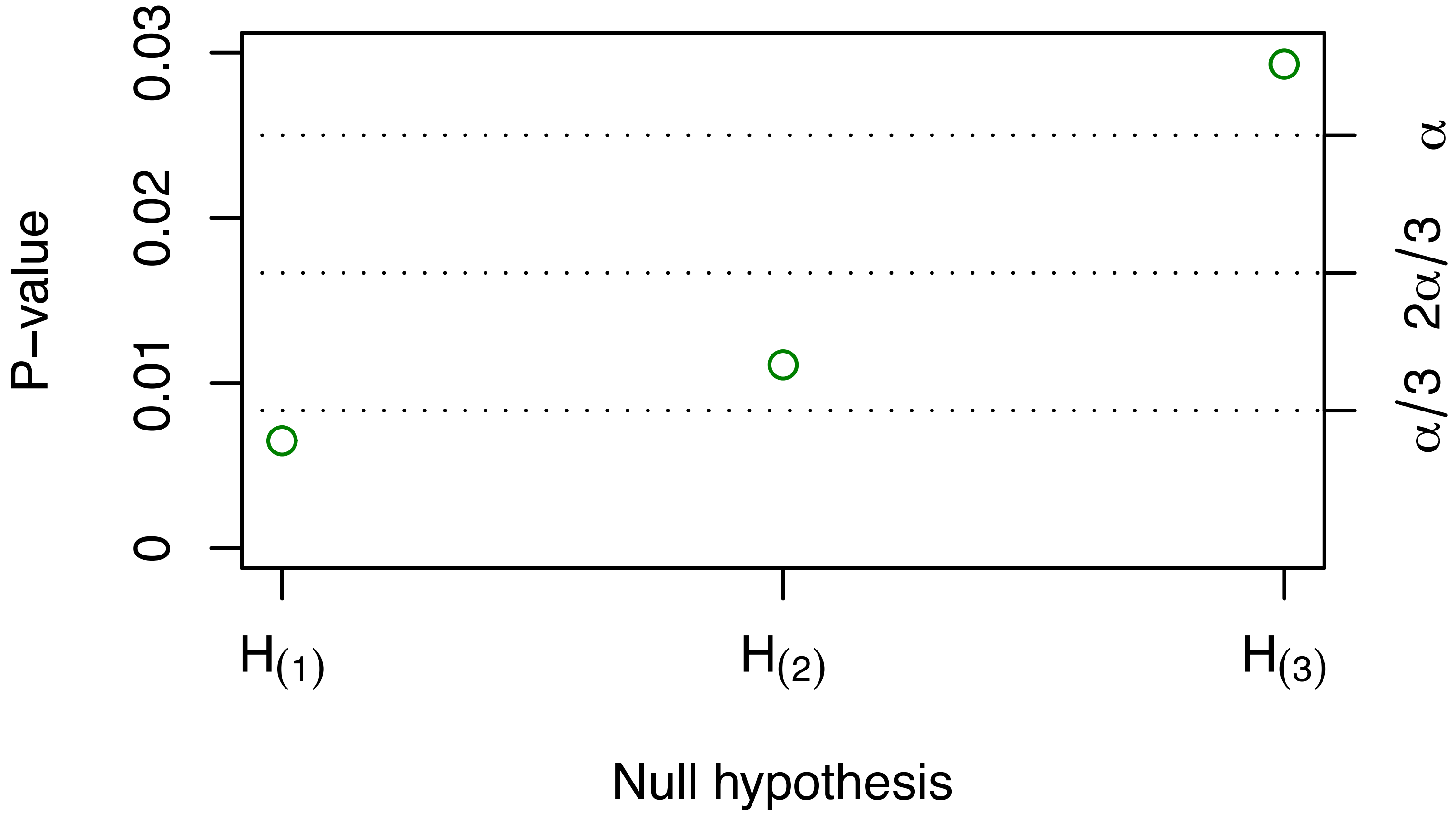
Evidence of treatment effect at Doses M and H

$p_{(1)} = p_2 = 0.0065$, $p_{(2)} = p_1 = 0.0111$ 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 procedure rejects $H_{(1)} = H_2$

Simes procedure rejects the global null hypothesis H_I

Type I error rate control

Bonferroni procedure

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

Simes global procedure

Simes procedure is more powerful but controls Type I error rate only for some joint distributions

Simes procedure may lead to Type I error rate inflation and its properties will be discussed later in this module

Type I error rate control

Example

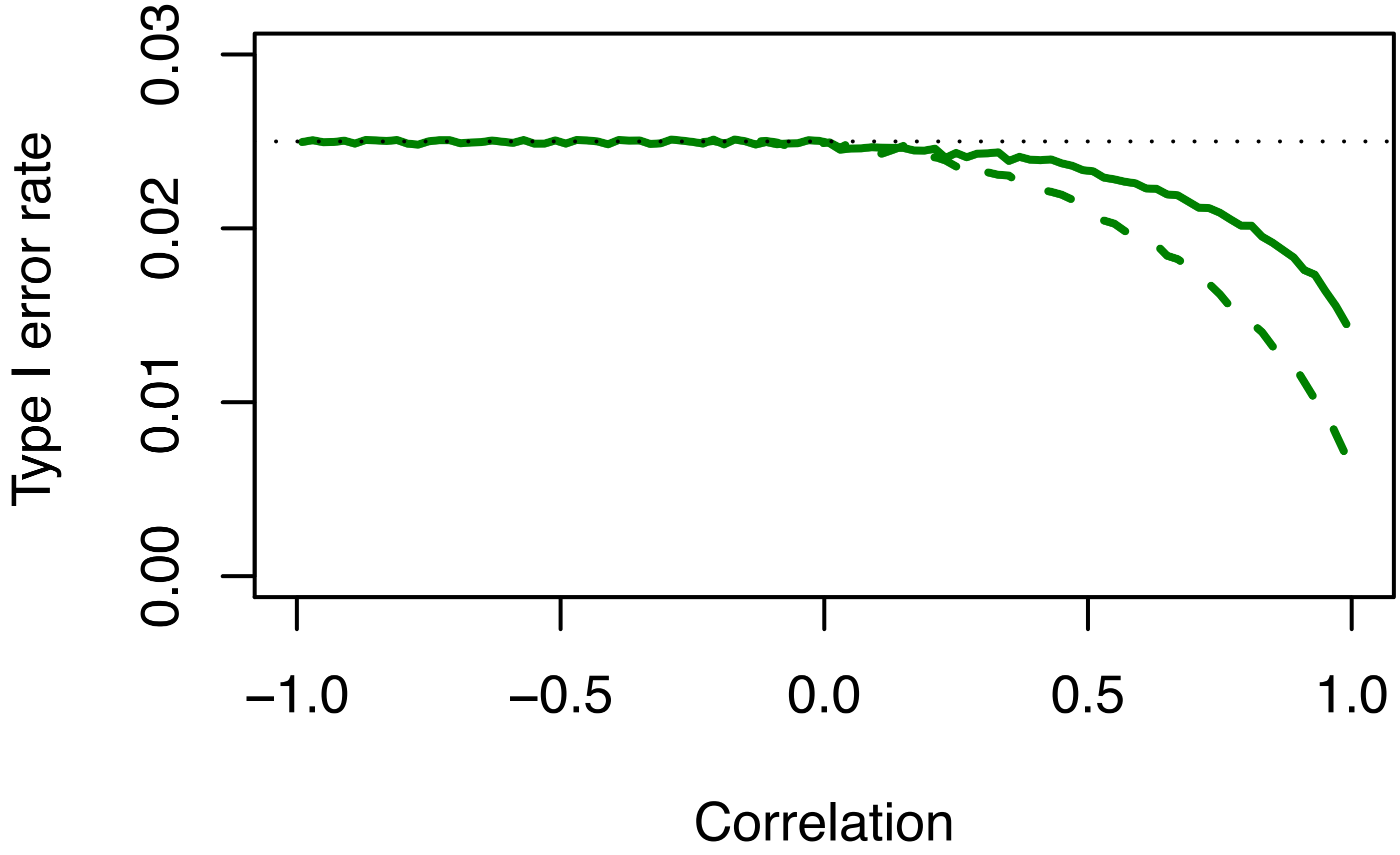
Multiple testing problem with $m = 2$ and $m = 5$ comparisons

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

$\alpha = 0.025$, Type I error rate

Bonferroni procedure

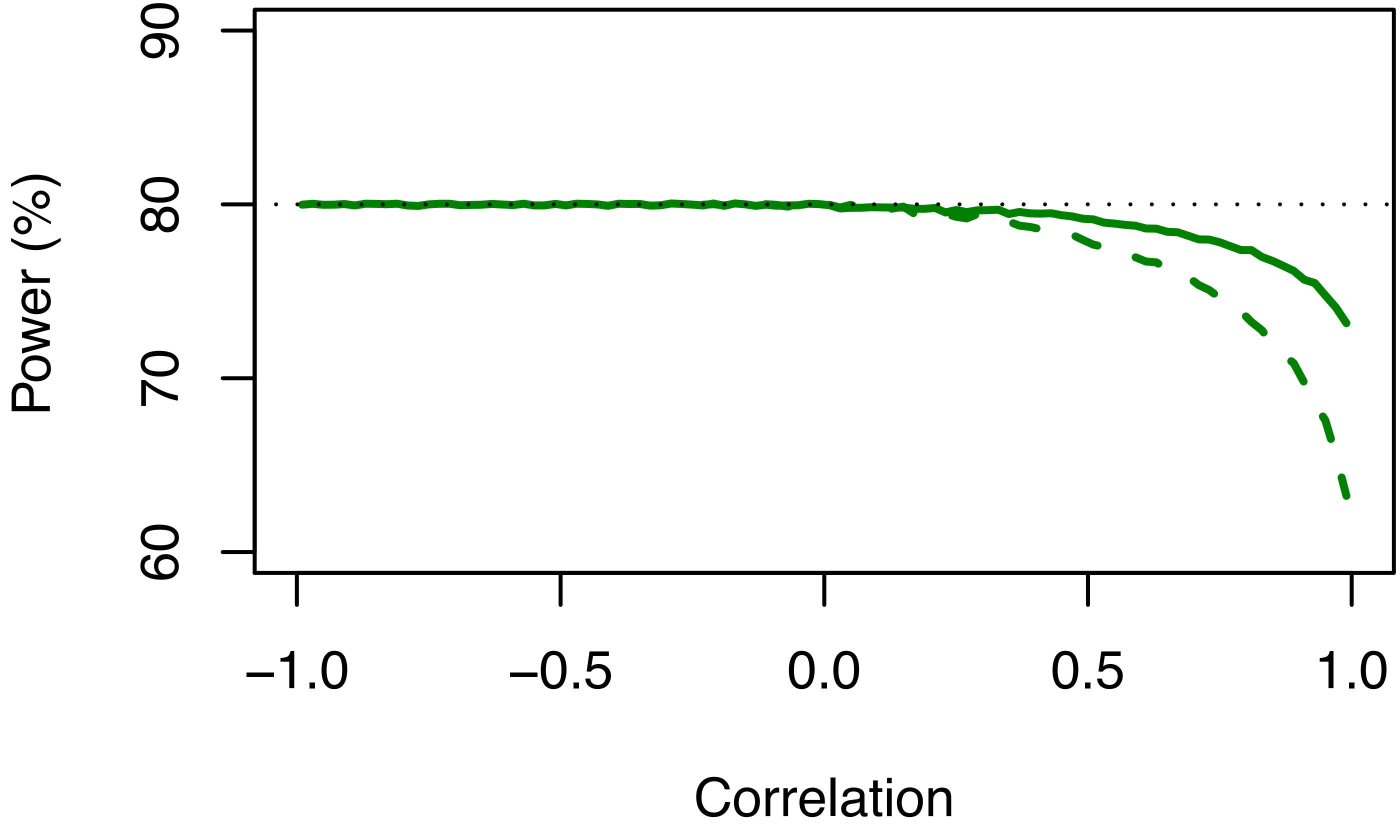
Type I error rate



— 2 comparisons, - - - 5 comparisons

Bonferroni procedure

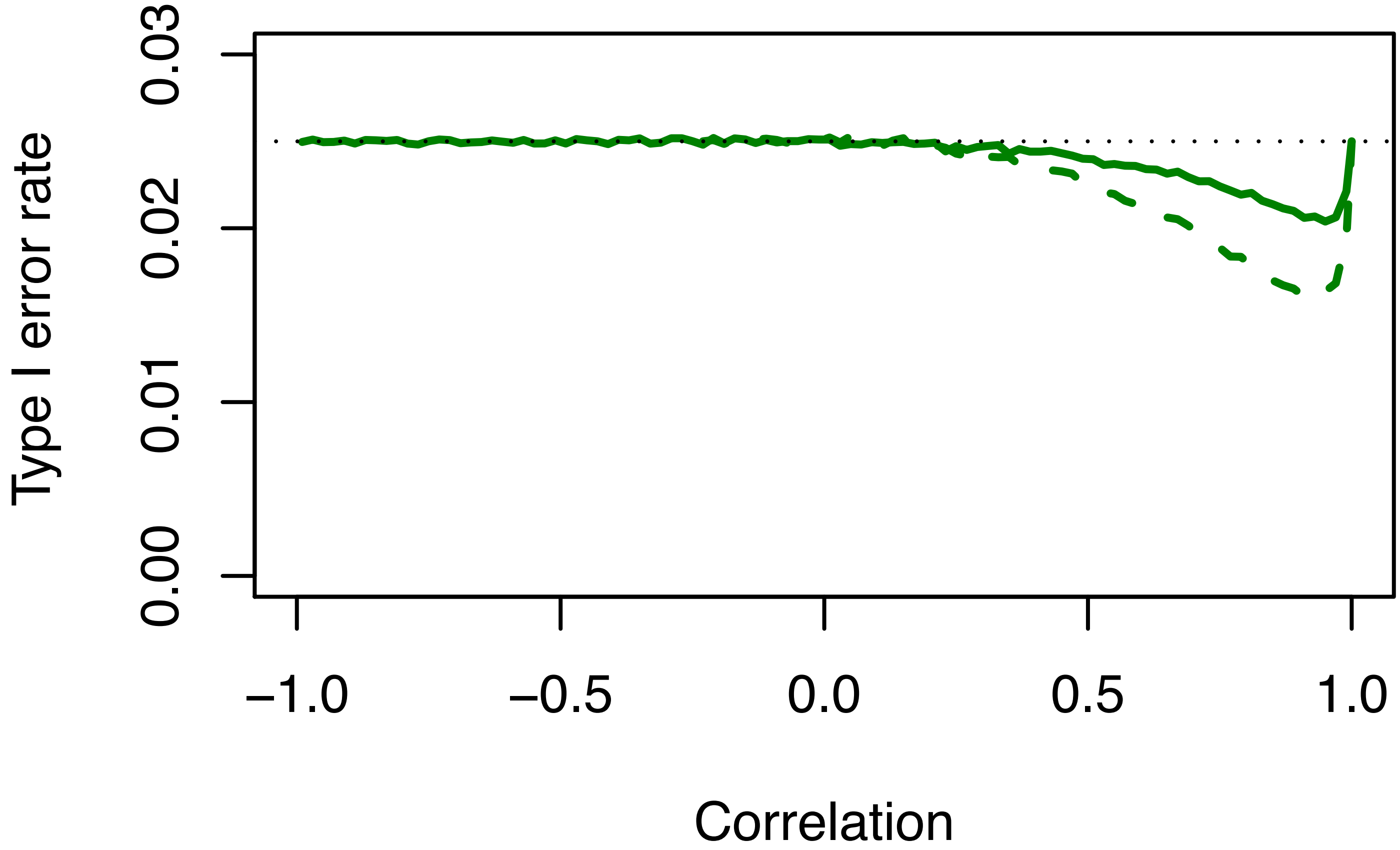
Power



— 2 comparisons, - - - 5 comparisons

Simes procedure

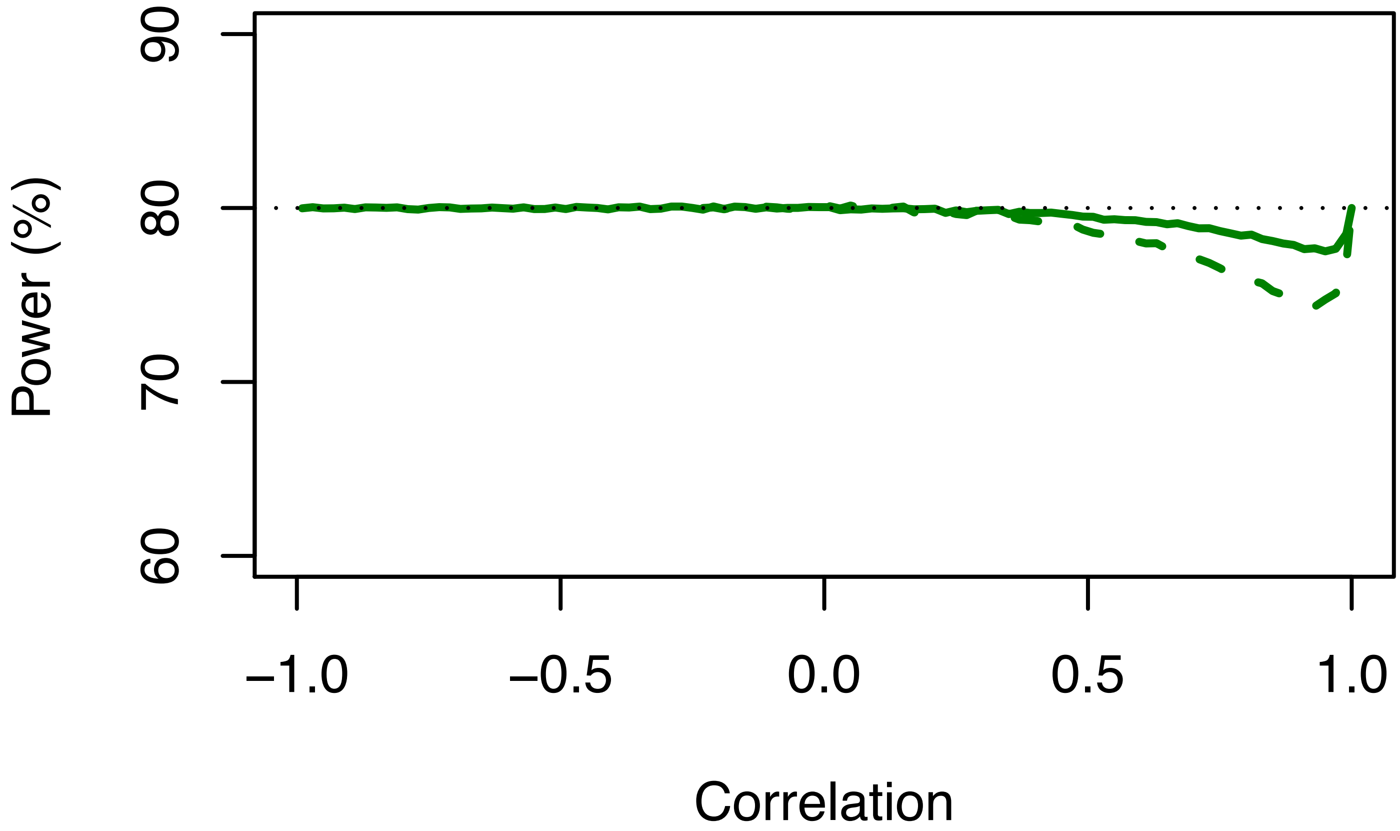
Type I error rate



— 2 comparisons, - - - 5 comparisons

Simes procedure

Power



— 2 comparisons, - - - 5 comparisons

2. Procedures with a data-driven hypothesis ordering

Data-driven hypothesis ordering

Distributional relationships

No distributional assumptions (any joint distribution of hypothesis test statistics)

Logical relationships

Stepwise procedures with a data-driven hypothesis ordering

Order in which null hypotheses are tested is not fixed

Stepwise procedures

Bonferroni-based procedure

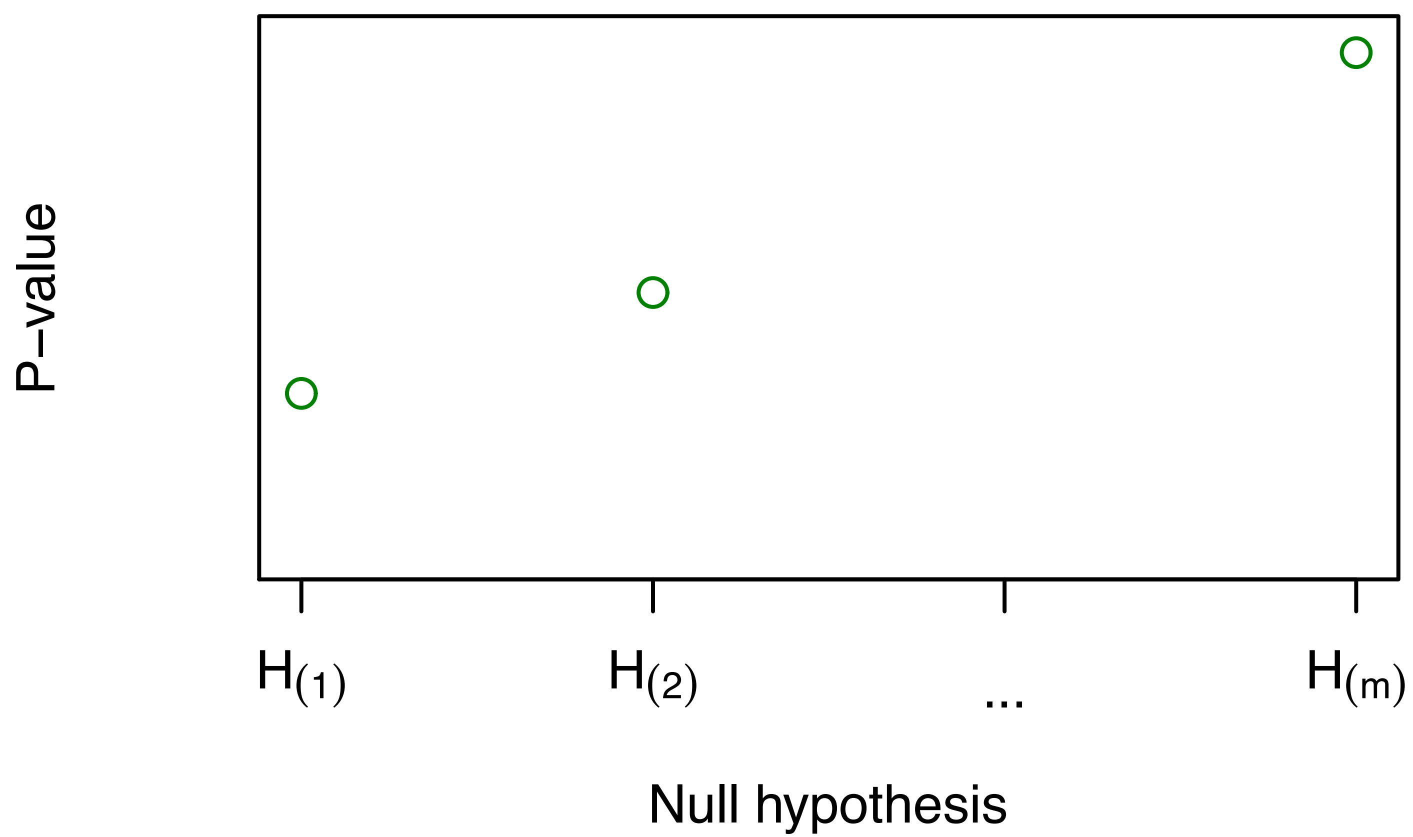
Holm procedure (step-down procedure)

Simes-based procedures

Hommel and Hochberg procedures (step-up procedures)

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-down procedures

Adjusted significance levels

Monotonically increasing sequence of adjusted significance levels

$$\alpha_{(1)} < \ldots < \alpha_{(m)}$$

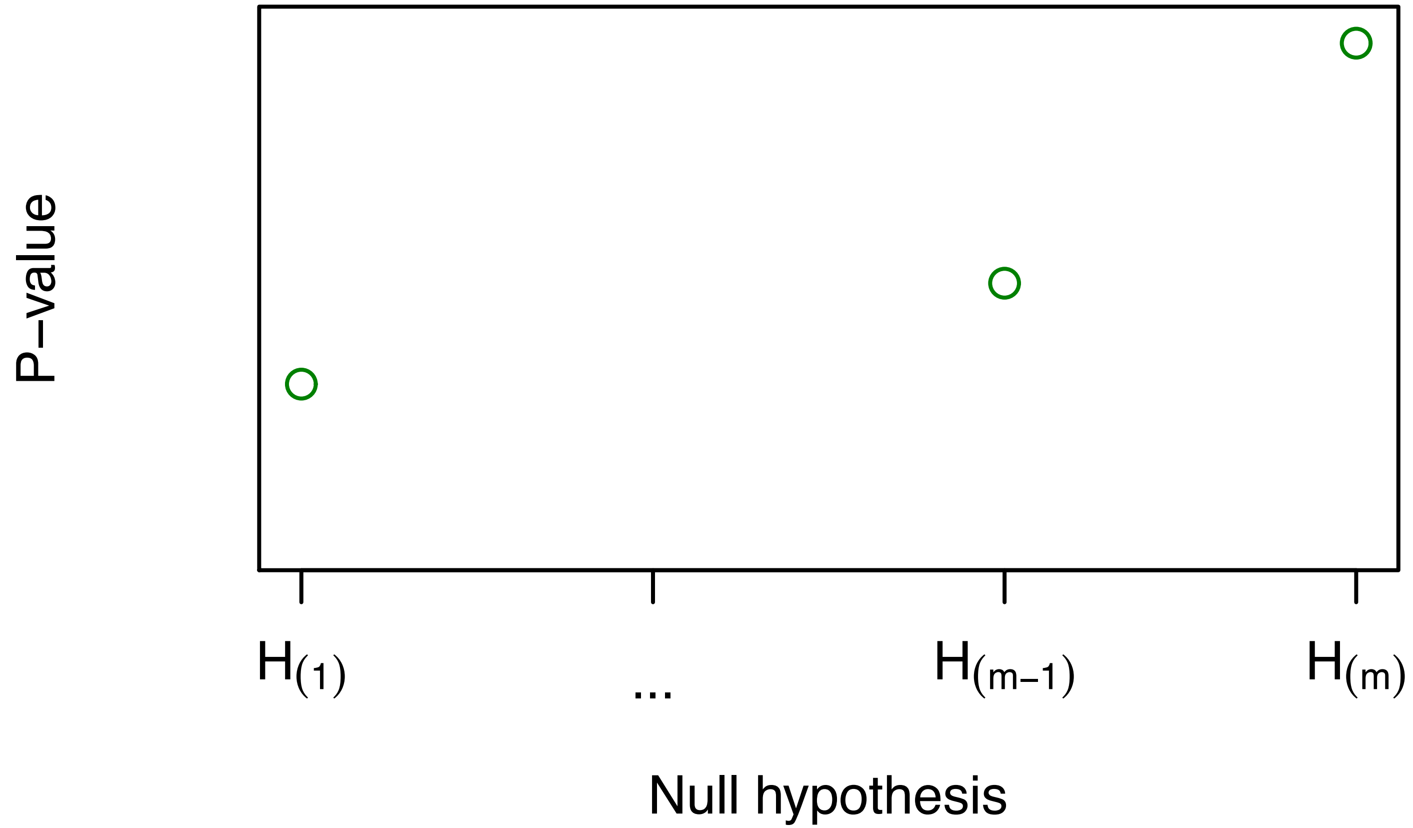
Conjunctive decision rules

Reject ordered hypothesis $H_{(i)}$ if

$$p_{(1)} \leq \alpha_{(1)} \text{ and } \ldots \text{ and } p_{(i)} \leq \alpha_{(i)}$$

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

Step-up procedures

Adjusted significance levels

Monotonically increasing sequence of adjusted significance levels

$$\alpha_{(1)} < \ldots < \alpha_{(m)}$$

Disjunctive decision rules

Reject ordered hypothesis $H_{(i)}$ if

$$p_{(i)} \leq \alpha_{(i)} \text{ or } \ldots \text{ or } p_{(m)} \leq \alpha_{(m)}.$$

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 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

Hochberg procedure

Simes-based procedure

Hochberg procedure is a closed testing procedure derived from simplified Simes procedure (which is less powerful than regular Simes procedure)

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 null 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 null hypotheses and stop

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

Example 4: Major depressive disorder trial

Scenario 2

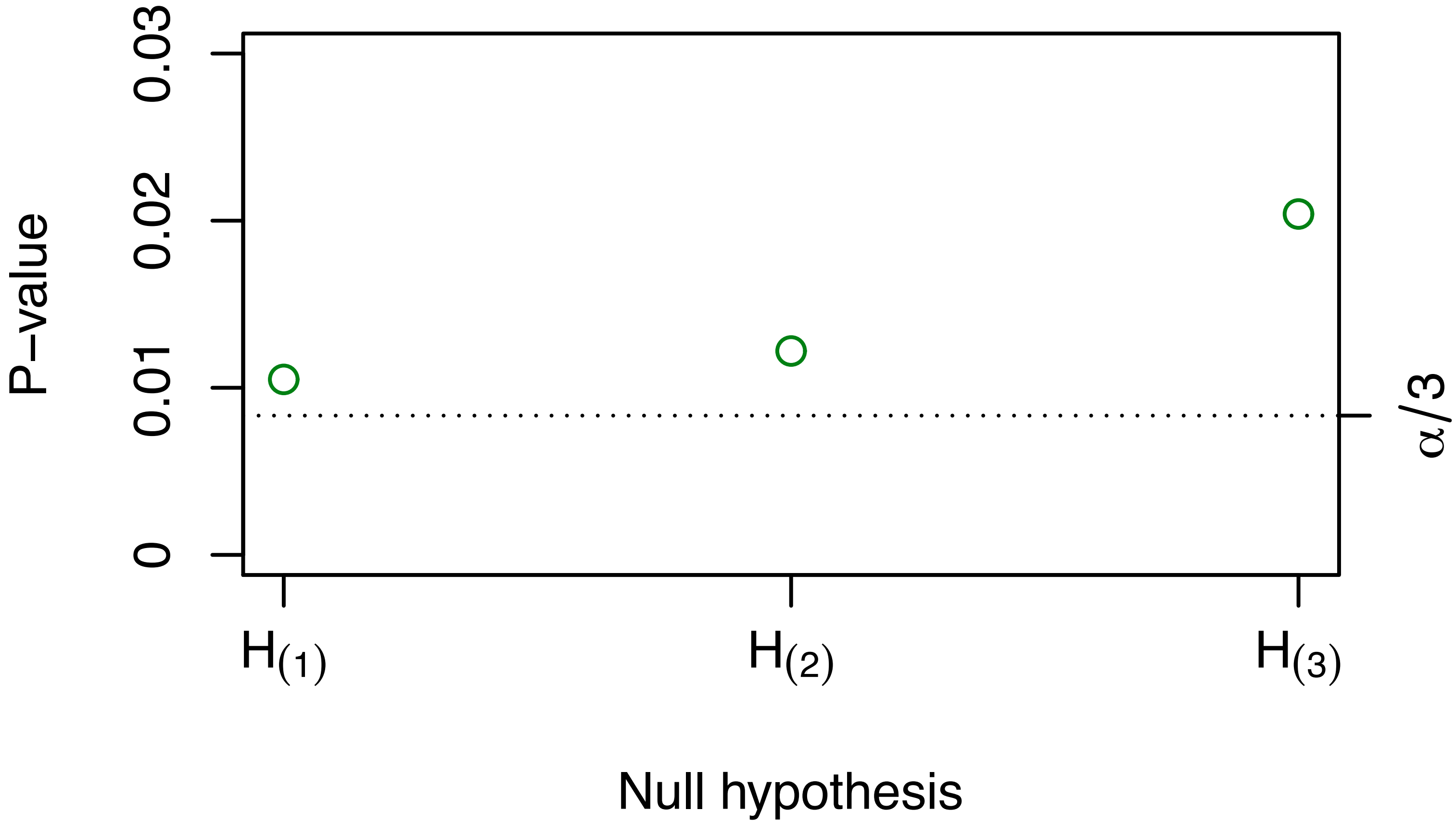
Comparison	P-value
Dose H vs Placebo (H_1)	0.0105
Dose M vs Placebo (H_2)	0.0122
Dose L vs Placebo (H_3)	0.0204

$p_{(1)} = p_1 = 0.0105$, $p_{(2)} = p_2 = 0.0122$ and $p_{(3)} = p_3 = 0.0204$, ordered p-values

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

Holm procedure

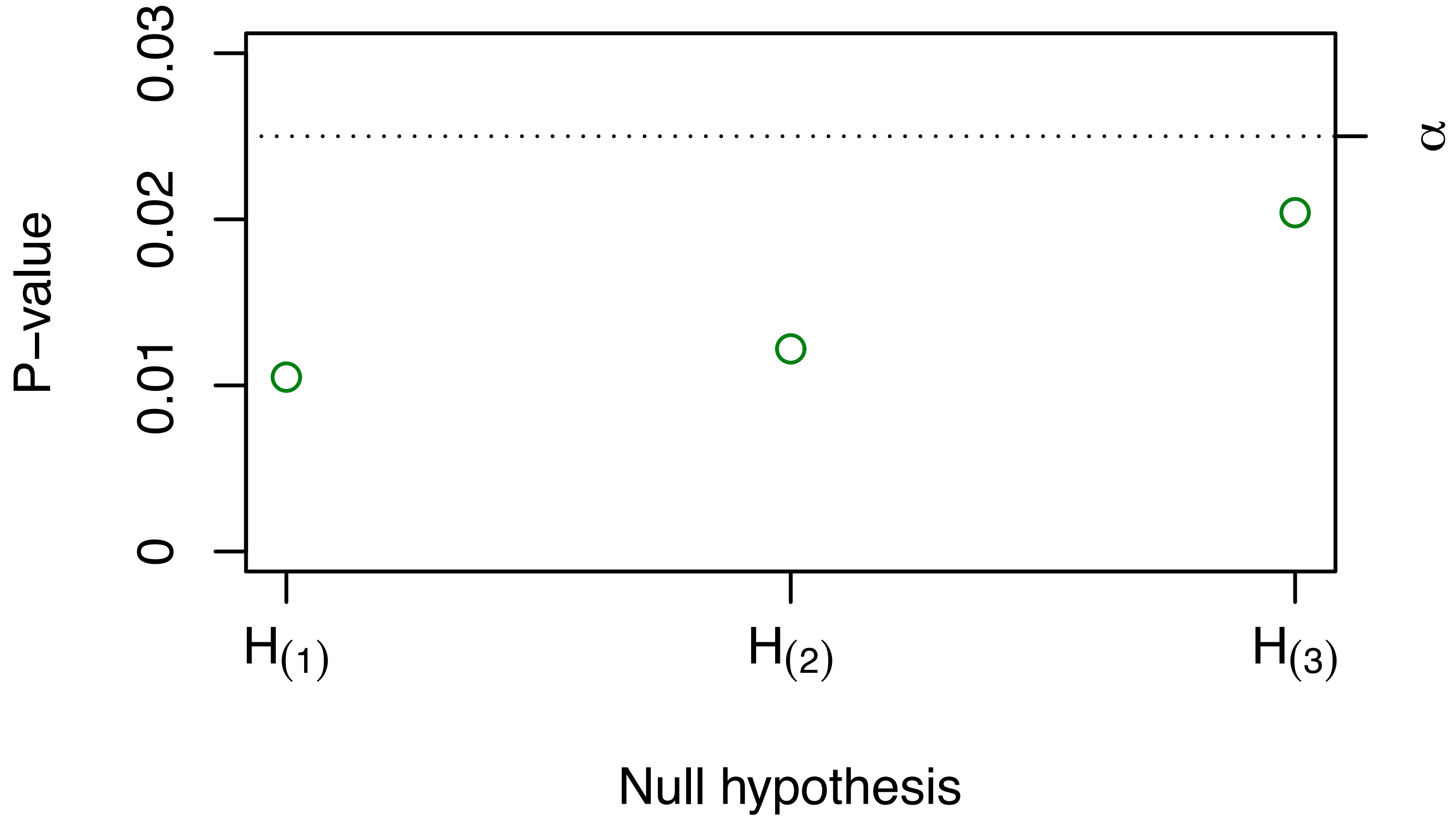
Decision rules in Example 4 ($\alpha = 0.025$)



Step 1: $H_{(1)}$, $H_{(2)}$ and $H_{(3)}$ are all accepted since $p_{(1)} > \alpha/3$

Hochberg procedure

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



Step 1: $H_{(1)}$, $H_{(2)}$ and $H_{(3)}$ are all rejected since $p_{(3)} < \alpha$

Hommel procedure

Simes-based procedure

Hommel procedure is a closed testing procedure derived from Simes procedure, i.e., each intersection hypothesis is tested using Simes procedure

Hommel procedure is based on a more complicated algorithm than Hochberg procedure

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 $p_{(m-i+j)} > j\alpha/i$ for all $j = 1, \ldots, i$, accept $H_{(m-i+j)}$ 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: Major depressive disorder trial

Scenario 4

Comparison	P-value
Dose H vs Placebo (H_1)	0.0291
Dose M vs Placebo (H_2)	0.0095
Dose L vs Placebo (H_3)	0.0153

No significant effect at Dose H 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

Example 4: Major depressive disorder trial

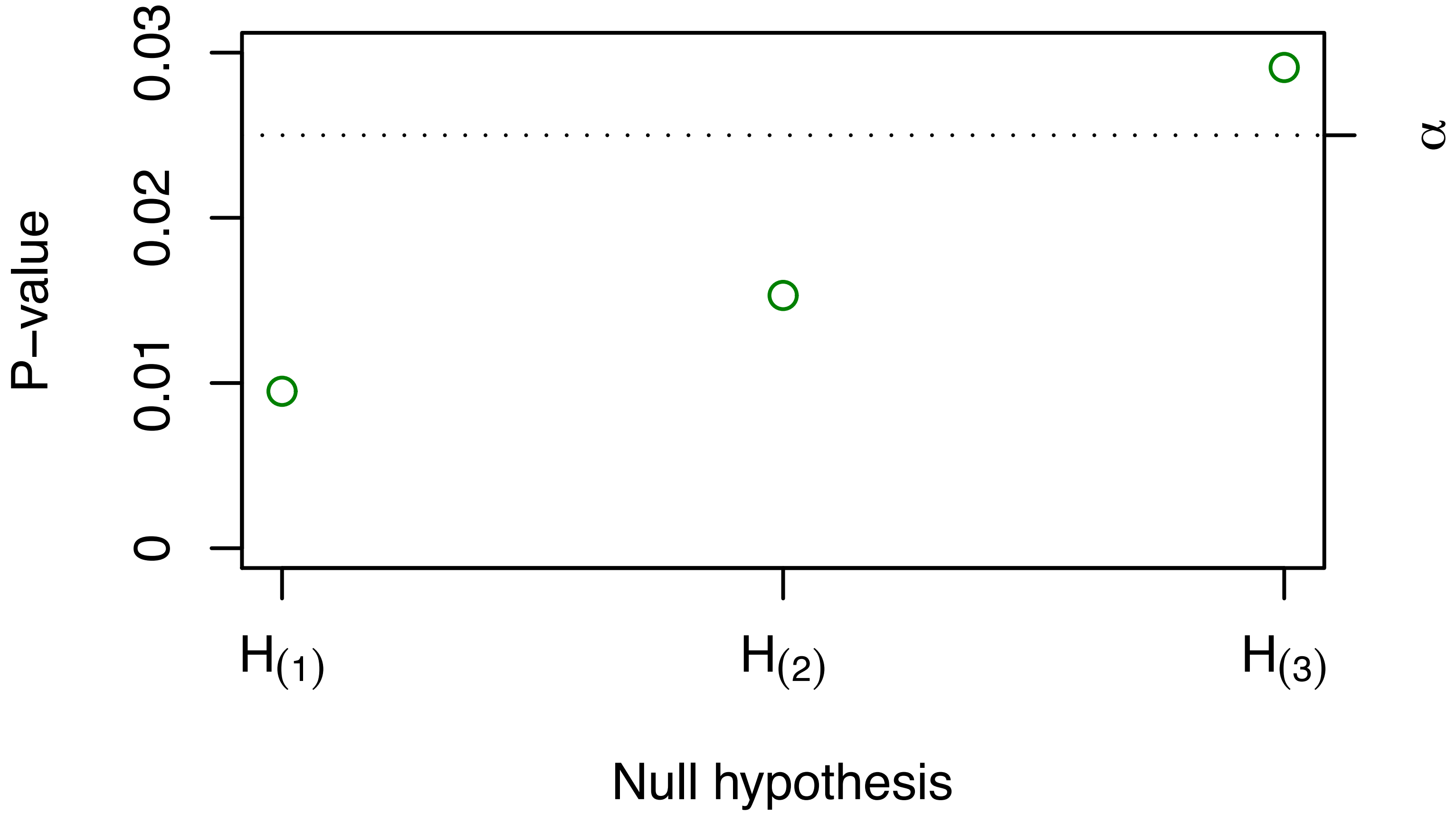
Simplified decision rule for Hommel procedure (step-up algorithm)

Step 1: If $p_{(3)} > \alpha$, accept $H_{(3)}$ and go to Step 2, otherwise reject all hypotheses and stop

Step 2: If $p_{(2)} > \alpha/2$, accept $H_{(2)}$ and go to Step 3, otherwise reject all remaining hypotheses and stop

Step 3: If (1) $\alpha/2 < p_{(2)} \leq 2\alpha/3$ and $p_{(1)} \leq \alpha/2$ or (2) $p_{(1)} \leq \alpha/3$, reject $H_{(1)}$; otherwise accept $H_{(1)}$

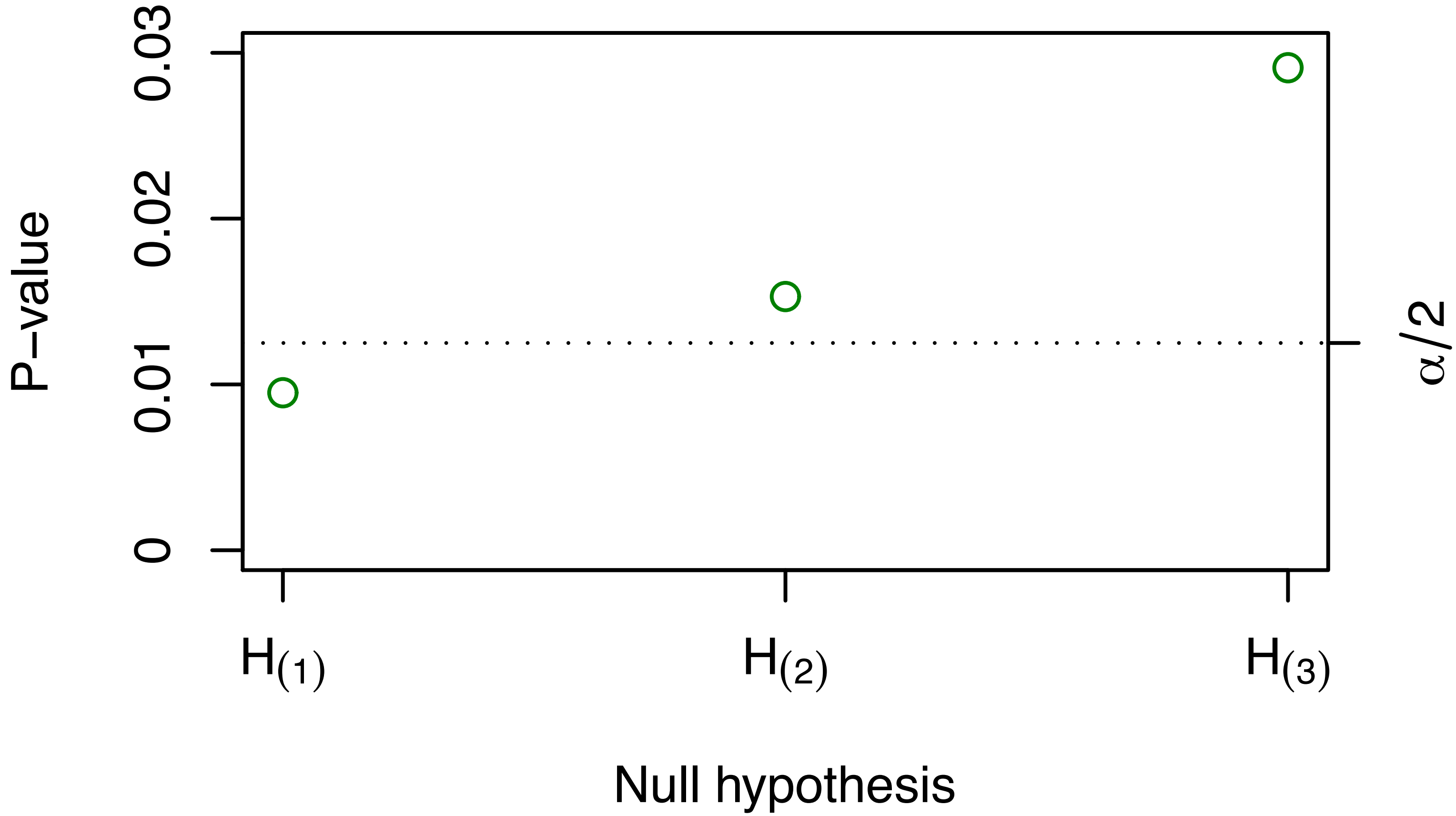
Decision rules in Example 4 ($\alpha = 0.025$)



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

Hochberg procedure

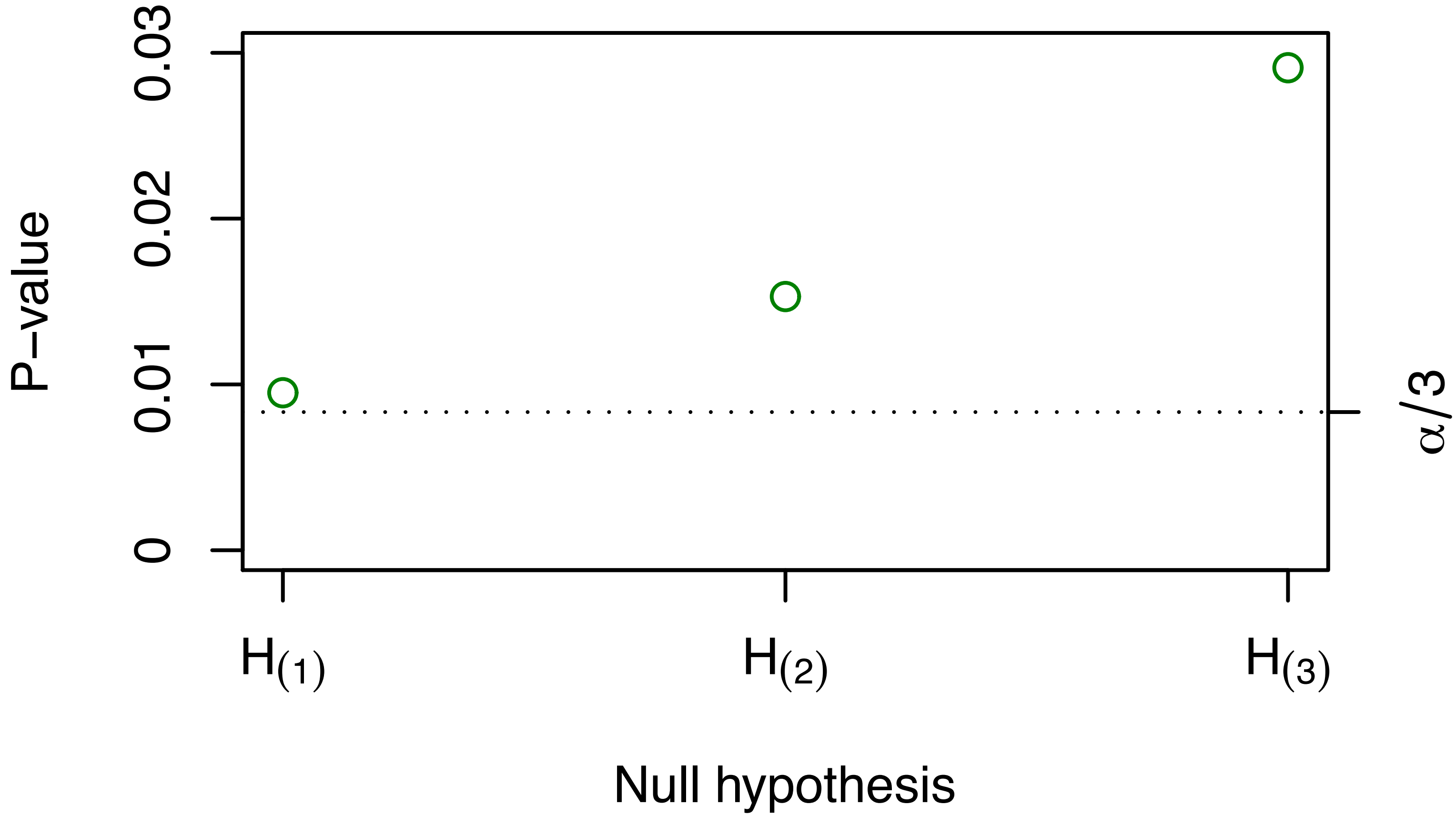
Decision rules in Example 4 ($\alpha = 0.025$)



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

Hochberg procedure

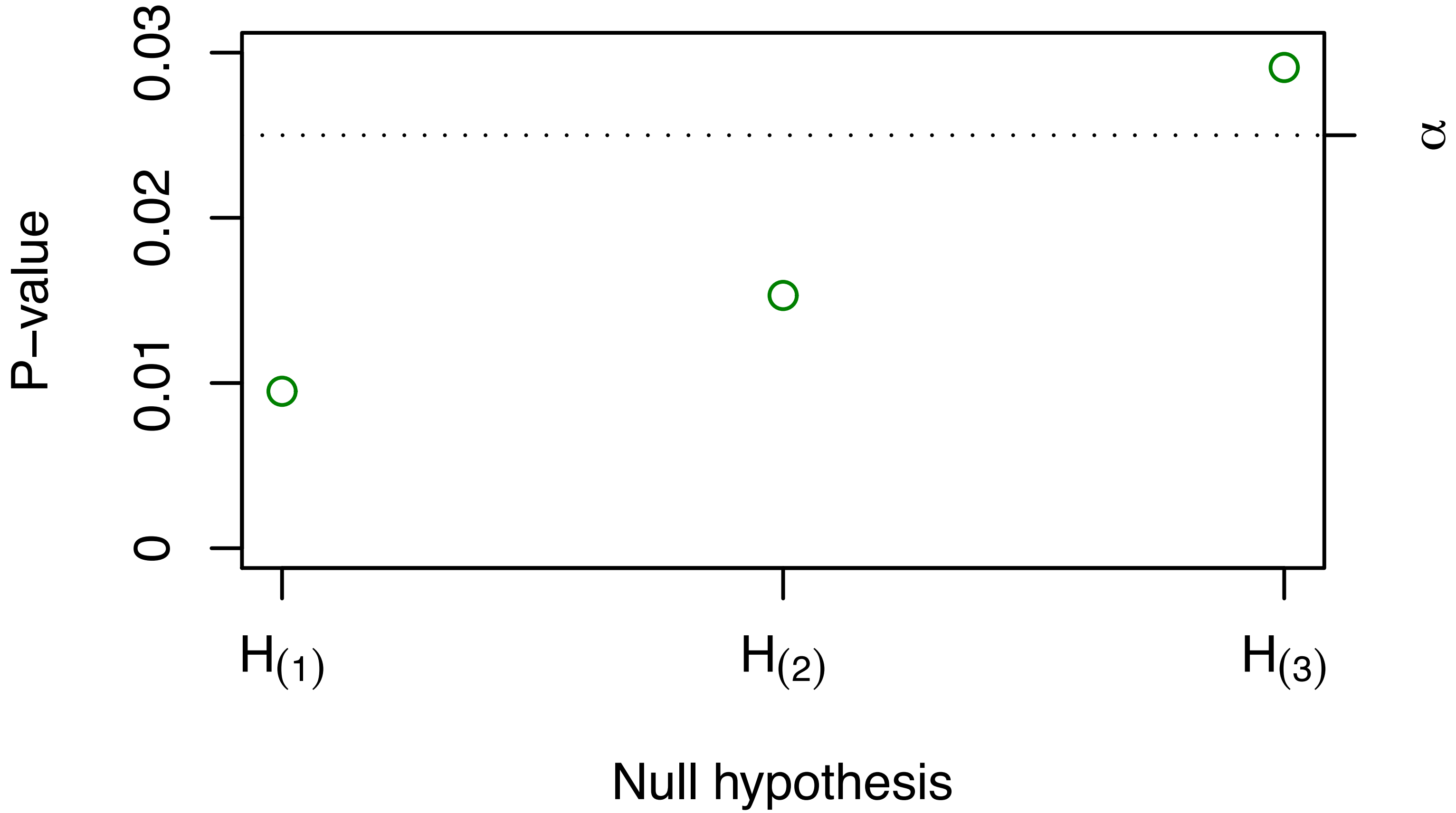
Decision rules in Example 4 ($\alpha = 0.025$)



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

Hommel procedure

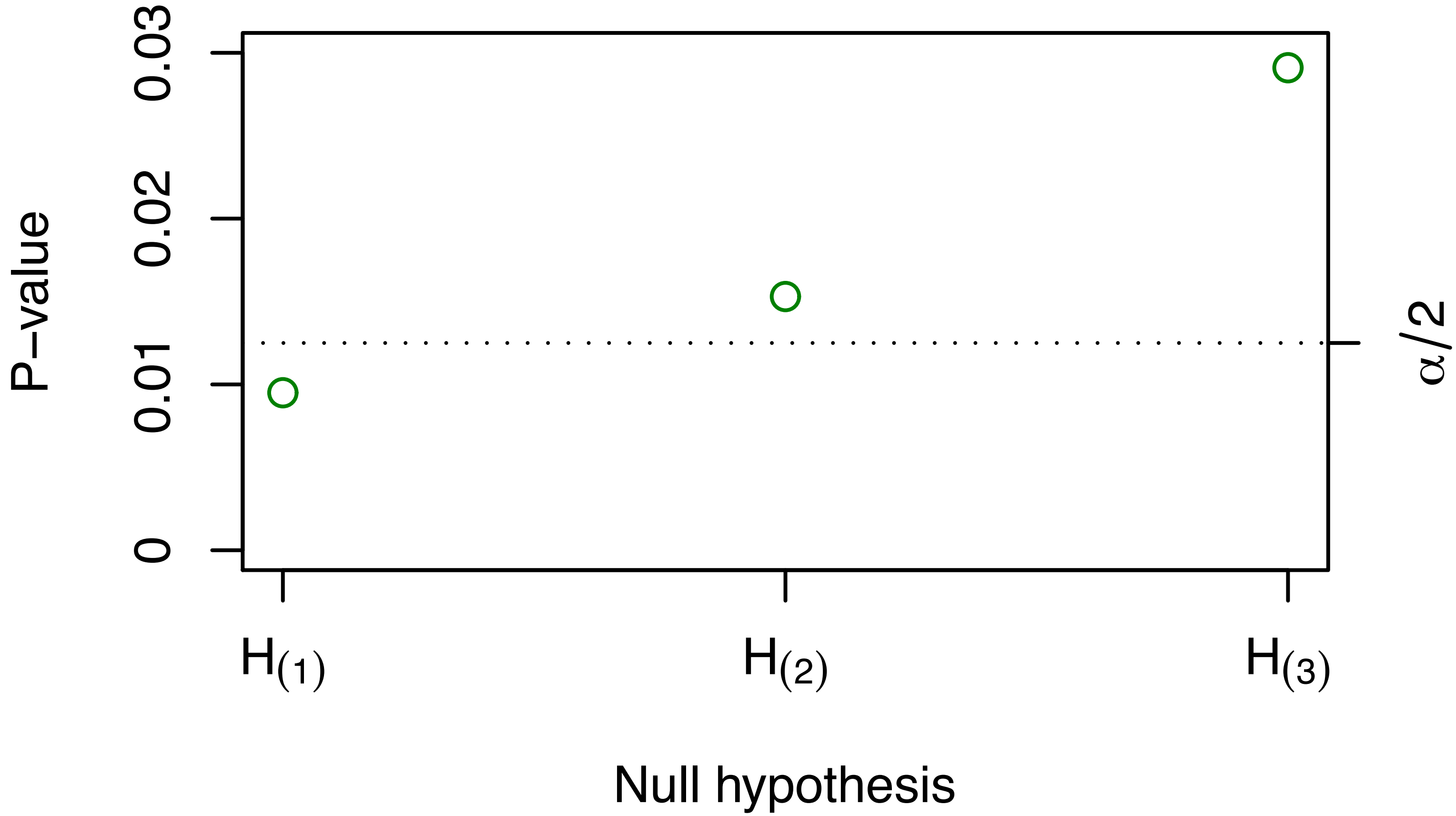
Decision rules in Example 4 ($\alpha = 0.025$)



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

Hommel procedure

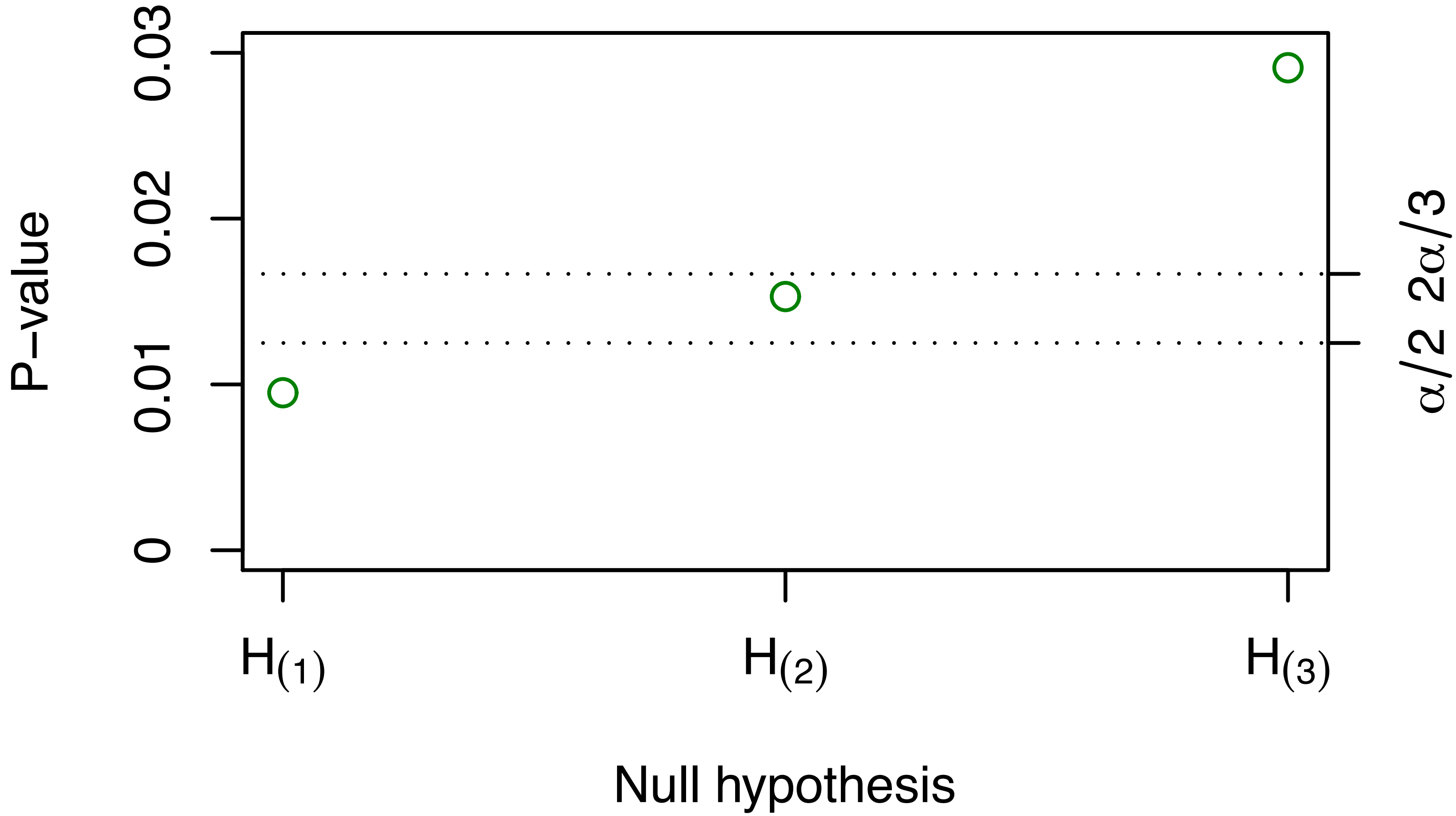
Decision rules in Example 4 ($\alpha = 0.025$)



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

Hommel procedure

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



Step 3: $H_{(1)}$ is rejected since $\alpha/2 < p_{(2)} < 2\alpha/3$ and $p_{(1)} < \alpha/2$ ($p_{(2)}$ is used to reject $H_{(1)}$)

Hommel and Hochberg procedures

Type I error rate

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

Hochberg and Hommel procedures

Type I error rate control

Hypothesis test statistics are independent or positively dependent (when Simes procedure controls Type I error rate)

Positive dependence

Positive dependence condition is satisfied if hypothesis test statistics follow a multivariate normal distribution with positive correlations (Sarkar, 2008)

Positive dependence condition

Example 1: Osteoporosis/breast cancer trial

Condition is satisfied if the two endpoints are positively correlated

Example 4: Major depressive disorder trial

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

Example 5: Schizophrenia trial

Condition is satisfied since the subpopulations are subsets of the overall population

Hommel and Hochberg procedures

Power

Hochberg procedure is uniformly more powerful than Holm procedure

Hommel procedure is uniformly more powerful than Hochberg procedure

Testing algorithms

Hochberg procedure tends to be more popular in clinical trial applications due to simple algorithm

Hommel procedure is recommended due to uniform improvement in power

Hommel and Hochberg procedures

Properties

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

Procedures reward consistency among test outcomes: It is easier to achieve significance if all marginal p-values are small

This property is important in trials with multiple dose-placebo comparisons and other clinical trial applications

Exercise

Cardiovascular clinical trial

A single primary endpoint

Interim analysis

A co-primary endpoint was added

Type I error rate

FDA recommended using a multiple test to control Type I error rate

Exercise

Proposal

Bonferroni test is too conservative

Alternative approaches: Holm, Hommel, and Hochberg tests (power comparison: Holm $<$ Hommel $<$ Hochberg)

Hommel test does not always control Type I error rate whereas Hochberg test does

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

3. Simulation study

Power comparison

Basic procedures

Bonferroni procedure

Procedure with a pre-specified hypothesis ordering

Fixed-sequence procedure

Procedure with a data-driven hypothesis ordering

Hochberg procedure

Power comparison

Clinical trial

Three arms: Two doses of experimental treatment (Dose L and Dose H) versus placebo

Sample size: 230 patients per arm

Endpoint

Single continuous endpoint

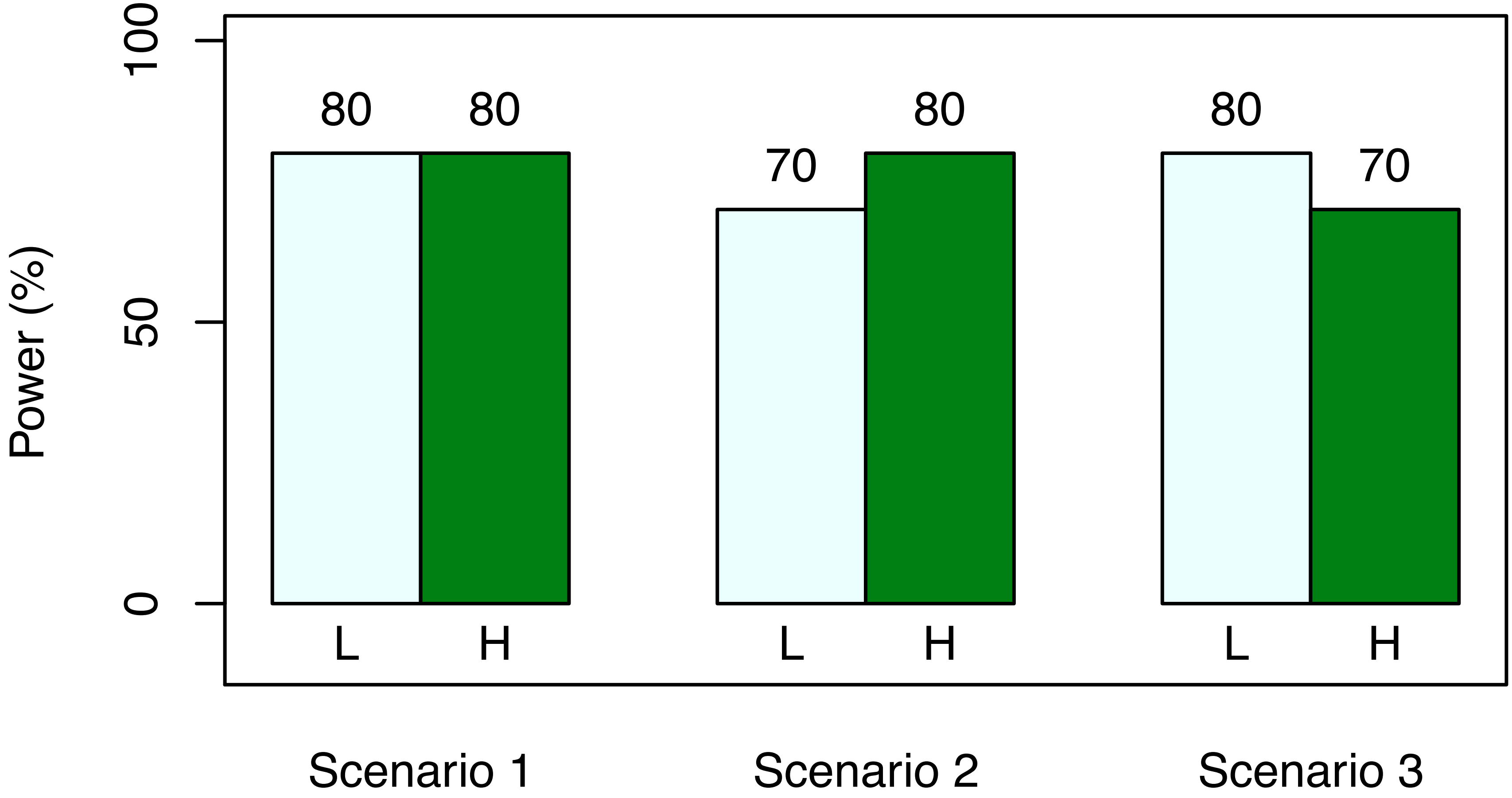
Dose-placebo tests

Test statistics follow a trivariate normal distribution

Common correlation is 0.5

Evaluation scenarios

Marginal power for dose-placebo tests



Scenario 1: Flat dose-response, Scenario 2: Positive dose-response, Scenario 3: Negative dose-response

Power comparison

Multiplicity penalty

Reduction in power after multiplicity adjustment

Example: If marginal power of 80% is reduced to 71% after multiplicity adjustment, multiplicity penalty is 9%

Outcomes of interest

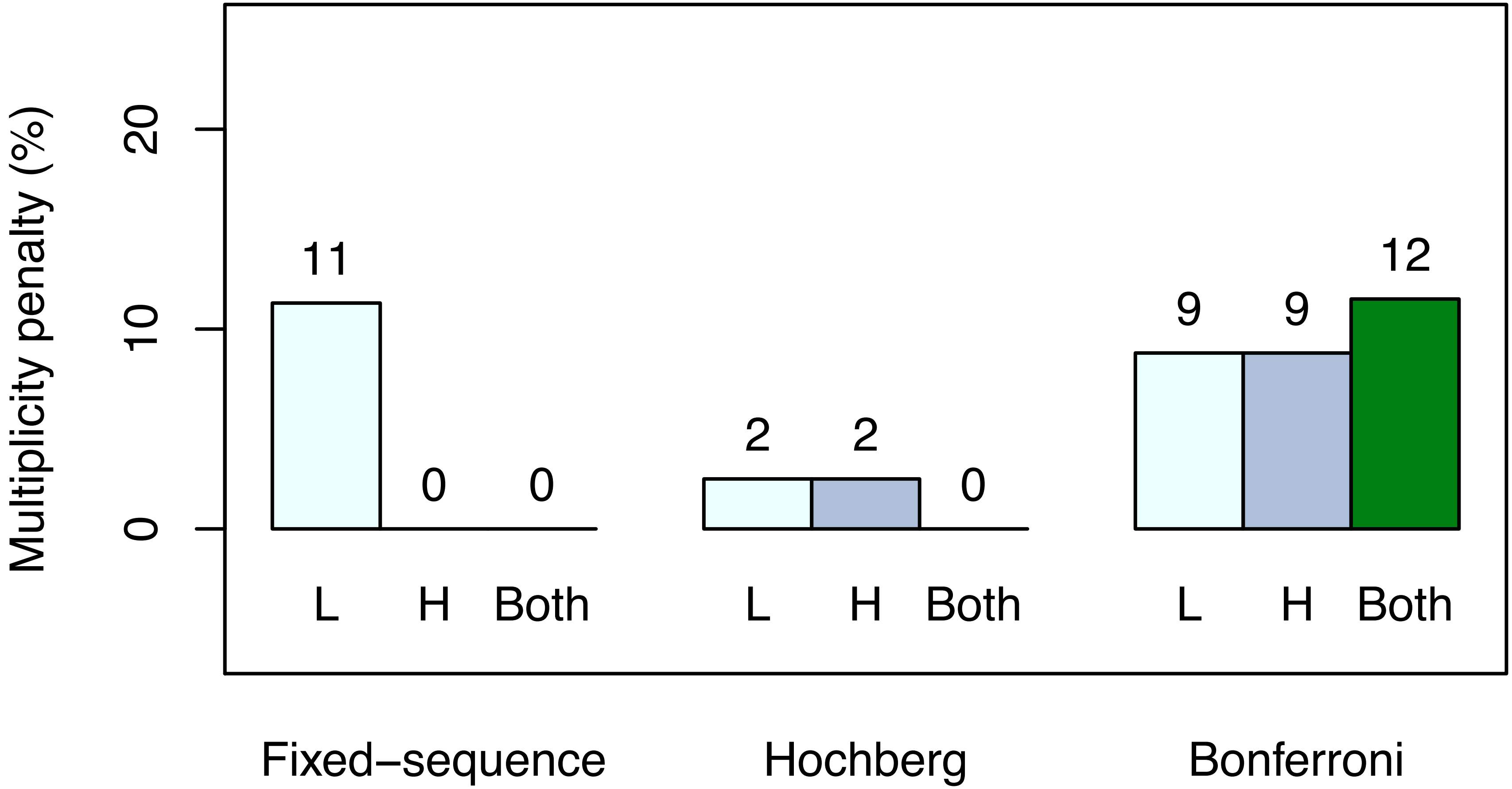
Assess significant effect at Dose L

Assess significant effect at Dose H

Assess significant effect at both doses

Scenario 1

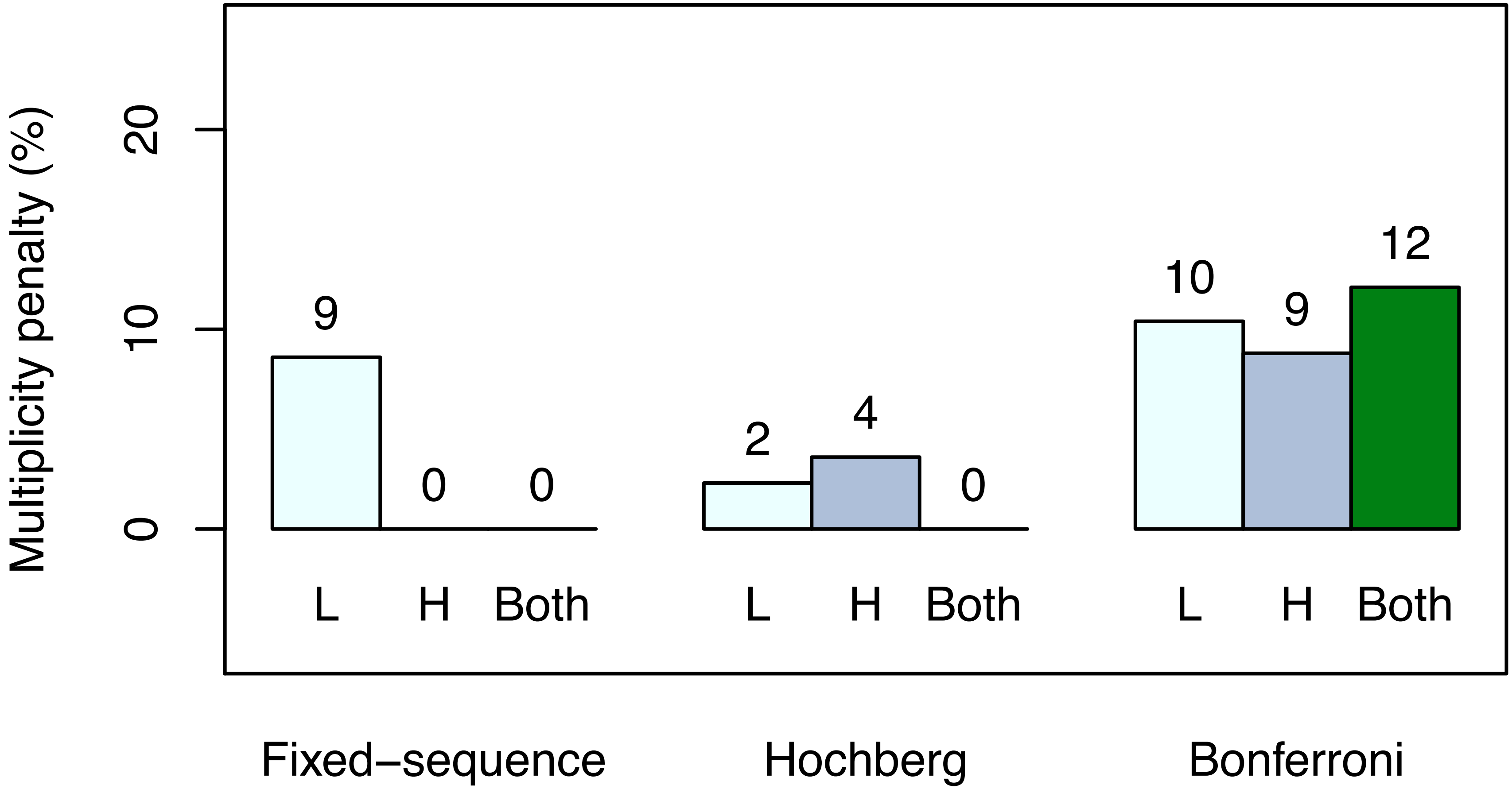
Flat dose-response



L: Significant effect at Dose L, H: Significant effect at Dose H, Both: Significant effect at both doses

Scenario 2

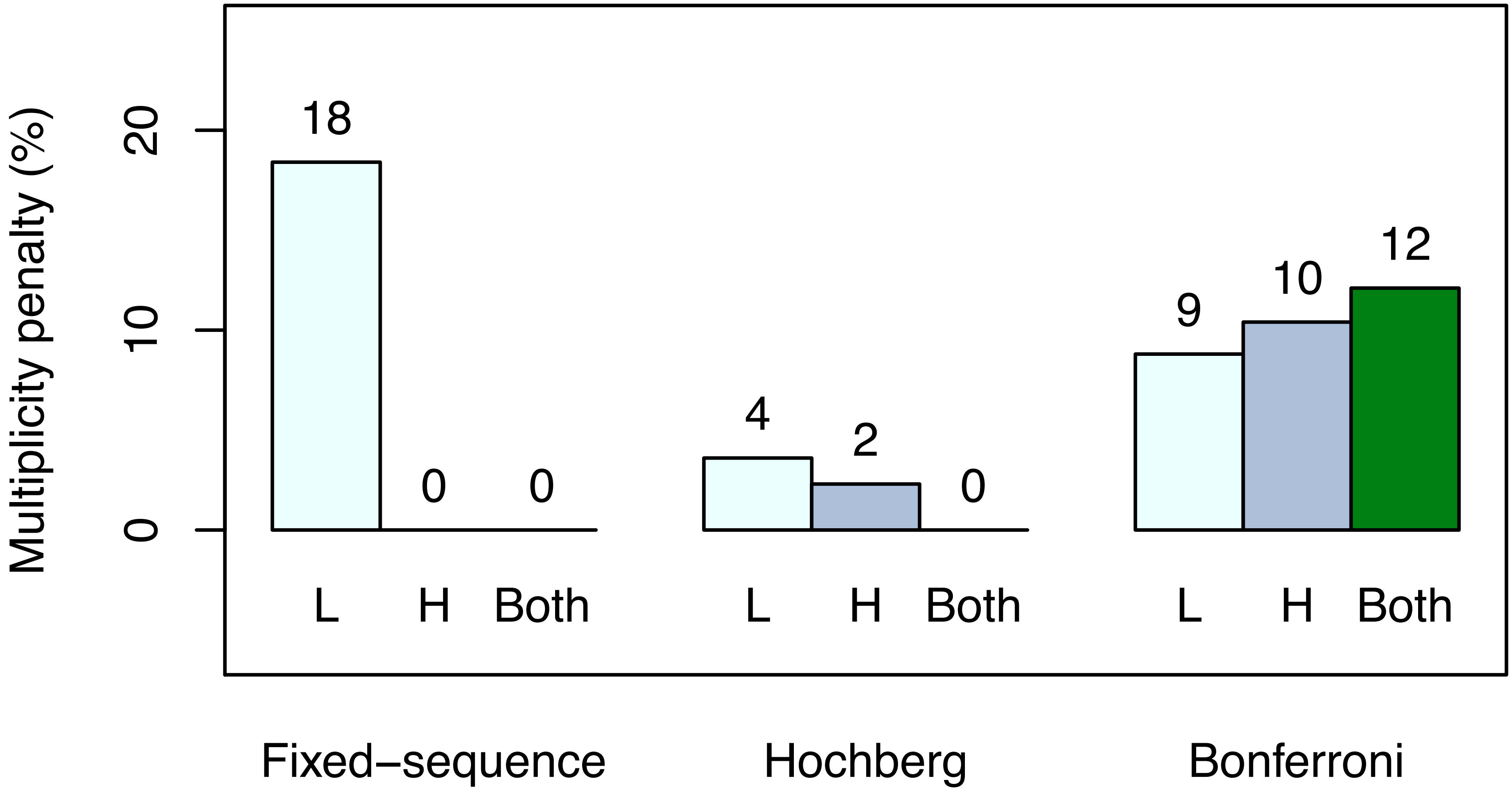
Positive dose-response



L: Significant effect at Dose L, H: Significant effect at Dose H, Both: Significant effect at both doses

Scenario 3

Negative dose-response



L: Significant effect at Dose L, H: Significant effect at Dose H, Both: Significant effect at both doses

Power comparison

Bonferroni procedure

Conservative in all scenarios

Fixed-sequence procedure

Performs well when the first test (Dose H versus placebo) is adequately powered (Scenarios 1 and 2)

Hochberg procedure

Performs well in all scenarios

4. Software implementation

Software implementation in SAS

SAS/STAT module

Selected nonparametric and semiparametric procedures: PROC MULTTEST

Custom macros

Nonparametric, semiparametric and parametric procedures

http://multxpert.com/wiki/Software

Software implementation in SAS

Custom macros

PvalProc macro: Adjusted p-values for popular nonparametric and semiparametric procedures (Bonferroni, Holm, fixed-sequence, fallback, Hommel and Hochberg procedures)

Chain macro: Adjusted p-values for chain procedures

http://multxpert.com/wiki/Software

Software implementation in R

MultComp package

Nonparametric and parametric procedures for linear and related models

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

MultXpert package

Nonparametric, semiparametric and parametric procedures

http://multxpert.com/wiki/MultXpert_package

Summary

Further reading

Multiple Testing Problems in Pharmaceutical Statistics (edited by Alex Dmitrienko, Ajit Tamhane and Frank Bretz)

Sections

1 (Basic procedures) and 2 (Procedures with a data-driven hypothesis ordering): Chapter 2

Summary

Modules B and C

Nonparametric and semiparametric multiple testing procedures

Single-step procedure

Bonferroni procedure

Stepwise procedures with pre-specified hypothesis ordering

Fixed-sequence, fallback and chain procedures

Stepwise procedures with data-driven hypothesis ordering

Holm, Hommel and Hochberg procedures

Summary

Type I error rate

Holm, fixed-sequence, fallback and chain procedures are nonparametric and control FWER for any joint distribution of test statistics

Hommel and Hochberg procedures are semiparametric and control FWER only for some joint distributions, including positively dependent test statistics

Power

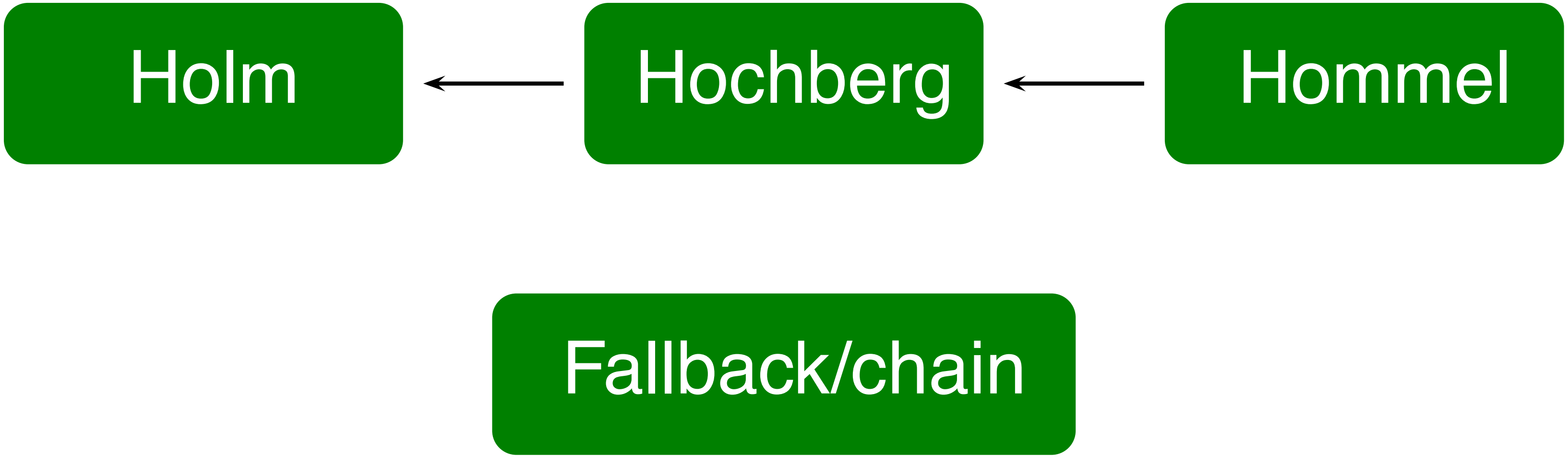
Nonparametric procedures make no assumptions about the joint distribution of test statistics which results in power loss

Summary

Power comparison

Less powerful

More powerful



Holm

Hochberg

Hommel

Fallback/chain

Module D
Parametric Procedures

Outline

1. **Dunnett family**

Dunnett family of parametric procedures

2. **Other parametric procedures**

Parametric fallback, parametric chain and feedback procedures

3. **Software implementation**

Software implementation of parametric procedures in SAS and R

1. Dunnett family

Parametric procedures

Distributional relationships

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

More powerful than nonparametric 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)

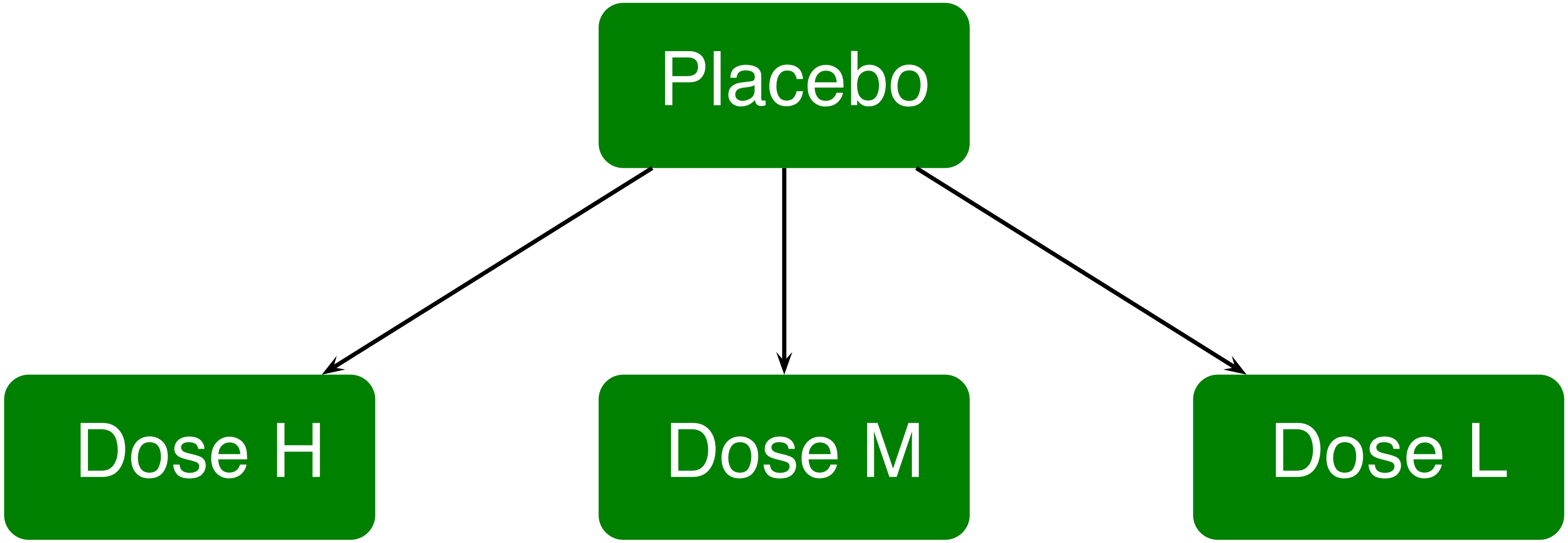
Assumptions

Responses are normally distributed

Balanced design (equal number of patients across treatment groups)

Example 4: Major depressive disorder 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

Nonparametric procedure

Bonferroni procedure

Single-step nonparametric 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

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

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

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

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: Major depressive disorder trial

Scenario 1

Comparison	Mean difference	Test statistic
Dose H vs Placebo (H_1)	2.3	2.30
Dose M vs Placebo (H_2)	2.5	2.50
Dose L vs Placebo (H_3)	1.9	1.90

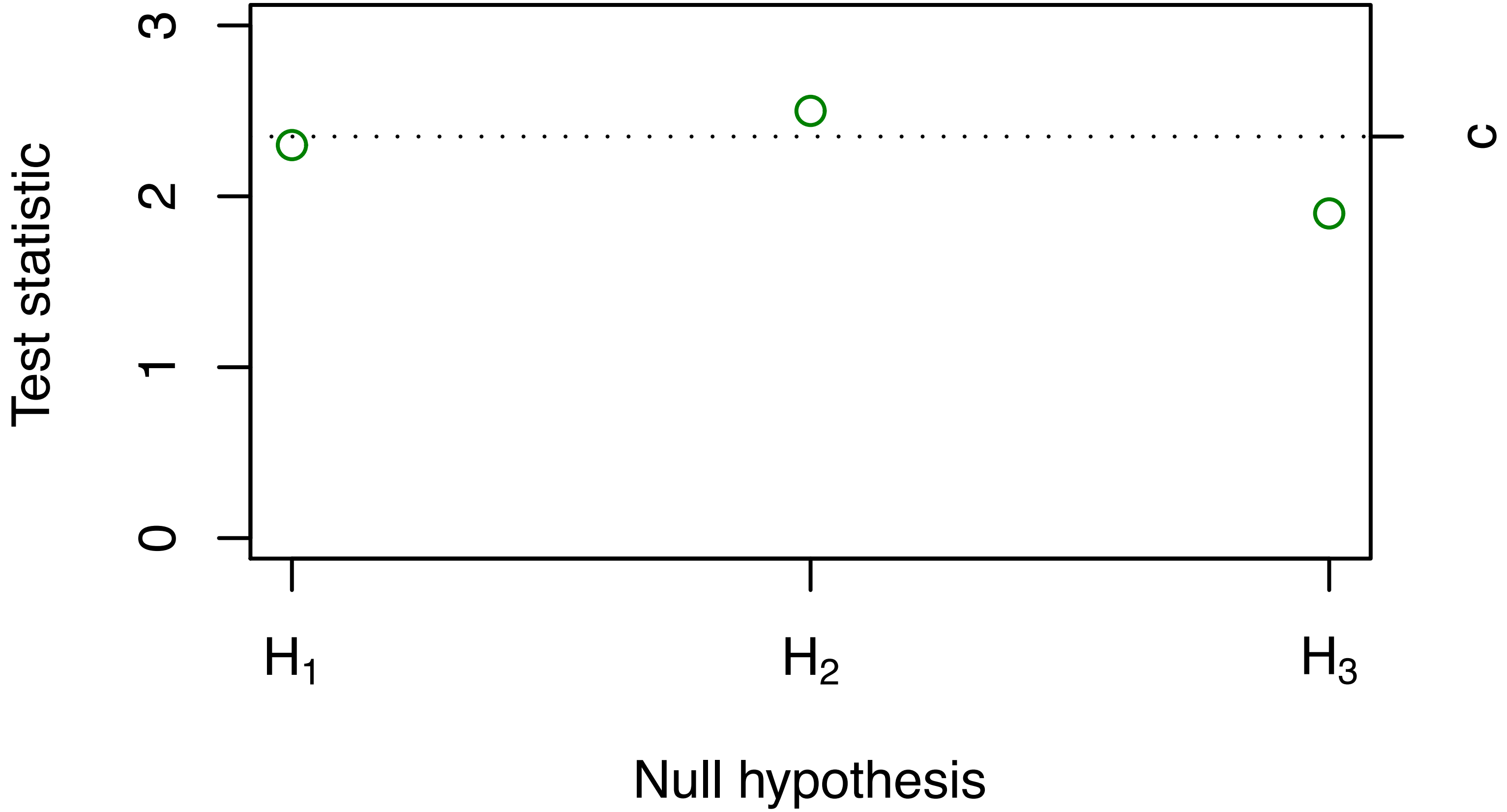
Sample size per group is 180 patients

Pooled standard deviation is 9.5

Evidence of treatment effect at Doses M and H

Single-step Dunnett procedure

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

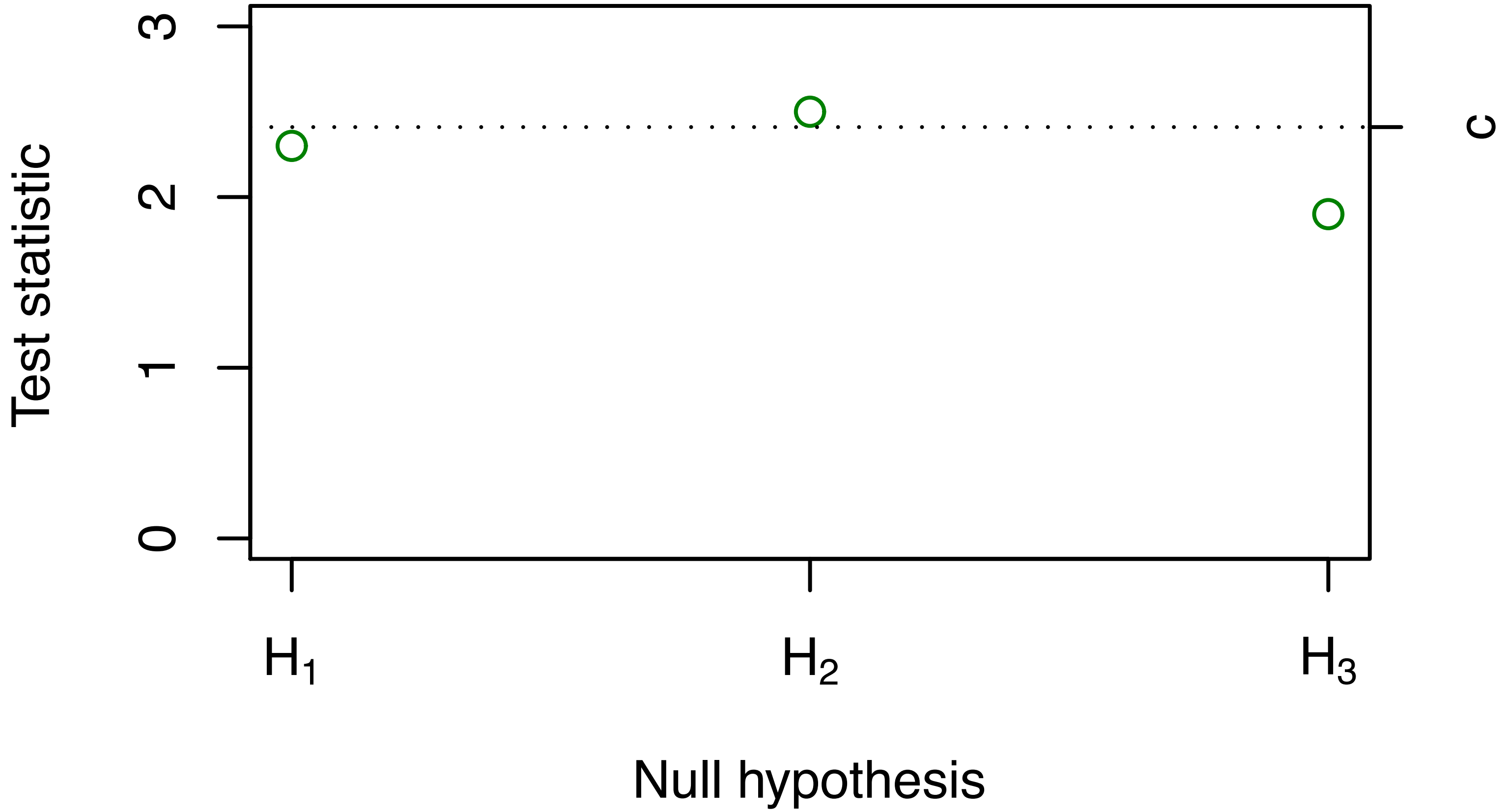


Dunnett critical value is $d_\alpha(m, \nu) = d_{0.025}(3, 716) = 2.35$

Dunnett procedure rejects H_2

Bonferroni procedure

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



Bonferroni critical value is $t_{\alpha/3}(2(n-1)) = t_{0.0083}(358) = 2.41$

Bonferroni procedure rejects H_2

Single-step Dunnett procedure

Type I error rate

Dunnett procedure controls FWER when hypothesis test statistics follow a multivariate t distribution, e.g., problems 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

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 rules (step-down algorithm)

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: Major depressive disorder trial

Scenario 1

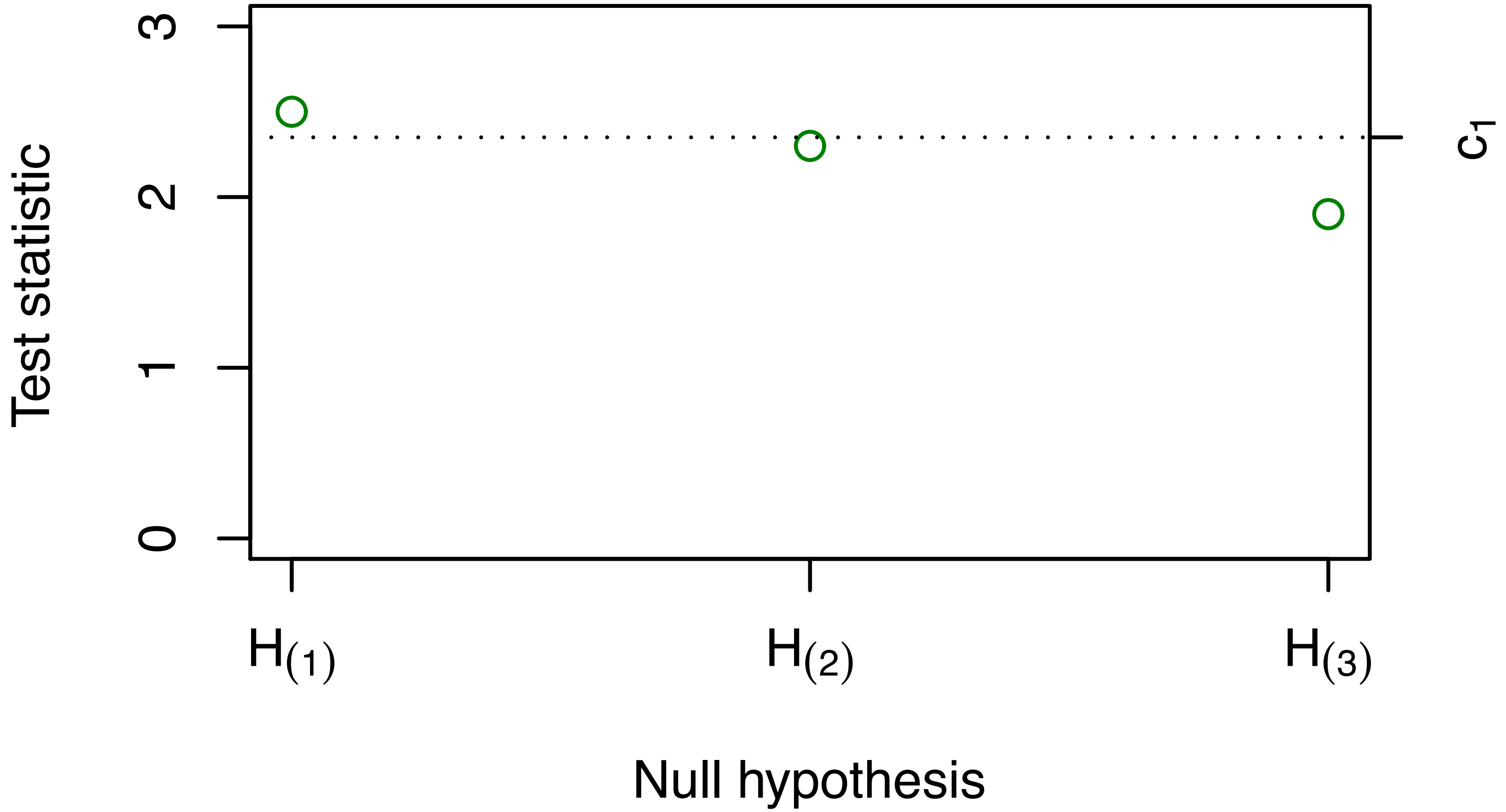
Comparison	Mean difference	Test statistic
Dose H vs Placebo (H_1)	2.3	2.30
Dose M vs Placebo (H_2)	2.5	2.50
Dose L vs Placebo (H_3)	1.9	1.90

$t_{(1)} = t_2 = 2.50$, $t_{(2)} = t_1 = 2.30$, $t_{(3)} = t_3 = 1.90$, ordered test statistics

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

Step-down Dunnett procedure

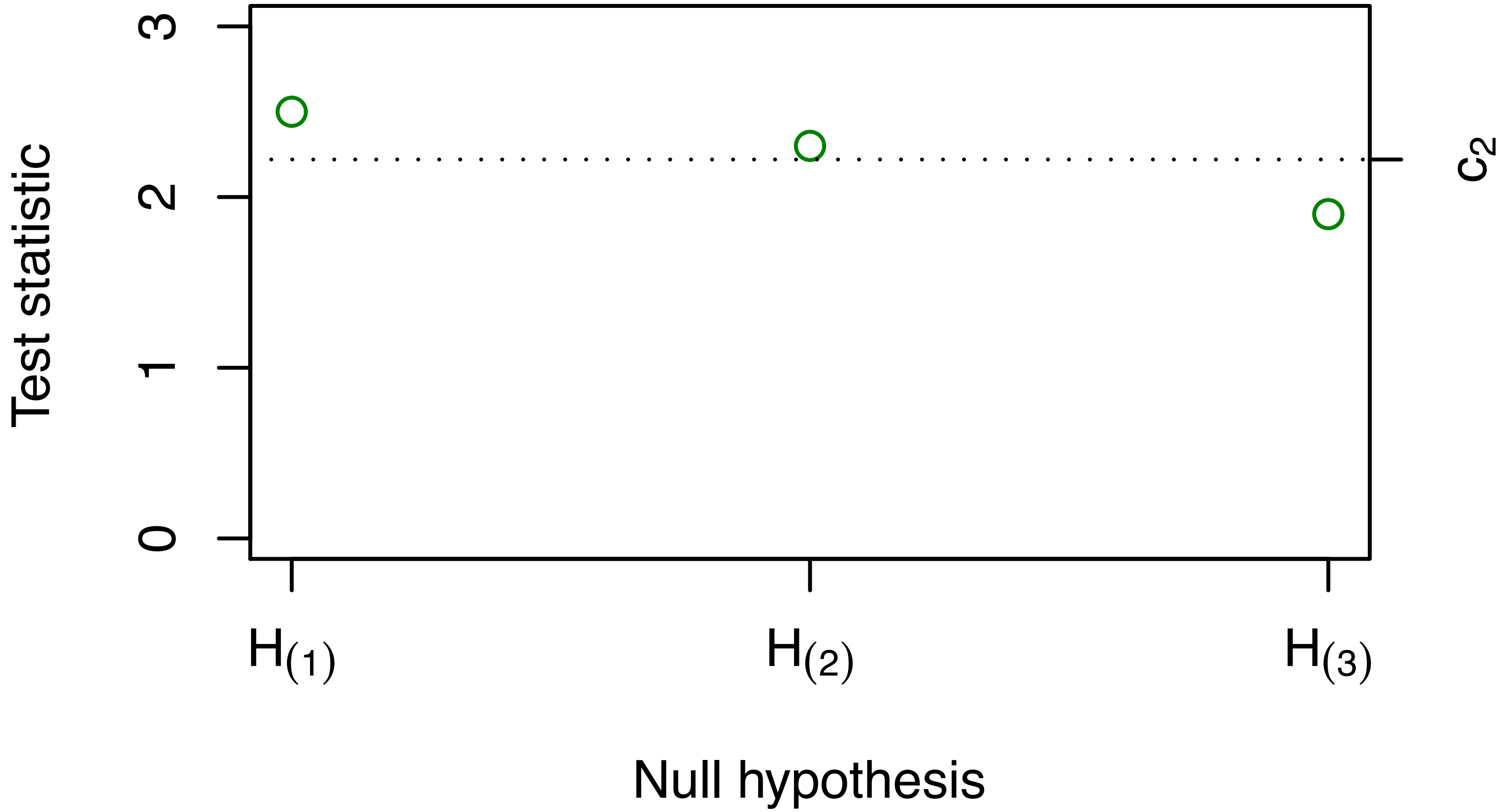
Decision rules in Example 4 ($\alpha = 0.025$)



Step 1: $H_{(1)}$ is rejected since $t_{(1)} > c_1 = d_\alpha(m, \nu)$ $= d_{0.025}(3, 716) = 2.35$

Step-down Dunnett procedure

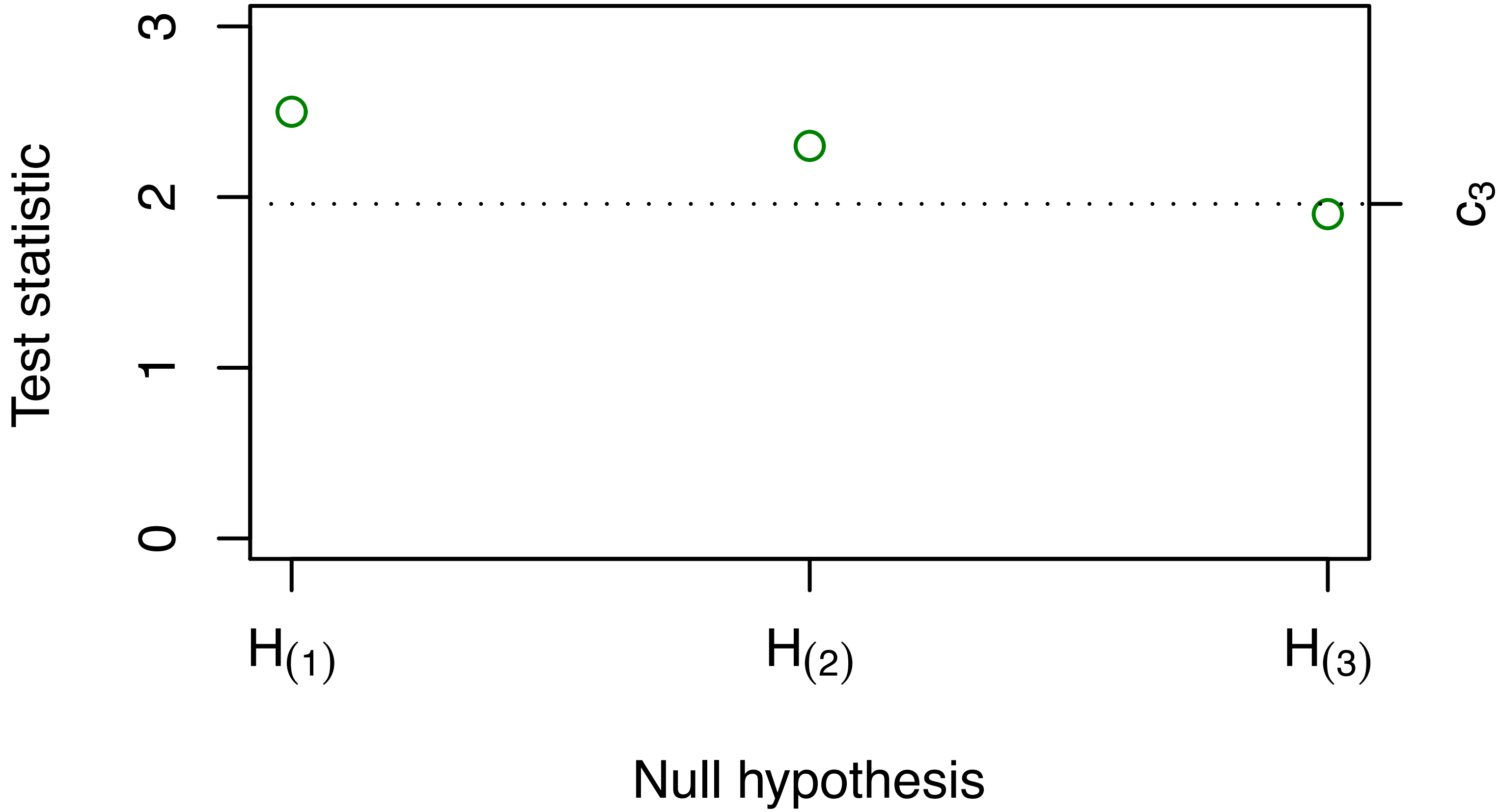
Decision rules in Example 4 ($\alpha = 0.025$)



Step 2: $H_{(2)}$ is rejected since $t_{(2)} > c_2 = d_\alpha(m-1,\nu)$ $= d_{0.025}(2,716) = 2.22$

Step-down Dunnett procedure

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



Step 3: $H_{(3)}$ is accepted since $t_{(3)} < c_3 = d_\alpha(m-2,\nu)$ $= d_{0.025}(1,716) = 1.96$

Step-down Dunnett procedure

Type I error rate

Step-down Dunnett procedure 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

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 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

2. Other parametric procedures

Other parametric procedures

Parametric fallback procedures

Extension of fallback procedures (Huque and Alosh, 2008)

Parametric chain procedures

Extension of chain procedures (Millen and Dmitrienko, 2011)

Feedback procedures

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

3. Software implementation

Software implementation in SAS and R

Custom SAS macros

ParProc macro: Adjusted p-values for popular parametric procedures (Dunnett and step-down Dunnett procedures)

http://multxpert.com/wiki/Software

MultXpert R package

ParAdjP function: Adjusted p-values for popular parametric procedures (Dunnett and step-down Dunnett procedures)

http://multxpert.com/wiki/MultXpert_package

Summary

Summary

Power

Parametric procedures make explicit assumptions about the joint distribution of hypothesis test statistics and are more powerful than nonparametric procedures

Type I error rate

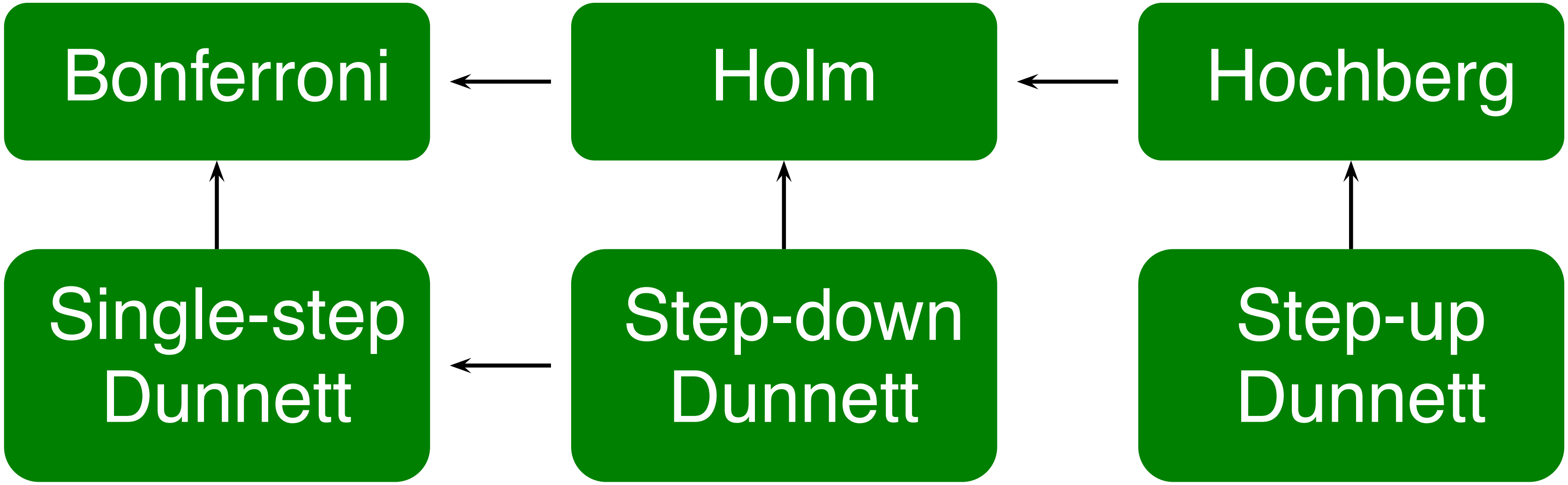
Single-step Dunnett, step-down Dunnett, step-up Dunnett procedures control FWER when hypothesis test statistics follow a multivariate t distribution

Summary

Power comparison

Less powerful

More powerful



Bonferroni

Holm

Hochberg

Single-step Dunnett

Step-down Dunnett

Step-up Dunnett

More powerful

Module D

Further reading

Multiple Testing Problems in Pharmaceutical Statistics (edited by Alex Dmitrienko, Ajit Tamhane and Frank Bretz)

Sections

1 (Dunnett family): Chapter 2

Module E
Simultaneous Confidence Intervals

Outline

1. Simultaneous confidence intervals

Simultaneous confidence intervals for single-step and stepwise procedures

2. Software implementation

Computation of simultaneous confidence intervals in SAS and R

1. Simultaneous confidence intervals

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

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

Stepwise procedures

Nonparametric stepwise procedures

Simultaneous confidence intervals are available for fixed-sequence, fallback and Holm procedures (Strassburger and Bretz, 2008; Guilbaud, 2008; Guilbaud, 2009)

Other procedures

Simultaneous confidence intervals can also be constructed for semiparametric (Hommel and Hochberg) and parametric procedures (step-down Dunnett) (Stefansson, Kim and Hsu, 1988; Guilbaud and Karlsson, 2011; Guilbaud, 2012)

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, Familywise error rate

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

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

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

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

Consistency

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

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 all null hypotheses
- Define simultaneous confidence intervals

Nonparametric procedures

Bonferroni procedure

Lower confidence 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 non-informative (set to 0)

Example 4: Major depressive disorder trial

Scenario 1

Comparison	Mean difference	P-value
Dose H vs Placebo (H_1)	2.3	0.0111
Dose M vs Placebo (H_2)	2.5	0.0065
Dose L vs Placebo (H_3)	1.9	0.0293

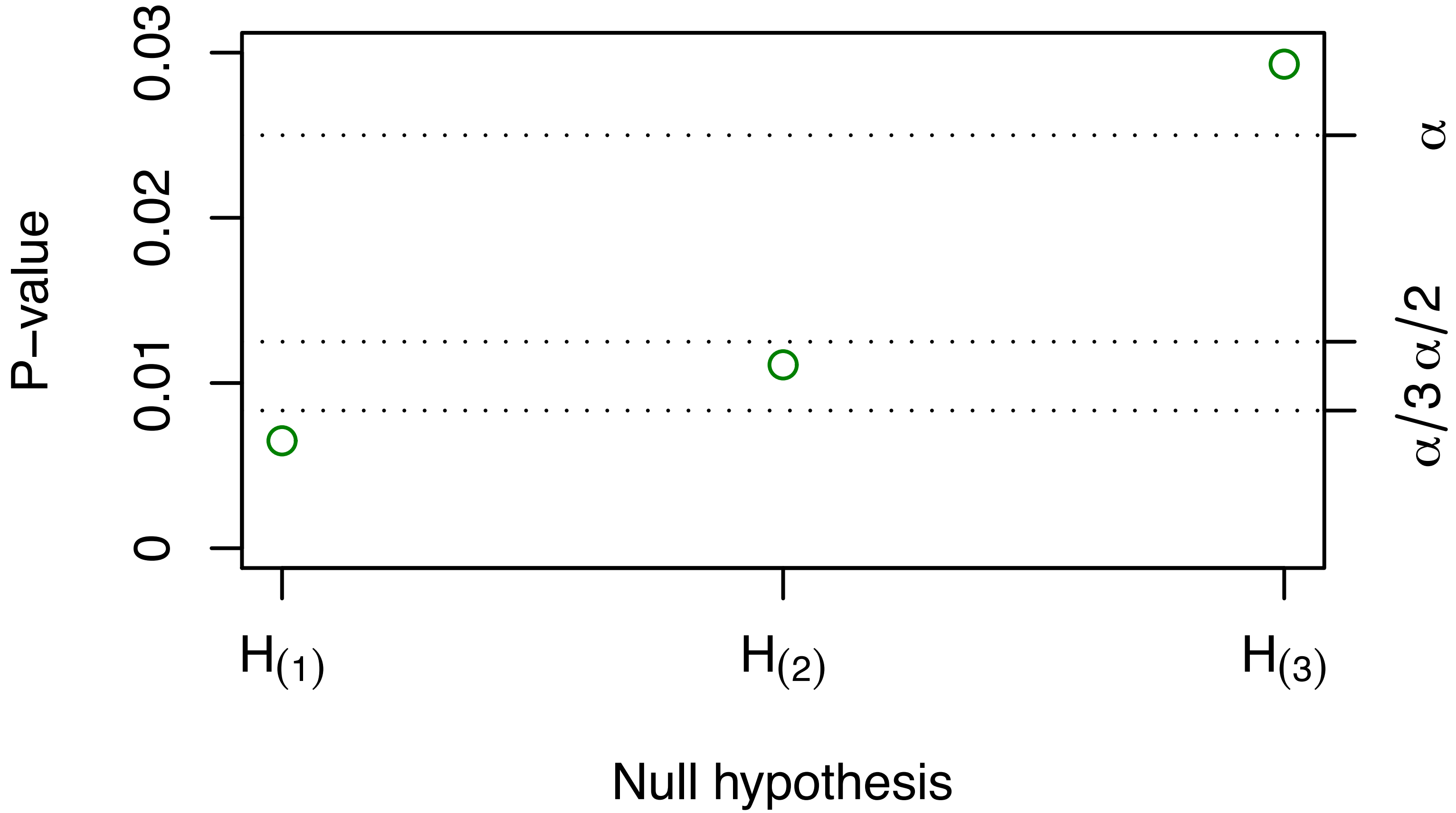
Sample size per group is 180 patients

Pooled standard deviation is 9.5

Evidence of treatment effect at Doses M and H

Bonferroni and Holm procedures

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

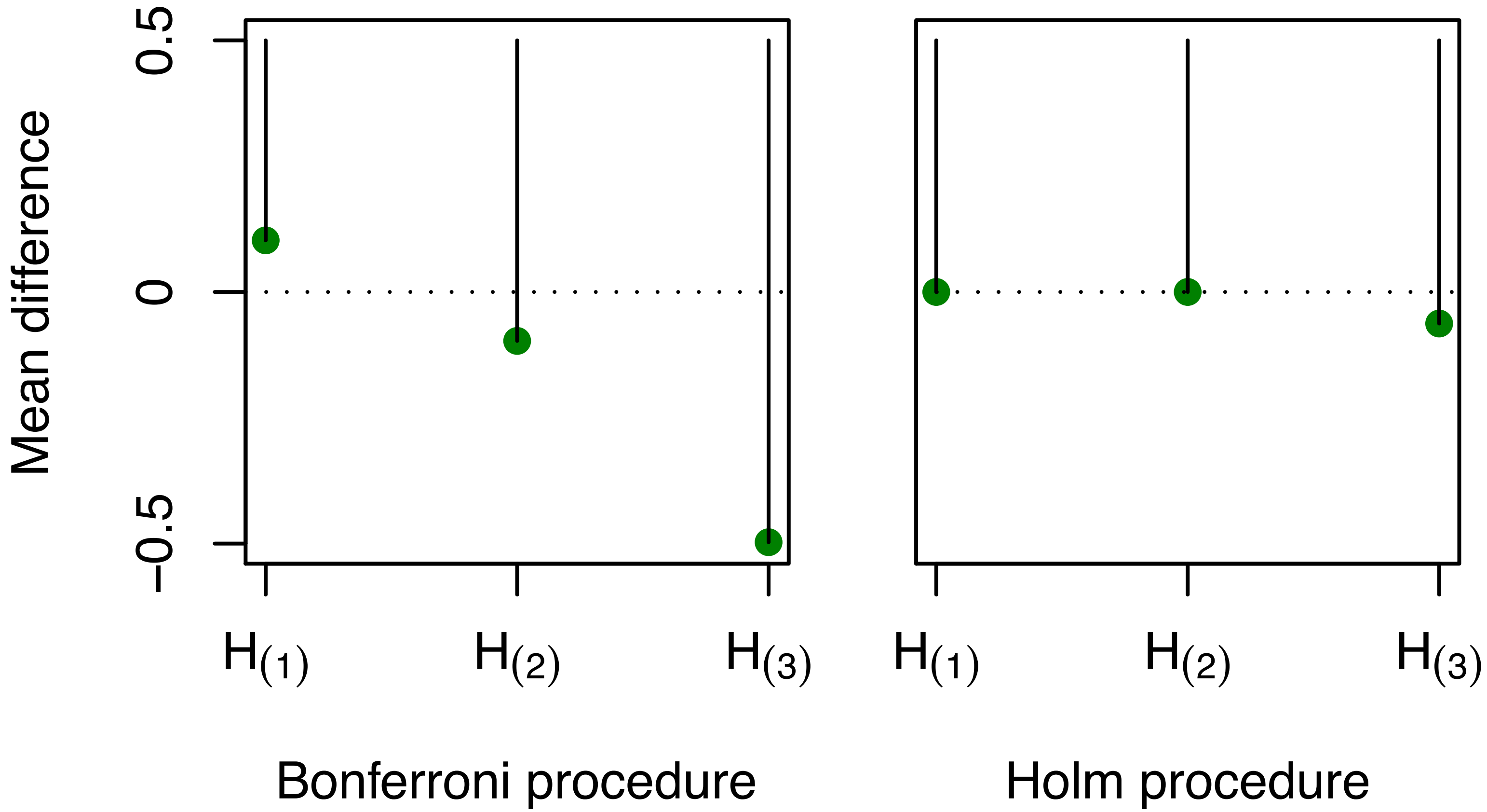


Bonferroni procedure rejects $H_{(1)} = H_2$

Holm procedure rejects $H_{(1)} = H_2$ and $H_{(2)} = H_1$

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

Bonferroni procedure

$0.10 < \theta_{(1)} < \infty$

$-0.10 < \theta_{(2)} < \infty$

$-0.50 < \theta_{(3)} < \infty$

Holm procedure

$0 < \theta_{(1)} < \infty$ (non-informative since $H_{(1)}$ is known to be rejected)

$0 < \theta_{(2)} < \infty$ (non-informative since $H_{(2)}$ is known to be rejected)

$-0.06 < \theta_{(3)} < \infty$

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

Simultaneous confidence intervals

Single-step Dunnett procedure

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

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

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 non-informative (set to 0)

Example 4: Major depressive disorder trial

Scenario 1

Comparison	Mean difference	Test statistic
Dose H vs Placebo (H_1)	2.3	2.30
Dose M vs Placebo (H_2)	2.5	2.50
Dose L vs Placebo (H_3)	1.9	1.90

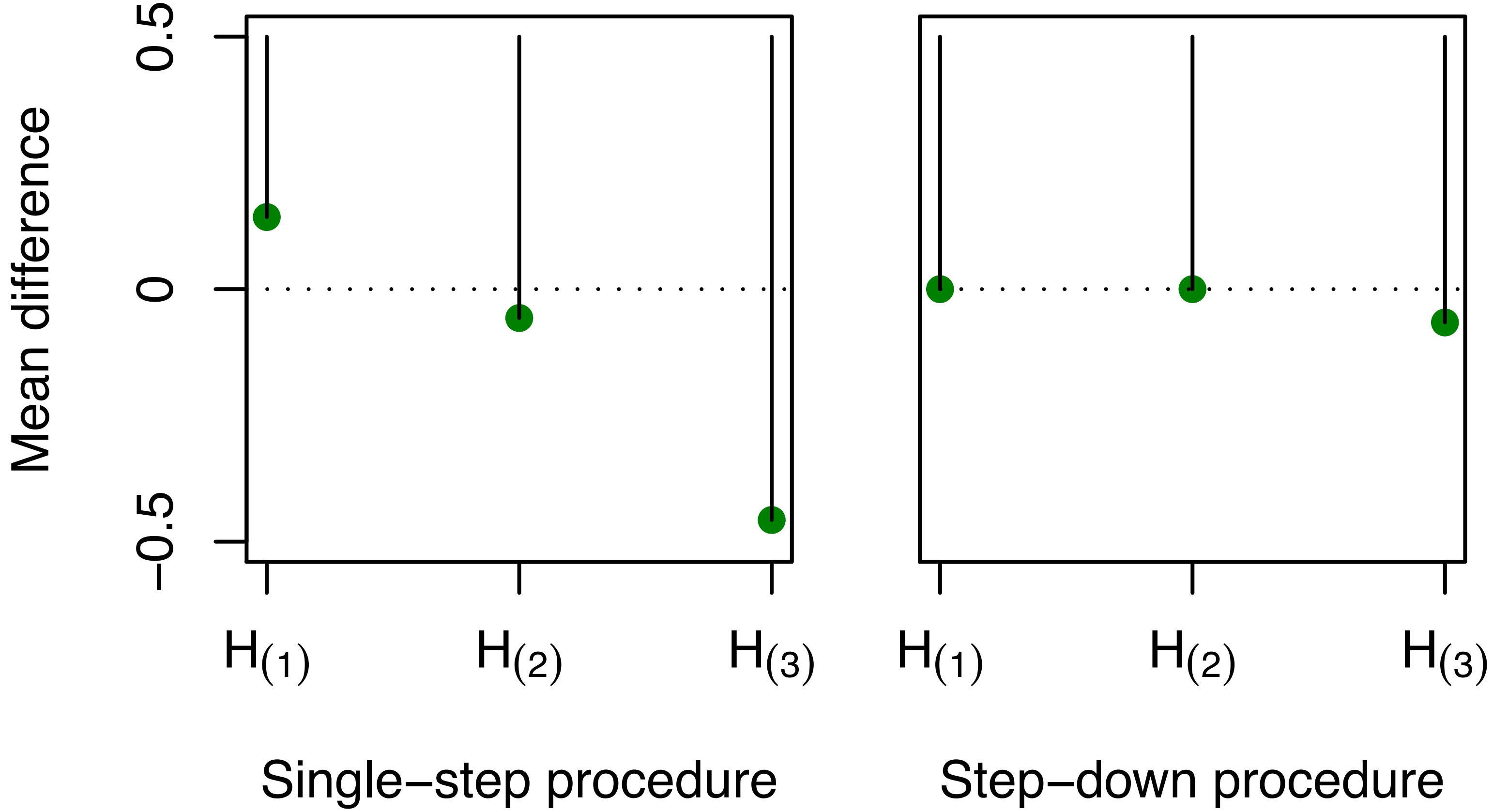
Sample size per group is 180 patients

Pooled standard deviation is 9.5

Evidence of treatment effect at Doses M and H

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

2. Software implementation

Software implementation in SAS

Custom macros

PvalCI macro: Simultaneous confidence intervals for commonly used nonparametric procedures (Bonferroni, fixed-sequence, fallback and Holm procedures)

ParCI macro: Simultaneous confidence intervals for parametric procedures (Dunnett and step-down Dunnett procedures)

http://multxpert.com/wiki/Software

Software implementation in R

MultXpert package

PvalCI function: Simultaneous confidence intervals for commonly used nonparametric procedures (Bonferroni, fixed-sequence, fallback and Holm procedures)

ParCI function: Simultaneous confidence intervals for parametric procedures (Dunnett and step-down Dunnett procedures)

http://multxpert.com/wiki/MultXpert_package

Summary

Simultaneous confidence intervals

Single-step procedures

Simultaneous confidence intervals are easily computed and provide information to support risk-benefit assessments

Stepwise procedures

Simultaneous confidence intervals for more powerful procedures are available but tend to be non-informative

Lower confidence limit is typically set to 0 when a hypothesis is rejected

Simultaneous confidence intervals

Alternative approach 1

A powerful procedure is used for inferential purposes, e.g., Holm-adjusted p-values are computed

Informative simultaneous confidence intervals are constructed using a basic procedure, e.g., Bonferroni procedure is applied

Results are not consistent

Simultaneous confidence intervals

Alternative approach 2

A powerful procedure is modified to produce more informative simultaneous confidence intervals, e.g., modified Holm, Hochberg or Hommel procedures can be applied (Guilbaud, 2012)

Modified procedures are less powerful than original procedures which lowers success probability

Simultaneous confidence intervals

Alternative approach 3

A powerful procedure is used for inferential purposes

Univariate confidence intervals are constructed

Results are not consistent but this approach is most commonly used

Module E

Further reading

Multiple Testing Problems in Pharmaceutical Statistics (edited by Alex Dmitrienko, Ajit Tamhane and Frank Bretz)

Sections

1 (Simultaneous confidence intervals): Chapter 2

Module F
Sample Size Calculations

Outline

1. Sample size calculations in trials with multiple objectives

Required parameters and success criteria

2. Case study

Sample size calculations in a trial with multiple dose-placebo comparisons

1. Sample size calculations in trials with multiple objectives

Clinical trial with a single objective

Clinical trial

Two arms: Experimental treatment versus placebo

Balanced design (1:1 randomization)

Endpoint

Single normally distributed endpoint

Clinical trial with a single objective

Parameters for sample size calculation

δ, Mean treatment difference

σ, Common standard deviation

α, One-sided Type I error rate

β, Type II error rate

Clinical trial with a single objective

Power function

$$\psi(n) = \Phi\left(\sqrt{\frac{n}{2}}\frac{\delta}{\sigma} - z_\alpha\right)$$

n, Sample size per treatment arm

z_α, Upper quantile of standard normal distribution

Sample size calculation

Sample size is found from $\psi(n) = 1 - \beta$

Closed-form solutions are often available

Clinical trial with multiple objectives

Clinical trial

Two arms: Experimental treatment versus placebo

Balanced design (1:1 randomization)

Objectives

At-least-one testing approach is applied

Multiple objectives such as multiple endpoints, multiple dose-placebo comparisons, multiple subgroups

Test statistics follow a multivariate normal distribution

Clinical trial with multiple objectives

Parameters for sample size calculation

One-sided familywise error rate

Type II error rate

Mean treatment differences

Common standard deviation

Correlations among test statistics

Success criterion (overall power)

Success criteria

Overall power function

Multiple ways to extend univariate power function to define success criterion (overall power function) for multiple objectives

Different approaches are defined in Bauer (1987), Millen and Dmitrienko (2011), Bretz et al. (2011), Dmitrienko et al. (2011)

Commonly used criteria

Disjunctive power (simple, generalized, subset)

Weighted power

Success criteria

Notation

$H_1, \dots, H_m$, Null hypotheses

$r_1, \dots, r_m$, Rejection indicators ($r_i = 1$ if H_i is rejected and $r_i = 0$ otherwise)

$\psi_i(n)$, Probability to reject H_i as a function of the sample size n

$\psi(n)$, Criterion function (overall power function)

Probabilities are evaluated under the alternative hypothesis of a beneficial effect on all individual objectives

Simple disjunctive power

Definition

Probability of rejecting at least one hypothesis

$$\psi(n) = P\left(\sum_{i=1}^{m} r_i \geq 1\right)$$

Properties

May not be sufficiently sensitive

Does not differentiate between clinically distinct outcomes (single hypothesis is rejected versus all hypotheses are rejected)

Generalized disjunctive power

Definition

Probability of rejecting at least $k \geq 2$ hypotheses

$$\psi(n) = P\left(\sum_{i=1}^{m} r_i \geq k\right)$$

Properties

More flexible than simple disjunctive power

Example: Demonstrate that a significant effect is present at two or more dose levels

Other types of disjunctive power

Subset disjunctive power

Probability of rejecting at least one hypothesis in each pre-defined subset of hypotheses

Example: Demonstrate that a significant effect on at least one primary endpoint and at least one secondary endpoint in advanced multiplicity problems

Conjunctive power

Probability of rejecting all hypotheses

Weighted power

Definition

Weighted sum of marginal power functions

$$\psi(n) = \sum_{i=1}^{m} \nu_i \psi_i(n)$$

$v_1, \ldots, v_m$, Hypothesis importance measures

Properties

Enables trial's sponsor to account for relative importance of individual hypotheses

Simplifies to expected number of rejected hypotheses with equal importance

Clinical trial with multiple objectives

Sample size calculation

Sample size is found from $\psi(n) = 1 - \beta$

Closed-form solutions are not available

Simulations are required to evaluate relevant criterion function and sample size is found by grid search

2. Case study

Setting

Clinical trial with three dose-placebo comparisons

Four arms: Three doses of experimental treatment (Dose L, Dose M and Dose H) versus placebo

Balanced design (1:1:1:1 randomization)

Endpoint

Single normally distributed endpoint (PANSS total score)

Parameters of sample size calculation

Mean differences and common standard deviation

Comparison	Mean diff (δ)	Common SD (σ)
Dose H vs Placebo (H_1)	5	18
Dose M vs Placebo (H_2)	5	18
Dose L vs Placebo (H_3)	3.5	18

Positive dose-response relationship is anticipated

Parameters of sample size calculation

Test statistics

Test statistics are equicorrelated and follow a trivariate normal distribution (common correlation is 0.5)

Test statistic means under the alternative hypothesis:

$$\mu_i = \delta_i/\sqrt{2\sigma/n},\ i = 1, 2, 3$$

Error rates

$\alpha = 0.025$, One-sided familywise error rate

$\beta = 0.1$, Type II error rate (90% power)

Parameters of sample size calculation

Success criteria

Simple disjunctive power: Probability of establishing a significant effect at one or more doses

Generalized disjunctive power: Probability of establishing a significant effect at two or more doses, Probability of establishing a significant effect at all three doses

Multiple testing procedures

Nonparametric procedures

Bonferroni and Holm procedures

Semiparametric procedure

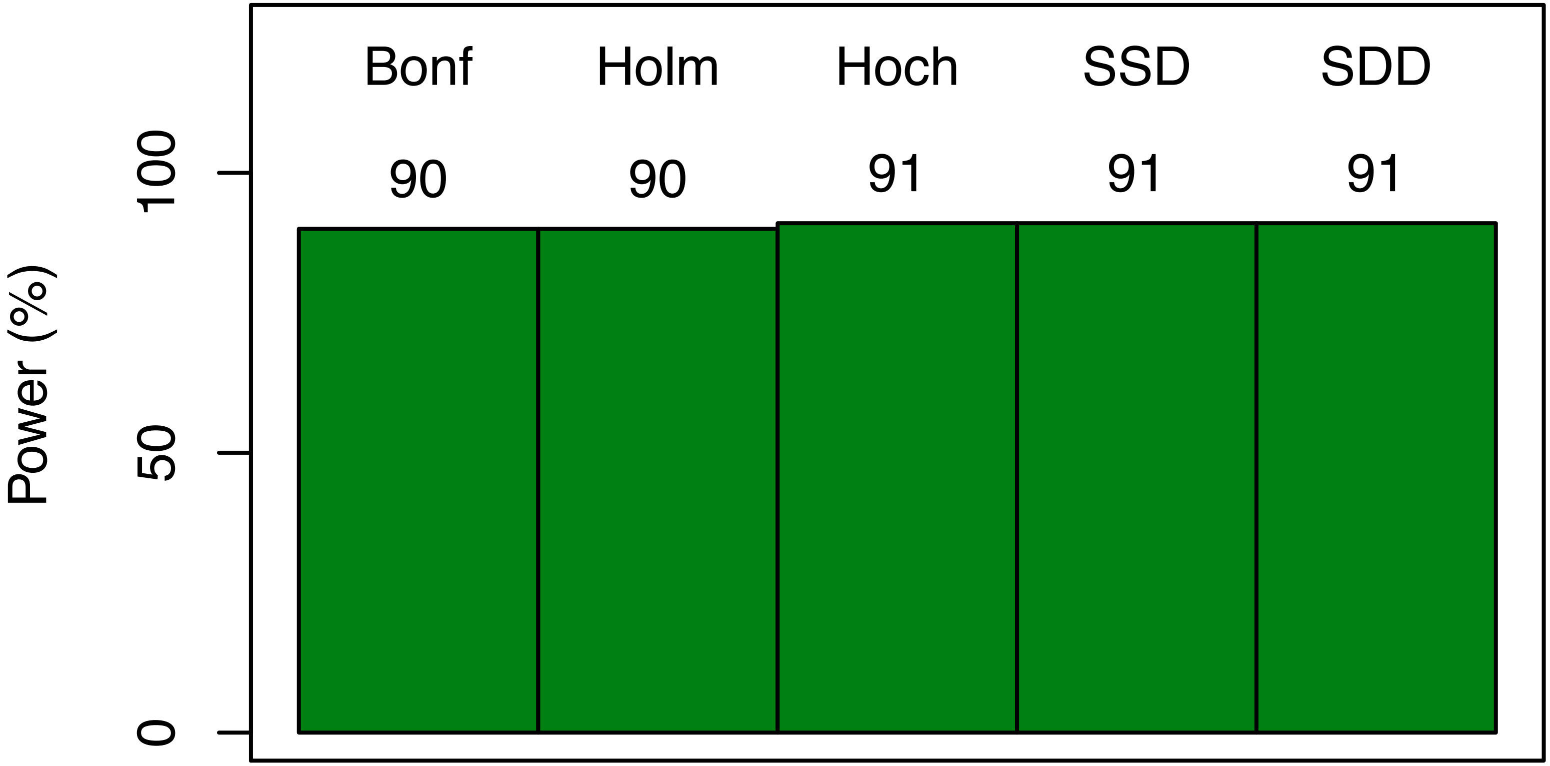
Hochberg procedure

Parametric procedures

Single-step Dunnett and step-down Dunnett procedures

Simple disjunctive power

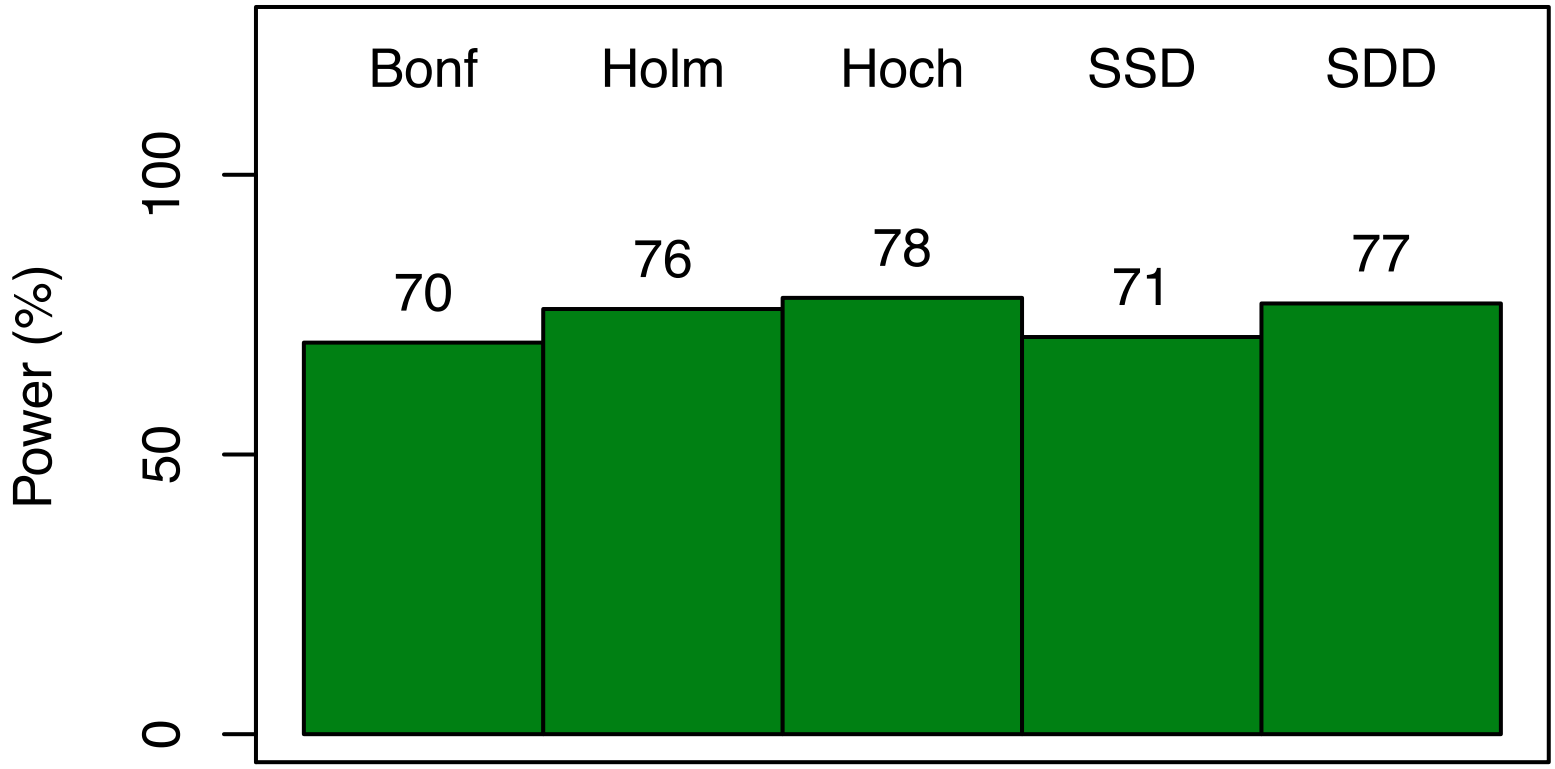
One, two or three significant doses ($n = 260$)



Bonf, Bonferroni procedure; Holm, Holm procedure; Hoch, Hochberg procedure; SSD, Single-step Dunnett procedure; SDD, Step-down Dunnett procedure

Generalized disjunctive power

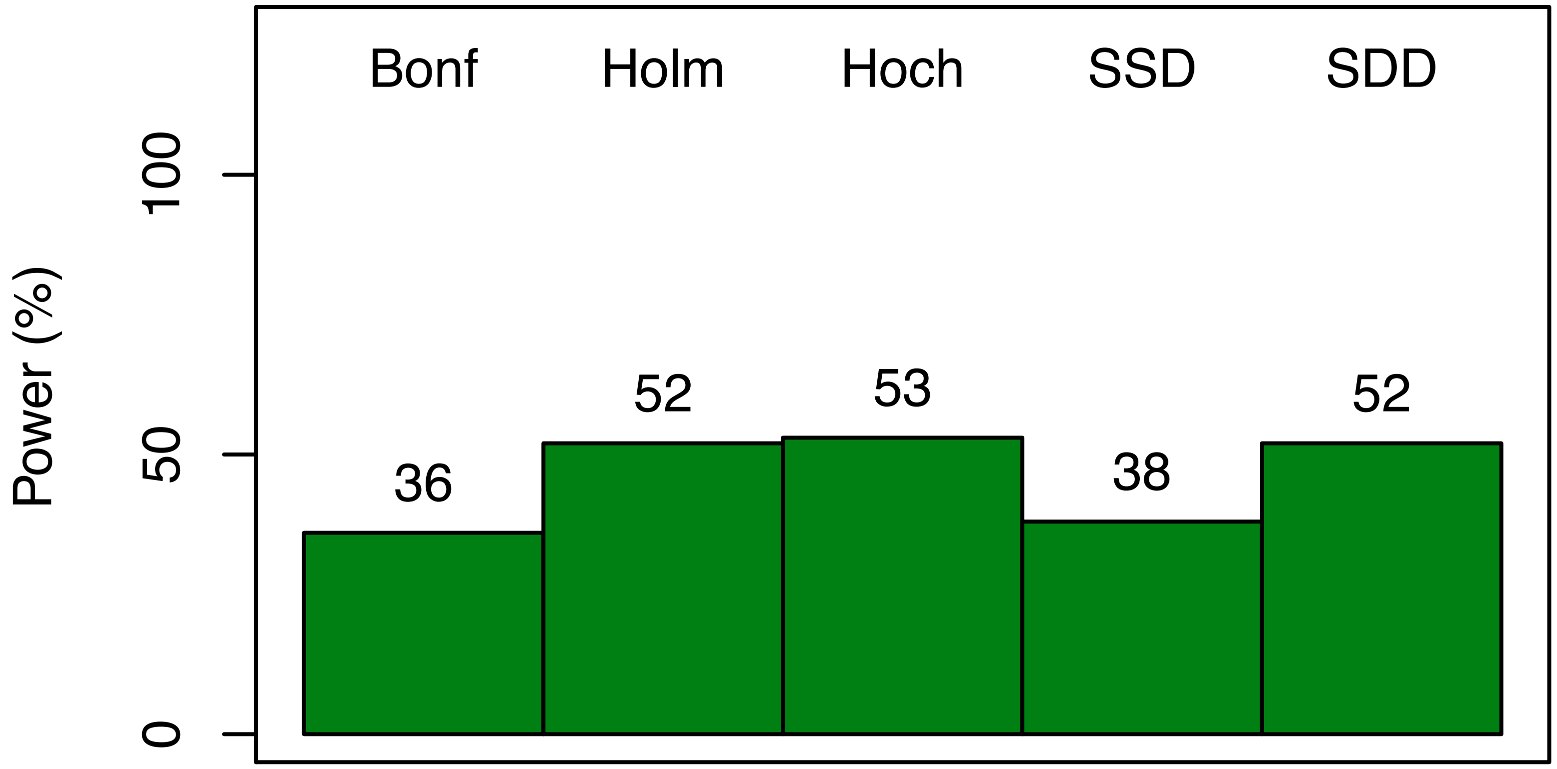
Two or three significant doses ($n = 260$)



Bonf, Bonferroni procedure; Holm, Holm procedure; Hoch, Hochberg procedure; SSD, Single-step Dunnett procedure; SDD, Step-down Dunnett procedure

Generalized disjunctive power

Three significant doses ($n = 260$)



Bonf, Bonferroni procedure; Holm, Holm procedure; Hoch, Hochberg procedure; SSD, Single-step Dunnett procedure; SDD, Step-down Dunnett procedure

Sample size calculations

Simple disjunctive power

Success criterion is not sensitive enough: Similar performance across all procedures

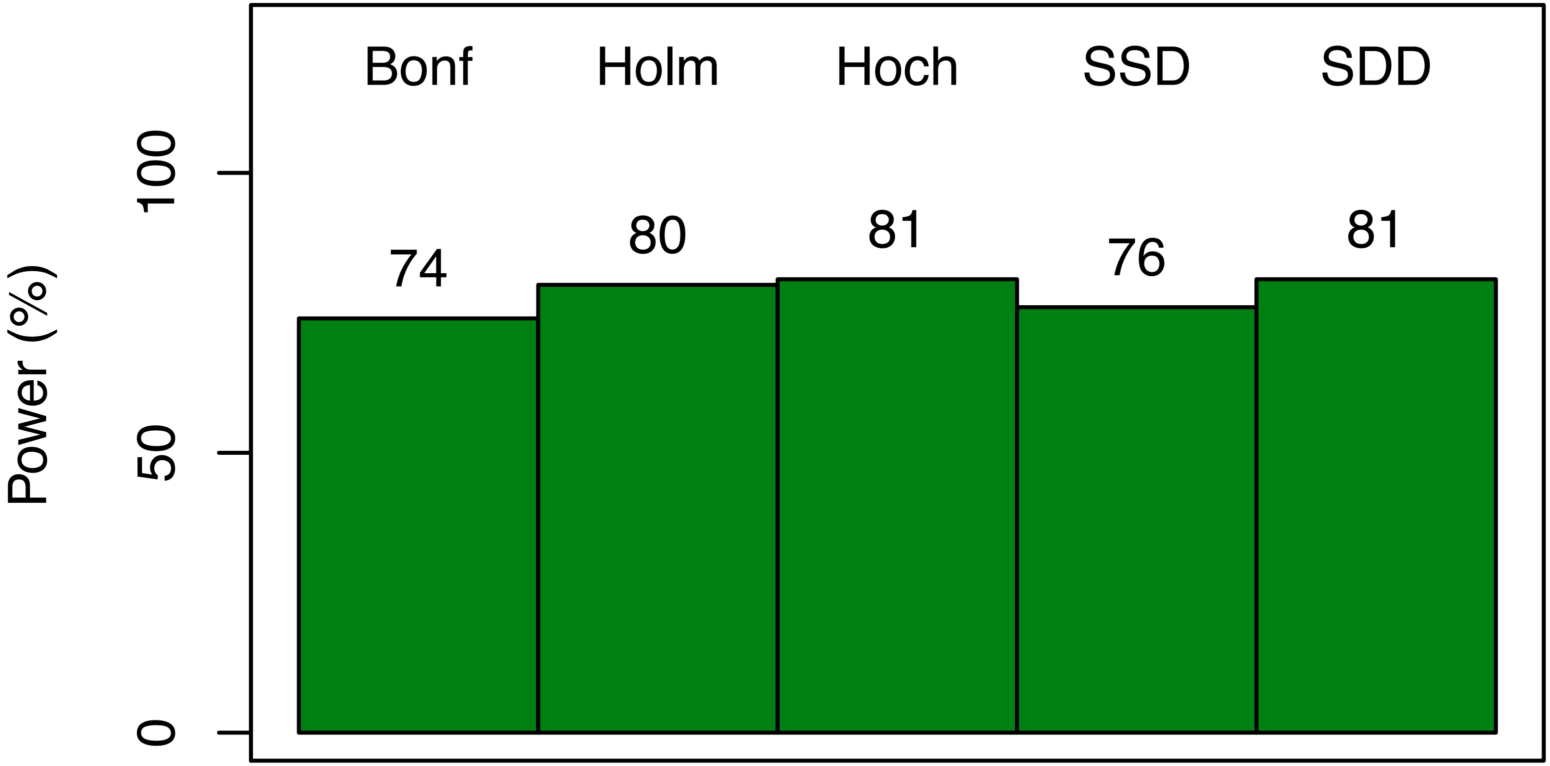
Generalized disjunctive power

Success criterion is more sensitive and demonstrates the importance of using more powerful procedures

Sample size calculations may be driven by generalized disjunctive power

Generalized disjunctive power

Two or three significant doses ($n = 280$)



Bonf, Bonferroni procedure; Holm, Holm procedure; Hoch, Hochberg procedure; SSD, Single-step Dunnett procedure; SDD, Step-down Dunnett procedure

Summary

Sample size calculations

Success criterion

Multiple ways to define overall power (success criterion) in clinical trials with multiple objectives

Success criterion needs to reflect the trial's clinical objectives

Sensitivity analysis

Recommended to evaluate criterion function under a broad range of plausible scenarios

Part II
Advanced Multiplicity Problems

Module G
Introduction to Gatekeeping Procedures

Outline

1. Classification of multiple testing problems

Problems with a single family and multiple families of null hypotheses

2. Clinical trial examples

3. Classification of gatekeepers

Gatekeepers with simple logical relationships and general logical relationships

4. Classification of testing strategies

Sequential testing, sequential testing with re-testing and simultaneous testing

1. Classification of multiple testing problems

Classification of multiple testing problems

Part I

Clinical trials with equally important objectives

A single family of null hypotheses

Part II

Clinical trials with ordered objectives

Multiple families of null hypotheses

Multiple families of null hypotheses

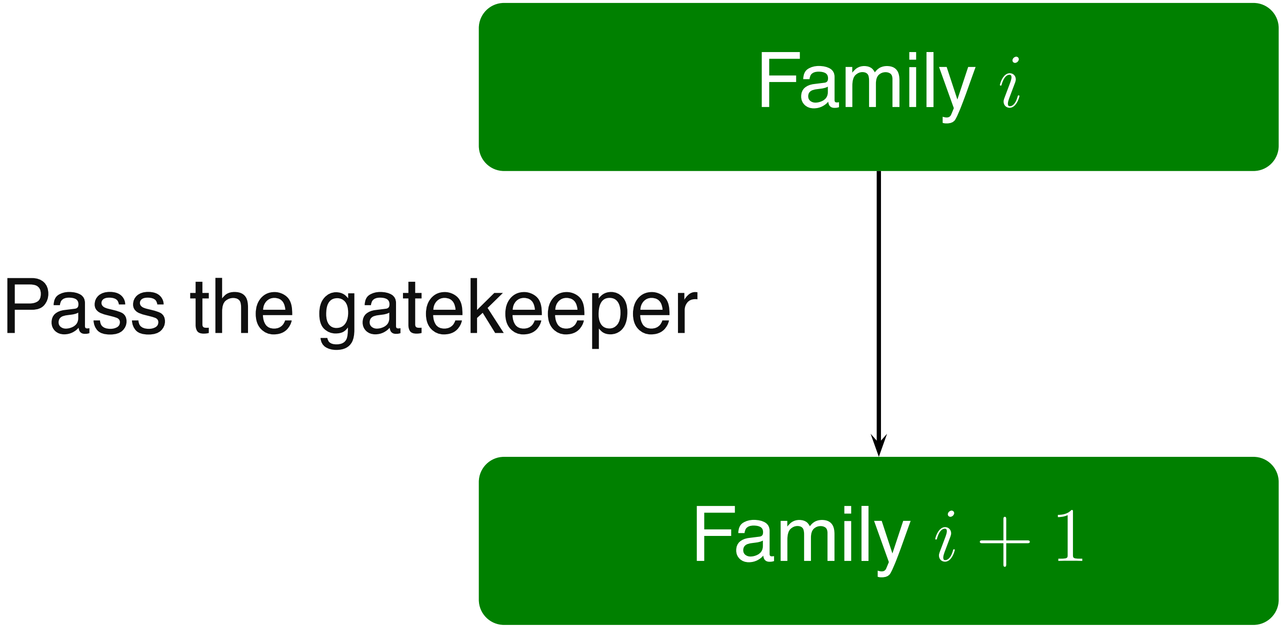
Family 1 — $H_1, \ldots, H_{k_1}$

$\ldots$

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

Gatekeeper

Family i ($i = 1, \ldots, m-1$) serves as a gatekeeper for Family $i+1$



Family i

Pass the gatekeeper

Family $i+1$

Multiple families of null hypotheses

Primary and secondary endpoints

Primary endpoints determine the trial’s outcome and key secondary endpoints provide useful supportive information about efficacy and safety

Primary and secondary populations

General population versus subgroups of patients who are more likely to benefit from treatment

Primary and secondary tests

Noninferiority assessment as the primary analysis followed by a superiority assessment

Regulatory position

FDA guidance document

Clinical studies section of labeling for prescription drugs and biologics

“The CLINICAL STUDIES section should present those endpoints that are essential to establishing the effectiveness of the drug (or that show the limitations of effectiveness) and those that provide additional useful and valid information about the activities of the drug”

Regulatory position

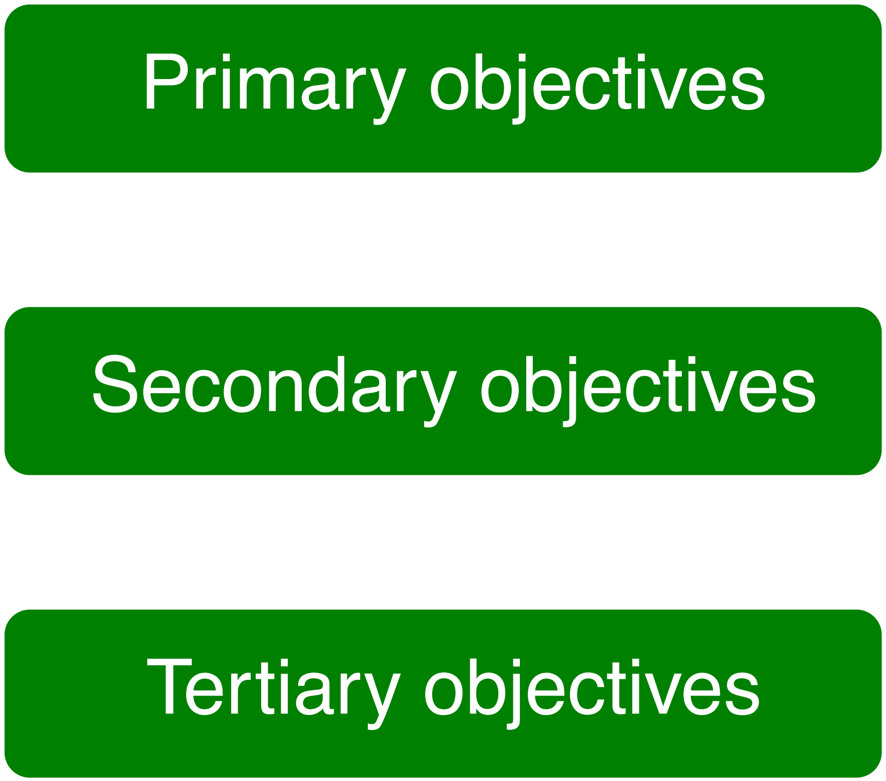
CPMP Points to Consider document

Points to consider on multiplicity issues in clinical trials

“Additional claims... [for] secondary variables... are possible only after the primary objective of the clinical trial has been achieved, and if the respective questions were pre-specified, and were part of an appropriately planned statistical analysis strategy”

Multiple objectives in clinical trials

Hierarchy of multiple objectives



Primary objectives

Secondary objectives

Tertiary objectives

Multiple objectives in clinical trials

Primary objectives

Directly related to the trial’s outcome and presented in product label using inferential statements (p-values and confidence intervals)

Secondary objectives

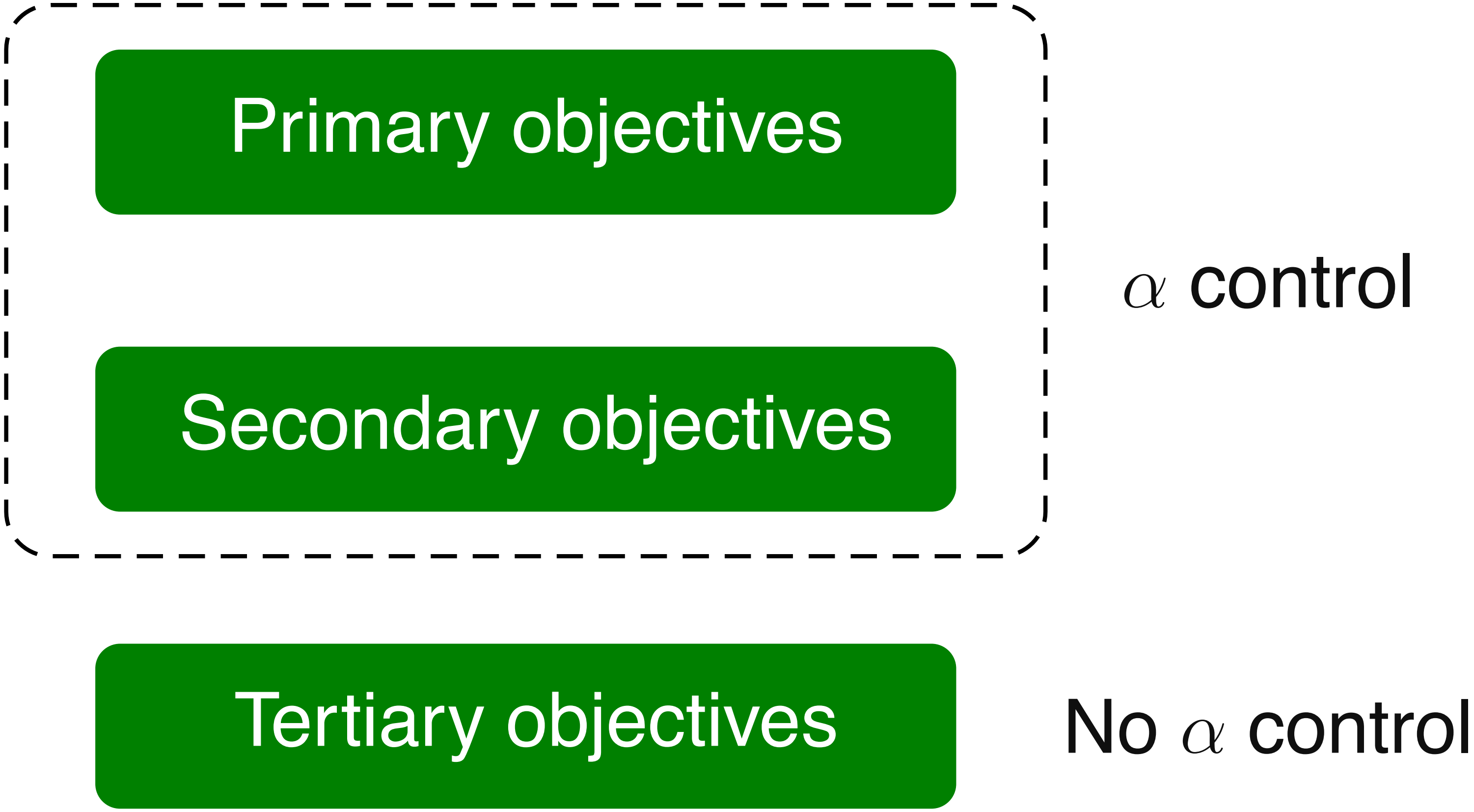
Provide key supportive evidence and presented in product label using inferential statements

Tertiary objectives

Play a general supportive role and presented in product label using descriptive statements (descriptive statistics)

Multiple objectives in clinical trials

Hierarchy of multiple objectives



Gatekeeping procedures

Definition

Multiple testing procedures for multiple families of null hypotheses

Global FWER control

Control familywise error rate over multiple families

Help enrich product labels by providing information on key secondary objectives

Optimal power

Maximize power by accounting for hierarchical structure of multiple families

2. Clinical trial examples

Clinical trial examples

Multiple sets of endpoints

Example 6: Alzheimer's disease trial

Example 7: Cardiovascular trial

Multiple doses and populations

Example 8: Schizophrenia trial

Multiple doses and test types

Example 9: Hypertension trial

Multiple sets of endpoints

Example 6: Alzheimer’s disease trial

Objective

Evaluate the effects of a treatment on cognition and global changes in patients with mild to moderate Alzheimer’s disease

Design

Treatment versus placebo

Multiple sets of endpoints

Primary endpoints

Endpoint 1: Cognition endpoint (ADAS-Cog)

Endpoint 2: Clinical global scale (CIBIC plus)

Treatment effect on both endpoints must be significant

Secondary endpoints

Endpoint 3: Biochemical marker

Endpoint 4: Imaging marker

Multiple sets of endpoints

Notation

δ_1, δ_2, δ_3 and δ_4, true mean treatment differences for four endpoints

Family 1

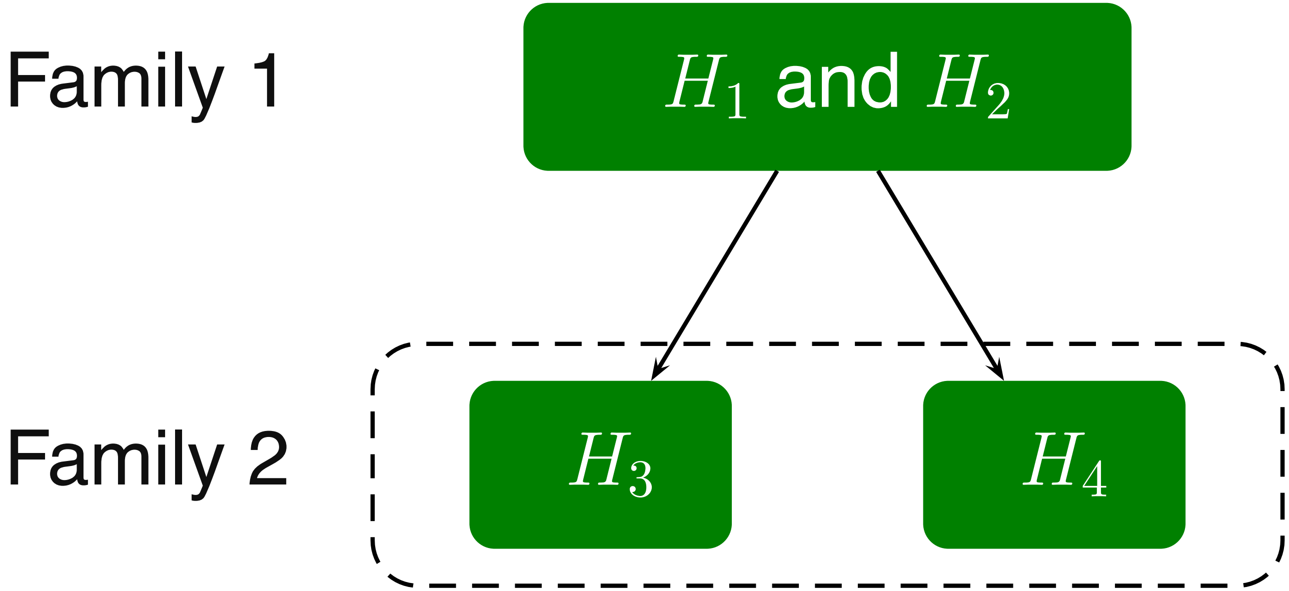
$H_1: \ \delta_1 \leq 0$ and $H_2: \ \delta_2 \leq 0$, null hypotheses of no effect for primary endpoints

Family 2

$H_3: \ \delta_3 \leq 0$ and $H_4: \ \delta_4 \leq 0$, null hypotheses of no effect for secondary endpoints

Multiple sets of endpoints

Decision tree



Null hypotheses in Family 2 are tested if both null hypotheses are rejected in Family 1

Multiple sets of endpoints

Example 7: Cardiovascular trial

Objective

Evaluate the effects of a treatment on morbidity and mortality in patients with acute myocardial infarction complicated by left ventricular dysfunction and heart failure

Design

Treatment versus placebo

Multiple sets of endpoints

Primary endpoints

Endpoint 1: All-cause mortality

Endpoint 2: Cardiovascular mortality and cardiovascular hospitalization

Treatment effect on at least one endpoint must be significant

Secondary endpoints

Endpoint 3: Cardiovascular mortality

Endpoint 4: All-cause mortality and all-cause hospitalization

Multiple sets of endpoints

Notation

δ_1, δ_2, δ_3 and δ_4, differences between true event rates for four endpoints

Family 1

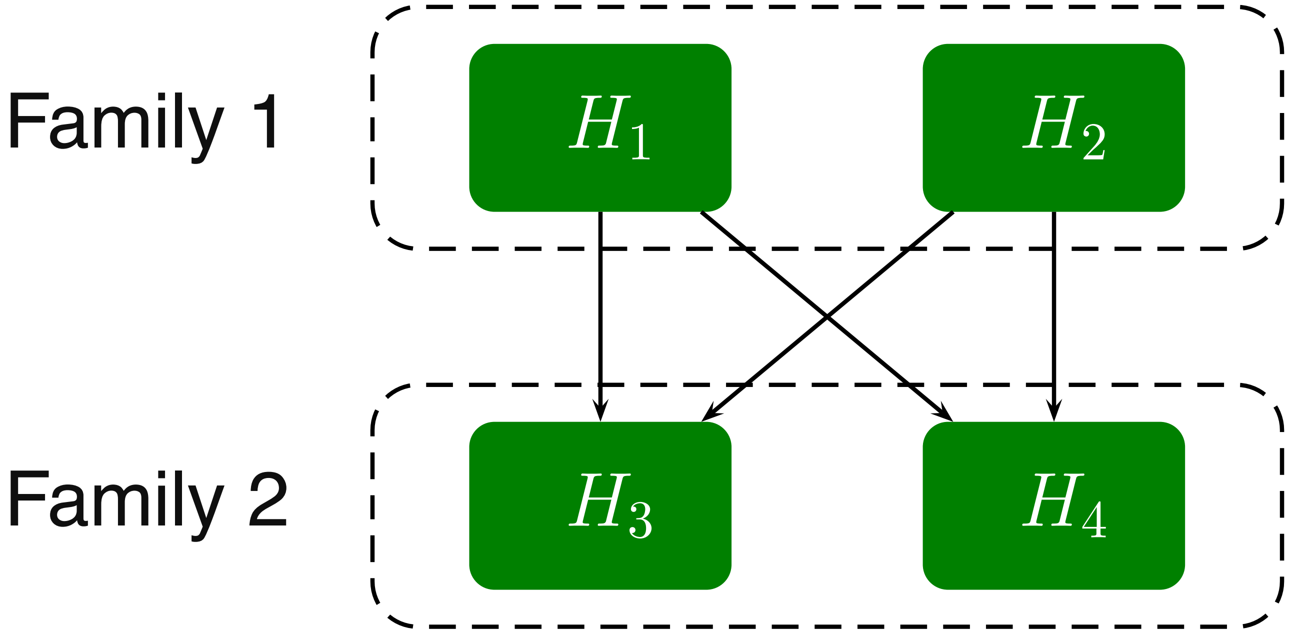
$H_1: \ \delta_1 \leq 0$ and $H_2: \ \delta_2 \leq 0$, null hypotheses of no effect for primary endpoints

Family 2

$H_3: \ \delta_3 \leq 0$ and $H_4: \ \delta_4 \leq 0$, null hypotheses of no effect for secondary endpoints

Multiple sets of endpoints

Decision tree



Null hypotheses in Family 2 are tested if at least one null hypothesis is rejected in Family 1

Multiple doses and populations

Example 8: Schizophrenia trial

Objective

Evaluate the efficacy of a treatment in patients diagnosed with schizophrenia

Design

Two doses of treatment (Doses L and H) versus placebo

Treatment effect on at least one dose must be significant

Multiple doses and populations

Primary endpoint

Positive and Negative Symptoms Scale (PANSS) total score

Two patient populations

General population and subpopulation (based on a genotypic classifier)

Multiple doses and populations

Notation

δ_1 and δ_2, true mean treatment differences for Doses L and H in general population

δ_3 and δ_4, true mean treatment differences for Doses L and H in subpopulation

Family 1

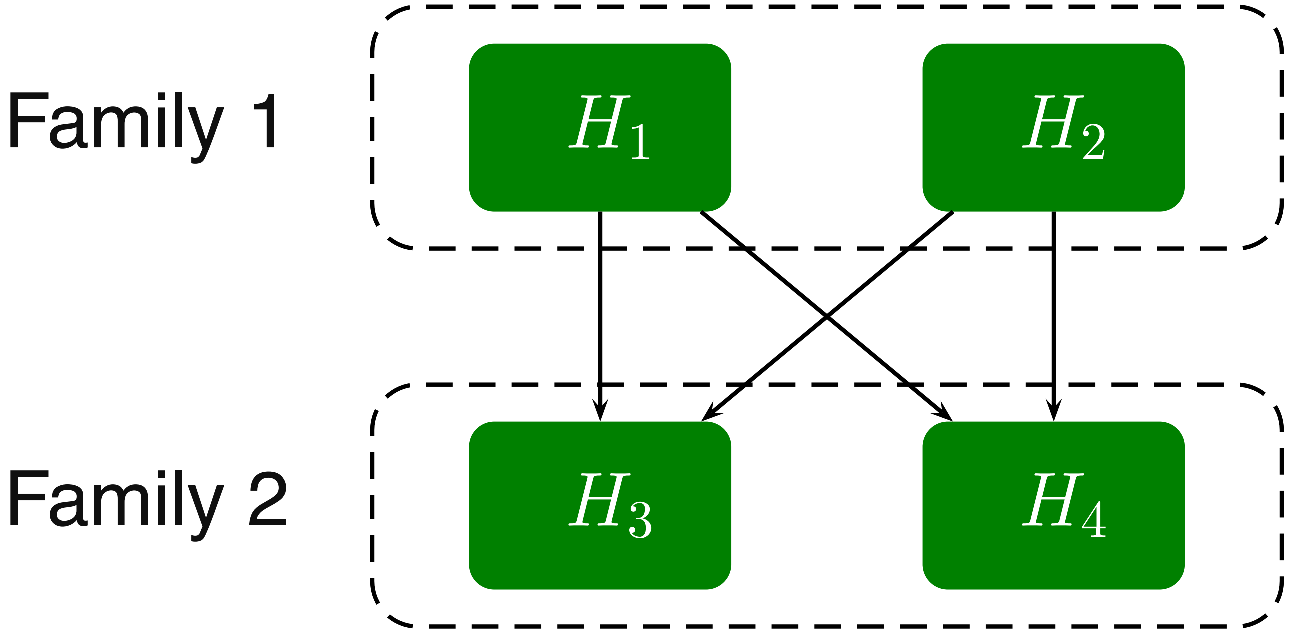
$H_1: \ \delta_1 \leq 0$ and $H_2: \ \delta_2 \leq 0$, null hypotheses of no effect for Doses L and H in general population

Family 2

$H_3: \ \delta_3 \leq 0$ and $H_4: \ \delta_4 \leq 0$, null hypotheses of no effect for Doses L and H in subpopulation

Multiple doses and populations

Decision tree



Null hypotheses in Family 2 are tested if at least one null hypothesis is rejected in Family 1

Multiple doses and test types

Example 9: Hypertension trial

Objective

Evaluate the effects of a treatment on blood pressure

Design

Two doses of treatment (Doses L and H) versus Active control

Treatment effect on at least one dose must be significant

Multiple doses and test types

Primary endpoint

Systolic blood pressure (based on ambulatory blood pressure monitoring)

Two test types

Noninferiority to Active control is evaluated first

Superiority to Active control is evaluated after noninferiority is established

Multiple doses and test types

Notation

δ_1 and δ_2, true mean treatment differences for Doses L and H

γ, positive noninferiority margin

Family 1

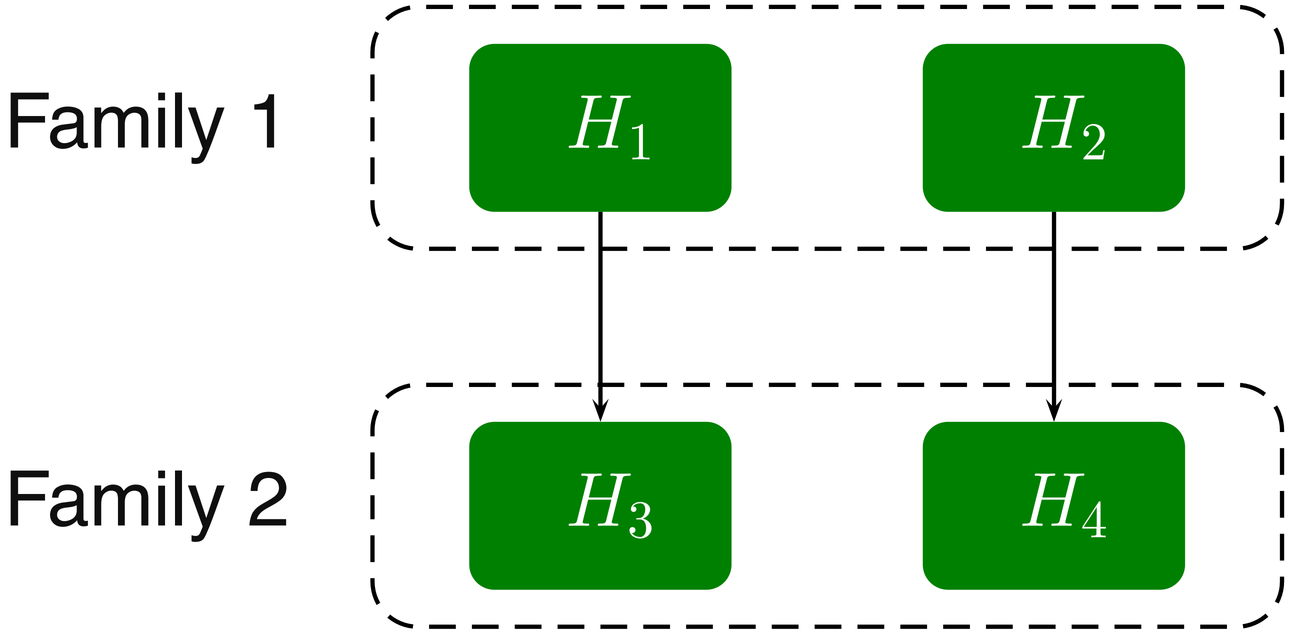
$H_1: \ \delta_1 \leq -\gamma$ and $H_2: \ \delta_2 \leq -\gamma$, null hypotheses of inferiority to Active control for Doses L and H

Family 2

$H_3: \ \delta_1 \leq 0$ and $H_4: \ \delta_2 \leq 0$, null hypotheses of lack of superiority to Active control for Doses L and H

Multiple doses and test types

Decision tree



Superiority test is carried out only if noninferiority is established (H_3 is tested only if H_1 is rejected and H_4 is tested only if H_2 is rejected)

3. Classification of gatekeepers

Classification of gatekeepers

Logical relationships

Gatekeepers define logical relationships among null hypotheses in different families

Simple logical relationships

Serial gatekeepers

Parallel gatekeepers

General logical relationships

General gatekeepers

References

Serial gatekeepers

Maurer, Hothorn and Lehmacher (1995), Westfall and Krishen (2001)

Parallel gatekeepers

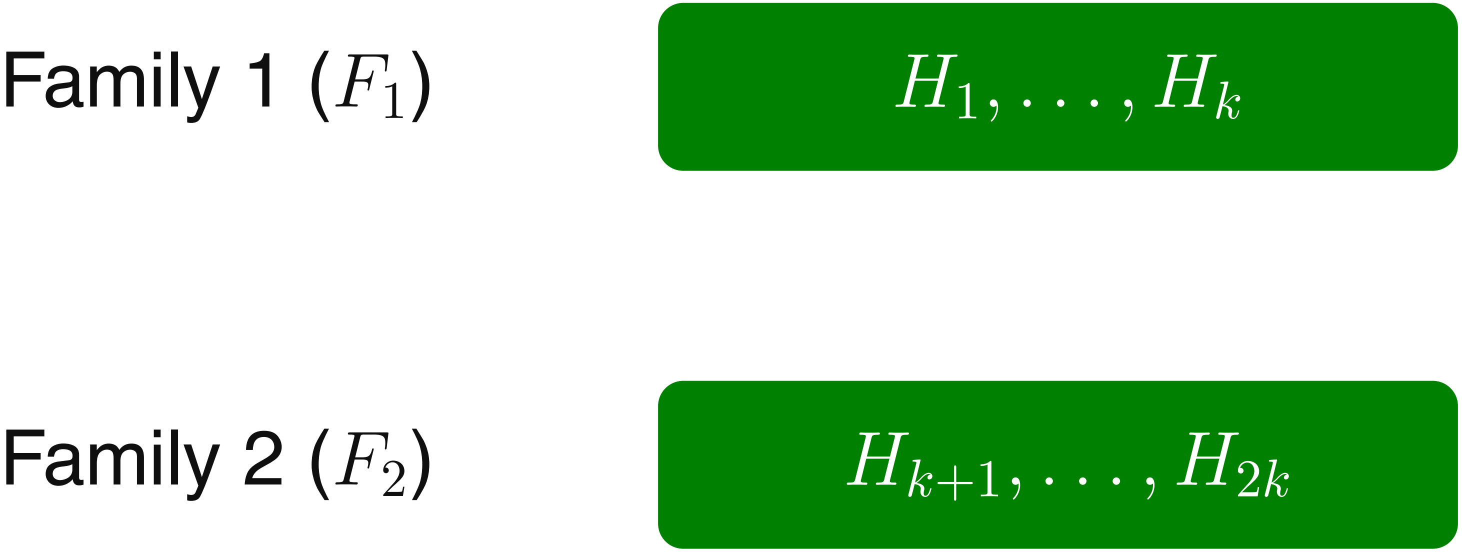
Dmitrienko, Offen and Westfall (2003), Dmitrienko and Tamhane (2009)

General gatekeepers

Dmitrienko and Tamhane (2011, 2013)

Two-family testing problem

Decision tree



Family 1 (F_1)

$H_1, \dots, H_k$

Family 2 (F_2)

$H_{k+1}, \dots, H_{2k}$

Two-family testing problem

Family 1

$F_1 = \{H_1, \ldots, H_k\}$, null hypotheses

$N_1 = \{1, \ldots, k\}$, index set

Family 2

$F_2 = \{H_{k+1}, \ldots, H_{2k}\}$, null hypotheses

$N_2 = \{k + 1, \ldots, 2k\}$, index set

Serial gatekeepers

Definition

Family 1 is a serial gatekeeper for Family 2, i.e., all hypotheses must be rejected in Family 1 to proceed to Family 2

Example

Example 6: Alzheimer’s disease trial

$F_1 = \{H_1, H_2\}$ is a serial gatekeeper for $F_2 = \{H_3, H_4\}$

Parallel gatekeepers

Definition

Family 1 is a parallel gatekeeper for Family 2, i.e., at least one hypothesis must be rejected in Family 1 to proceed to Family 2

Examples

Example 7: Cardiovascular trial

Example 8: Schizophrenia trial

$F_1 = \{H_1, H_2\}$ is a parallel gatekeeper for $F_2 = \{H_3, H_4\}$

General gatekeepers

Testable null hypotheses

Arbitrary logical relationships in multiple families can be defined by specifying a set of testable null hypotheses for each set of rejected hypotheses in Family 1

Special cases

Serial gatekeepers

Parallel gatekeepers

General gatekeepers

Restriction functions

Logical relationships between hypotheses in Families 1 and 2 are specified using a family of binary restriction functions

$L_i(I_1) = 0$ or 1, where $I_1 \subseteq N_1$, for any hypothesis H_i in Family 2

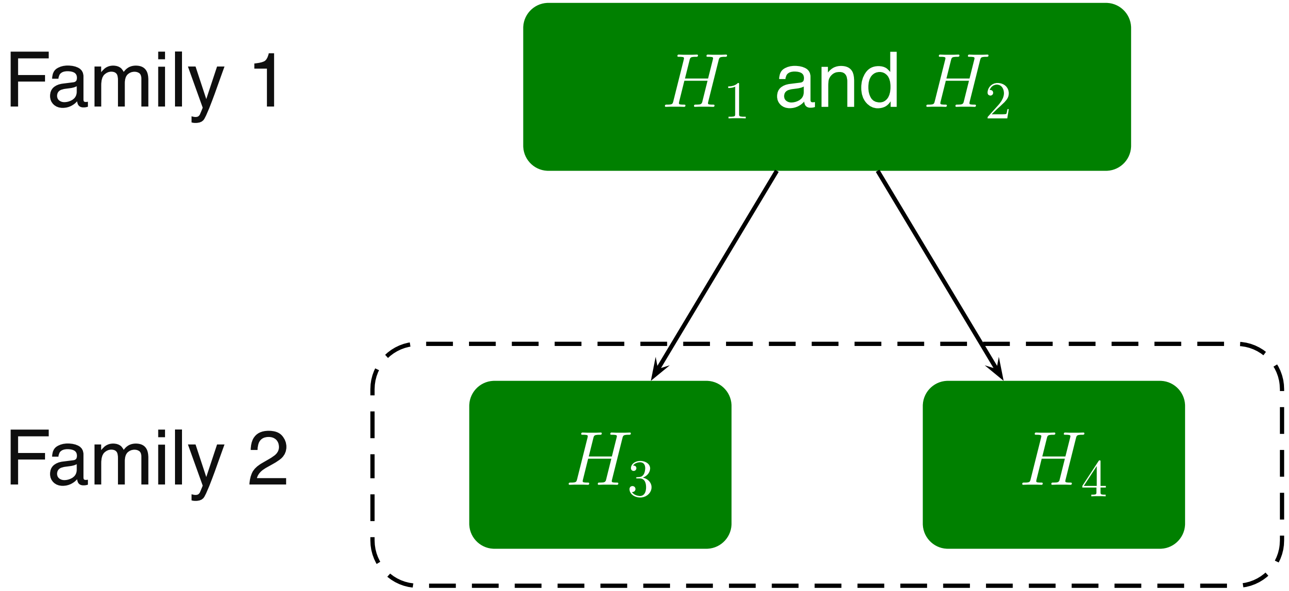
Testable null hypotheses

Null hypothesis H_i in Family 2 is testable if and only if $L_i(R_1) = 1$

R_1, index set of hypotheses rejected in Family 1

Example 6: Alzheimer's disease trial

Family 1 is a serial gatekeeper



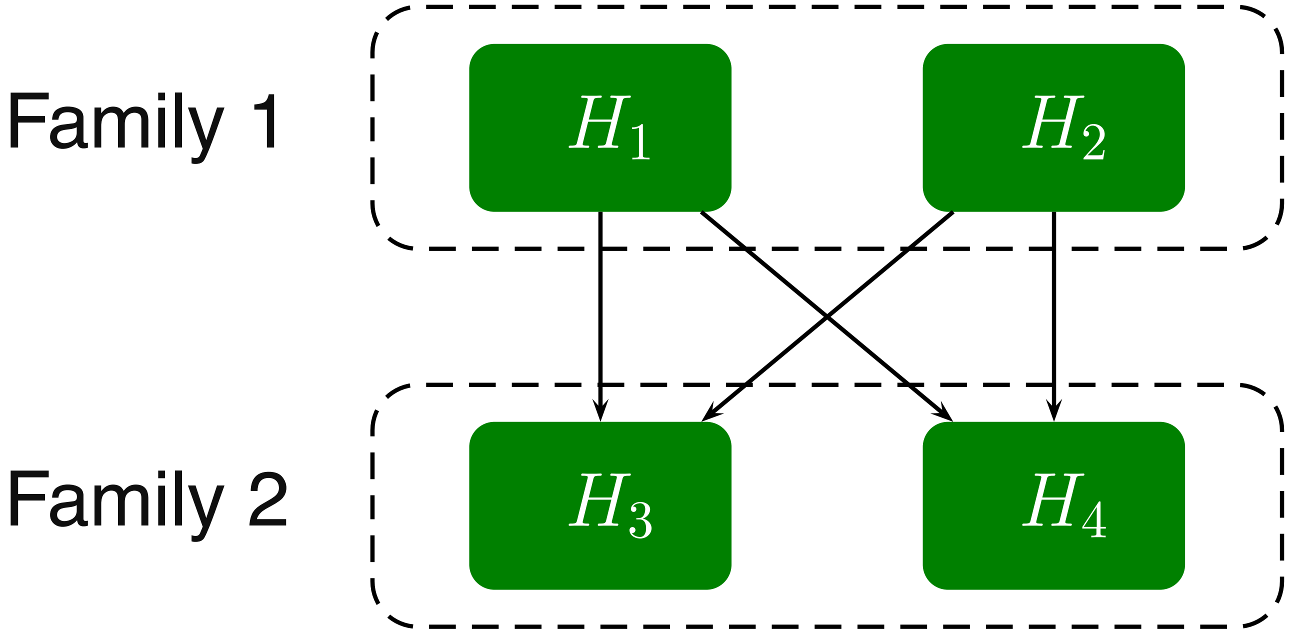
Example 6: Alzheimer's disease trial

Restriction functions

Set of rejected null hypotheses	Null hypothesis	Testable
H_1, H_2	H_3	Yes
H_1	H_3	No
H_2	H_3	No
Empty	H_3	No
H_1, H_2	H_4	Yes
H_1	H_4	No
H_2	H_4	No
Empty	H_4	No

Example 7: Cardiovascular trial

Family 1 is a parallel gatekeeper



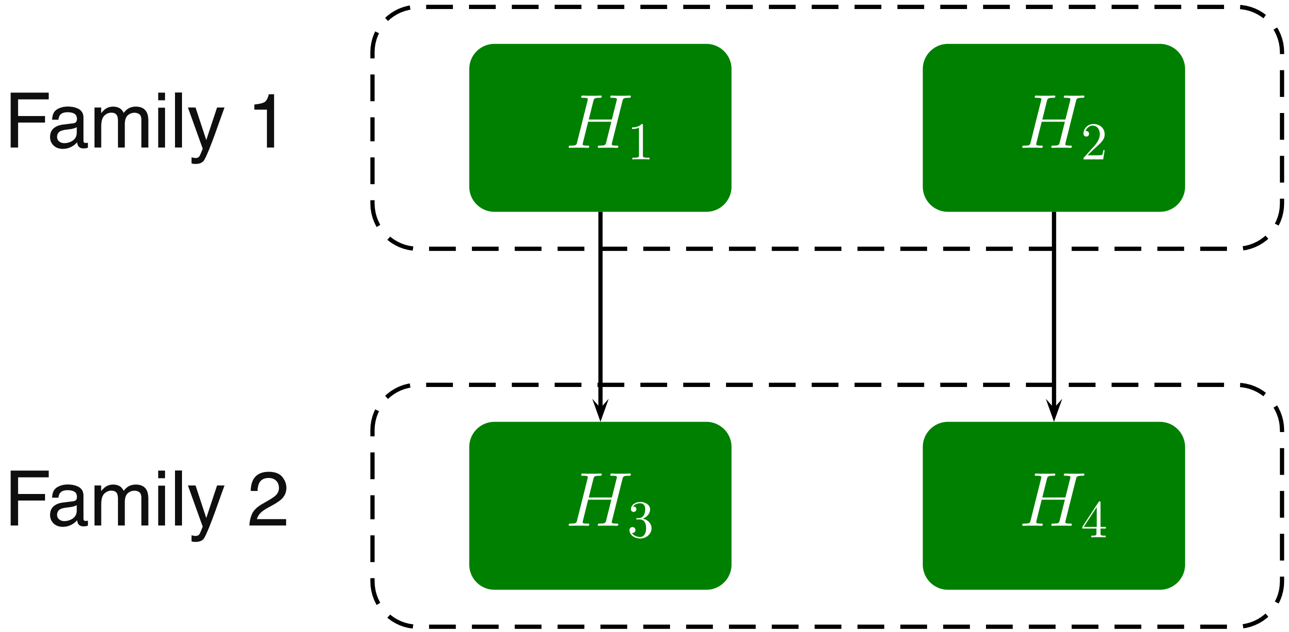
Example 7: Cardiovascular trial

Restriction functions

Set of rejected null hypotheses	Null hypothesis	Testable
H_1, H_2	H_3	Yes
H_1	H_3	Yes
H_2	H_3	Yes
Empty	H_3	No
H_1, H_2	H_4	Yes
H_1	H_4	Yes
H_2	H_4	Yes
Empty	H_4	No

Example 9: Hypertension trial

Family 1 is a general gatekeeper



Example 9: Hypertension trial

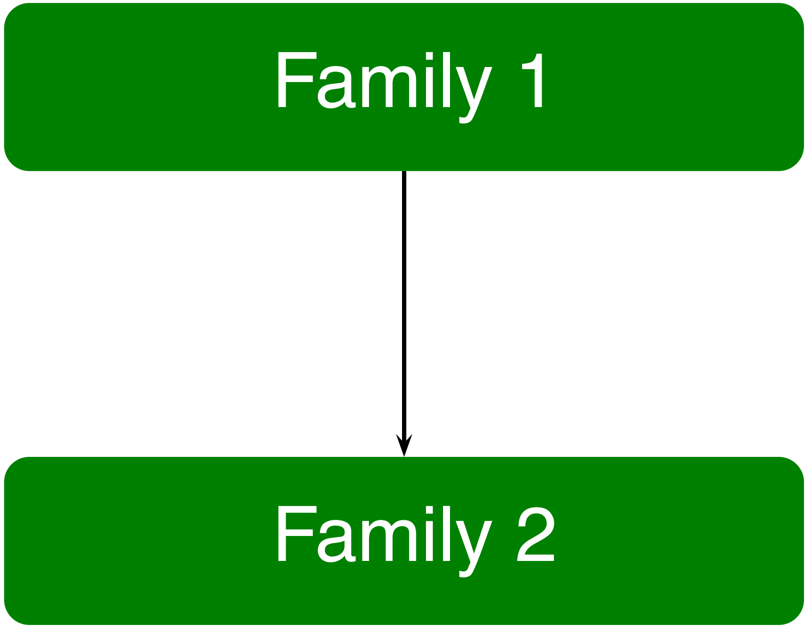
Restriction functions

Set of rejected null hypotheses	Null hypothesis	Testable
H_1, H_2	H_3	Yes
H_1	H_3	Yes
H_2	H_3	No
Empty	H_3	No
H_1, H_2	H_4	Yes
H_1	H_4	No
H_2	H_4	Yes
Empty	H_4	No

4. Classification of testing strategies

Sequential testing

Decision tree

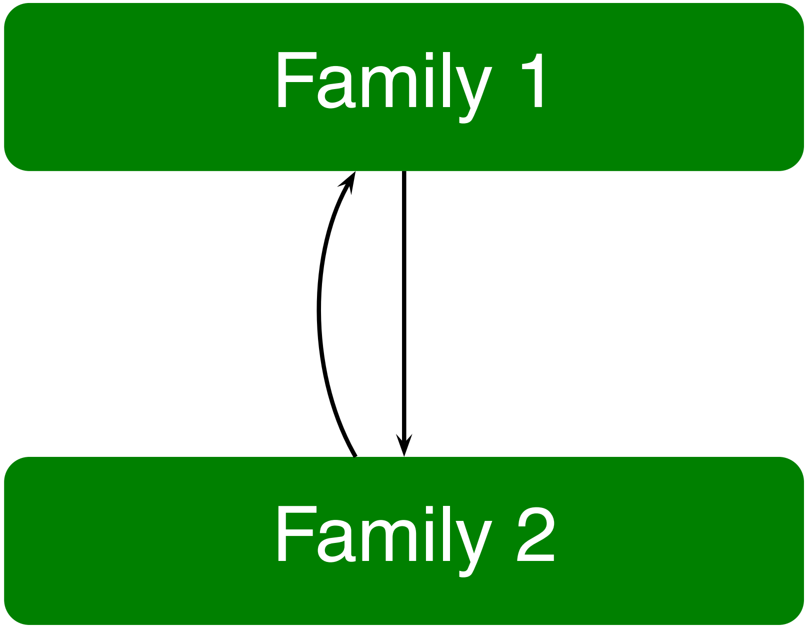


Families of hypotheses are tested sequentially starting with Family 1

Error rate is transferred along the sequence

Sequential testing with re-testing

Decision tree

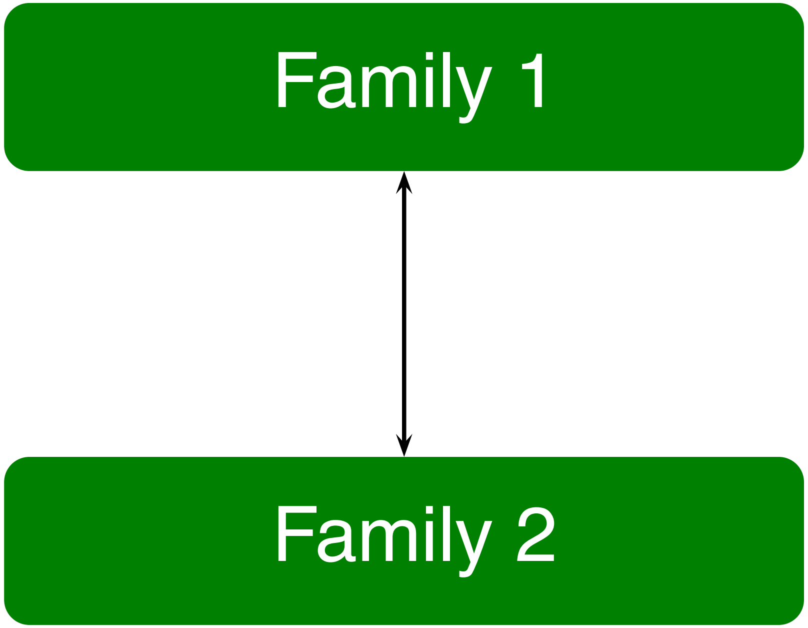


Families of hypotheses are tested sequentially starting with Family 1 with a re-testing loop

Error rate is transferred along the sequence and then back to Family 1

Simultaneous testing

Decision tree



Families of hypotheses are tested simultaneously

Error rate is transferred among families

Module H
Problems with Serial and Parallel Gatekeepers

Outline

1. **Serial gatekeepers**

Serial gatekeeping procedures

2. **Parallel gatekeepers**

Simple and advanced parallel gatekeeping procedures

3. **Sequential testing with re-testing**

Gatekeeping procedures with re-testing loop

Outline

4. Simulation study

Comparison of nonparametric and semiparametric gatekeeping procedures

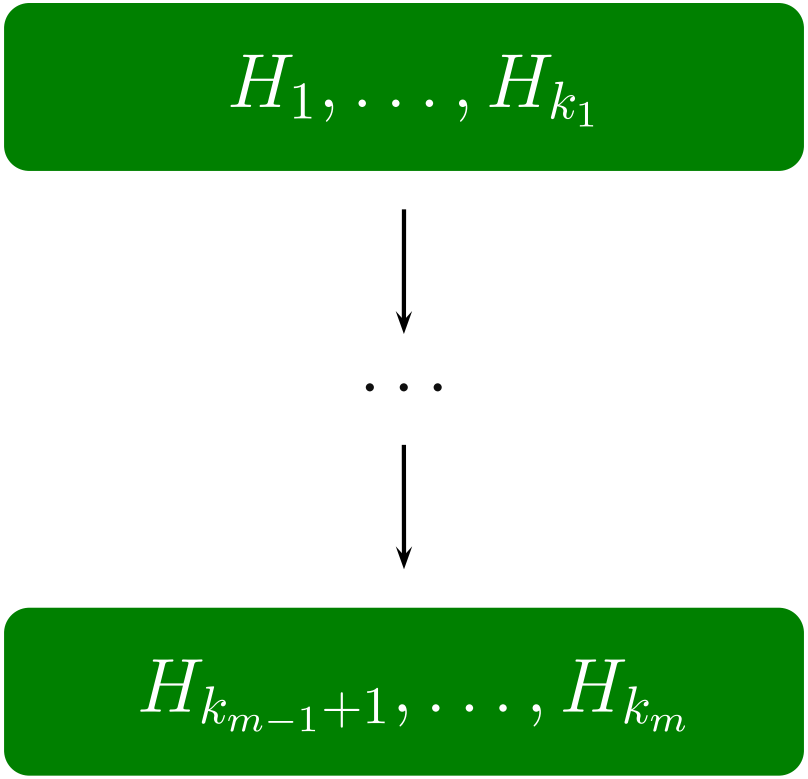
5. Software implementation

Software implementation of gatekeeping procedures in SAS and R

1. Serial gatekeepers

General setting

Multiple families of null hypotheses



Family 1 (F_1)

$H_1, \dots, H_{k_1}$

$\dots$

Family m (F_m)

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

Serial gatekeeping procedures

Serial gatekeepers

F_j is a serial gatekeeper for F_{j+1} (all hypotheses must be rejected in F_j to proceed to F_{j+1}, $j = 1, \ldots, m-1$)

Gatekeeping procedure

Build from component procedures applied within each family to control global familywise error rate (FWER) at α

Serial gatekeeping procedures

Component procedures

Family F_1: All-or-none procedure (all tests are carried out at α)

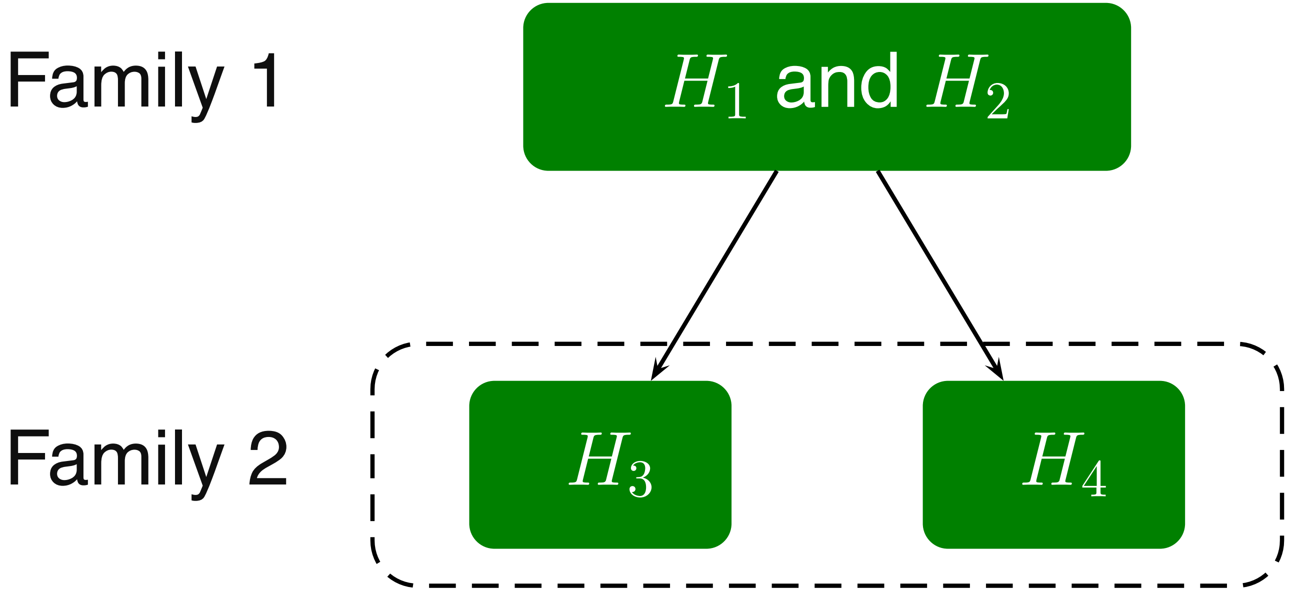
$\ldots$

Family F_{m-1}: All-or-none procedure (all tests are carried out at α)

Family F_m: Any procedure with local FWER controlled at α

Example 6: Alzheimer's disease trial

Decision tree



Example 6: Alzheimer’s disease trial

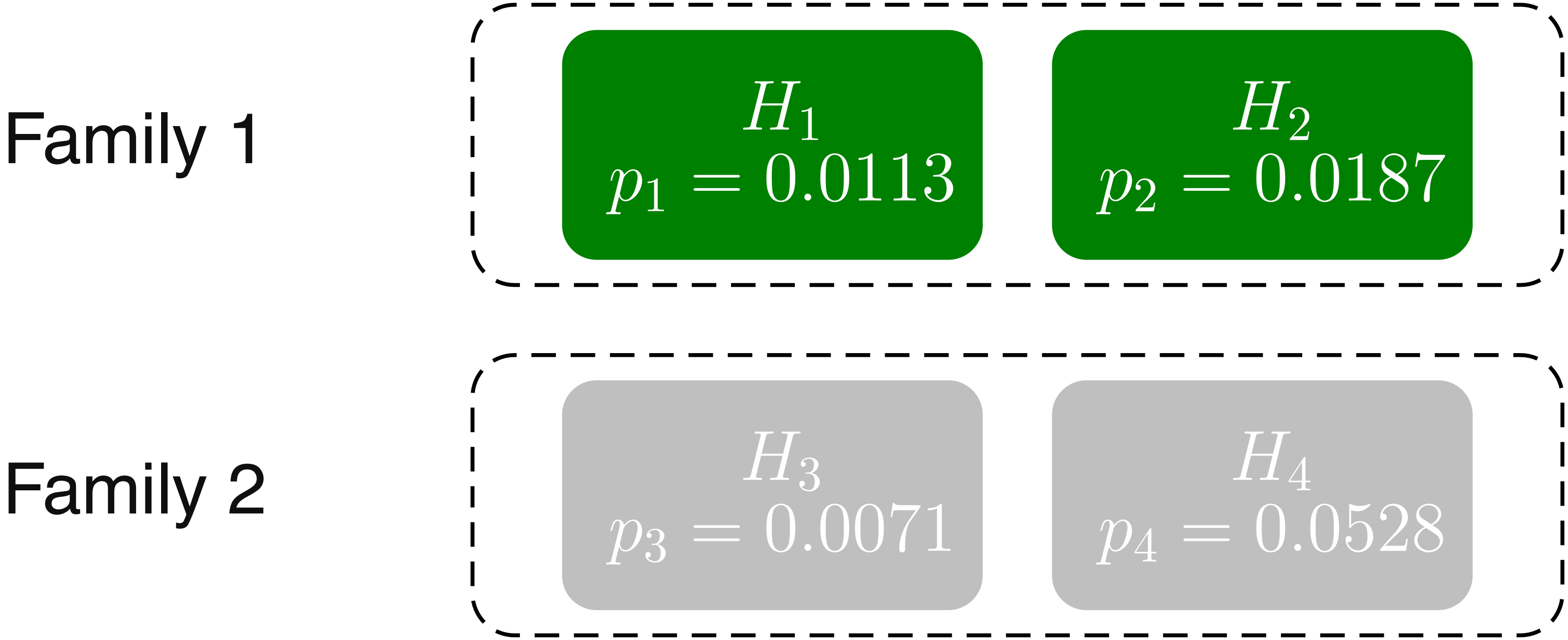
Four hypothesis tests

Family	Hypothesis	Raw p-value
Family 1	H_1	0.0113
	H_2	0.0187
Family 2	H_3	0.0071
	H_4	0.0528

$\alpha = 0.025$, Global familywise error rate

Example 6: Alzheimer's disease trial

Family 1: All-or-none procedure



Family 1

H_1
$p_1 = 0.0113$

H_2
$p_2 = 0.0187$

Family 2

H_3
$p_3 = 0.0071$

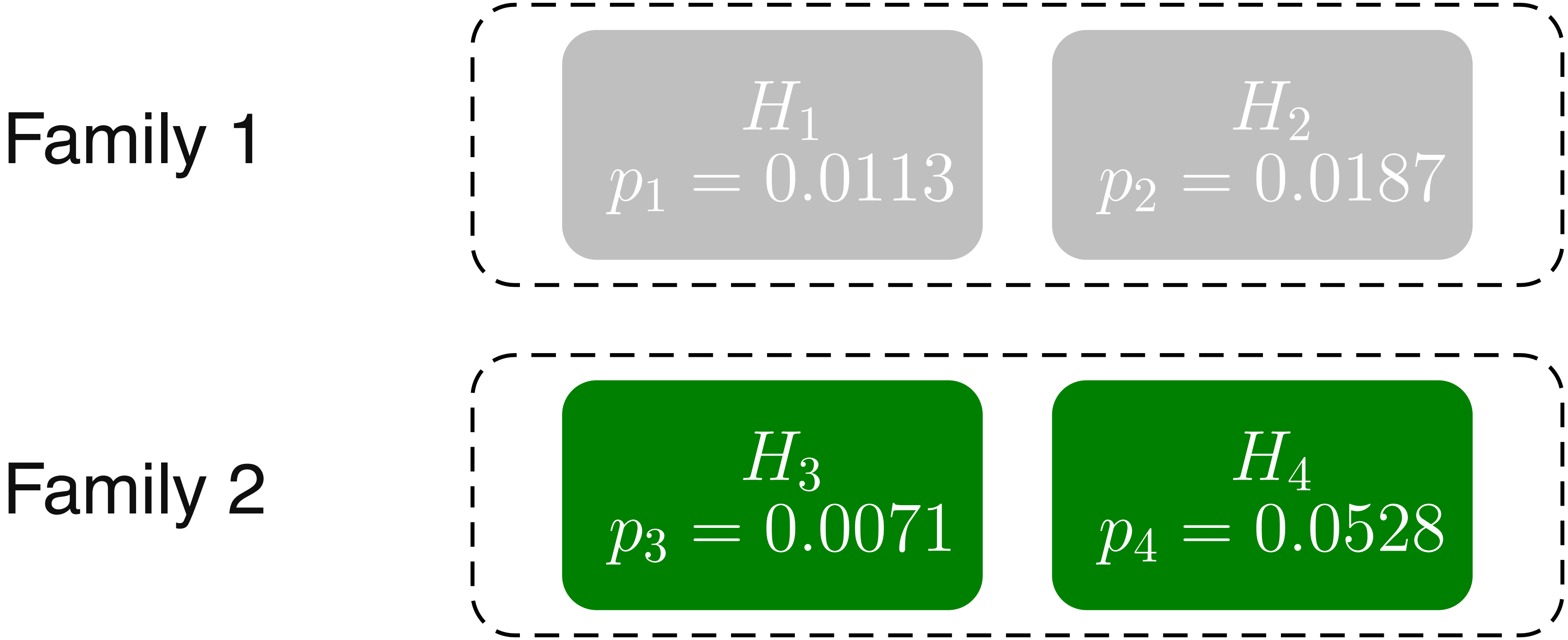
H_4
$p_4 = 0.0528$

Test H_1 and H_2 at $\alpha = 0.025$

H_1 and H_2 are rejected

Example 6: Alzheimer's disease trial

Family 2: Holm procedure



Family 1

H_1 $p_1 = 0.0113$

H_2 $p_2 = 0.0187$

Family 2

H_3 $p_3 = 0.0071$

H_4 $p_4 = 0.0528$

Test H_3 at $\alpha/2 = 0.0125$

Test H_4 at $\alpha = 0.025$ if H_3 is rejected

H_3 is rejected and H_4 is accepted

Serial gatekeepers

Clinical trial applications

Commonly used in clinical trials

Described in CPMP guidance document (Points to consider on multiplicity issues in clinical trials, CPMP/EWP/908/99, Sep 19, 2002)

Caveats

Important to have sufficient historical data to prioritize families of hypotheses

Null hypotheses toward end of testing sequence are likely to be accepted

2. Parallel gatekeepers

Parallel gatekeepers

Simple gatekeeping procedures

Gatekeeping procedures derived from nonparametric chain procedures

Advanced gatekeeping procedures

Multistage gatekeeping procedures derived from general component procedures (nonparametric, semiparametric and parametric)

Two-family testing problem

Decision tree

Family 1 $H_1, \ldots, H_k$

Family 2 $H_{k+1}, \ldots, H_{2k}$

Parallel gatekeeping procedures

Parallel gatekeeper

Family 1 is a parallel gatekeeper for Family 2, i.e., at least one null hypothesis must be rejected in Family 1 to proceed to Family 2

Gatekeeping procedures

Methods for constructing parallel gatekeeping procedures are considerably more complicated than those used for serial gatekeeping procedures

Parallel gatekeeping procedures

Naive approach

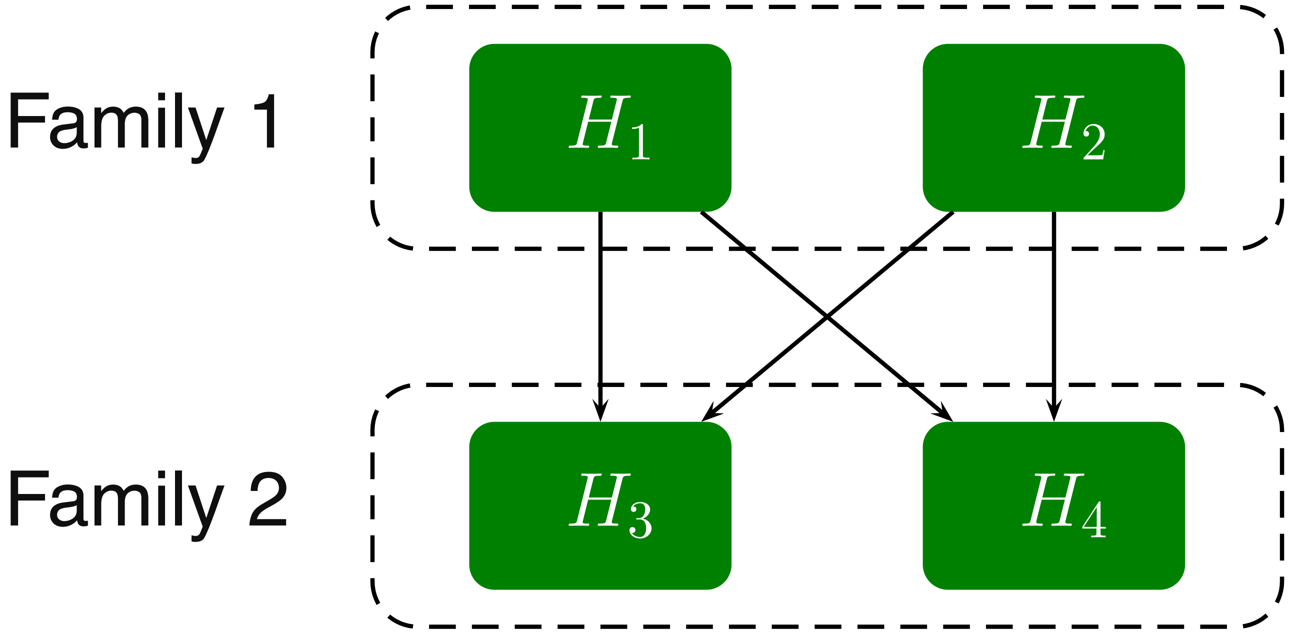
In trials with serial gatekeepers, component procedures control FWER locally at α within each family

Resulting gatekeeping procedure protects global FWER at α

Is it sufficient to control FWER locally in problems with parallel gatekeepers to protect global FWER?

Example 8: Schizophrenia trial

Decision tree



Family 1: Doses L and H versus Placebo in overall population

Family 2: Doses L and H versus Placebo in subpopulation

Naive gatekeeping procedure

Example 8: Schizophrenia trial

Family 1: Bonferroni procedure

Reject H_1 if $p_1 \leq \alpha/2$

Reject H_2 if $p_2 \leq \alpha/2$

Gatekeeper is passed if at least one null hypothesis is rejected

Family 2: Bonferroni procedure

Reject H_3 if $p_3 \leq \alpha/2$

Reject H_4 if $p_4 \leq \alpha/2$

Naive gatekeeping procedure

Global FWER control

Global FWER in Families 1 and 2 is controlled at α if probability of incorrectly rejecting a true null hypothesis $\leq \alpha$ for all configurations of true and false null hypotheses

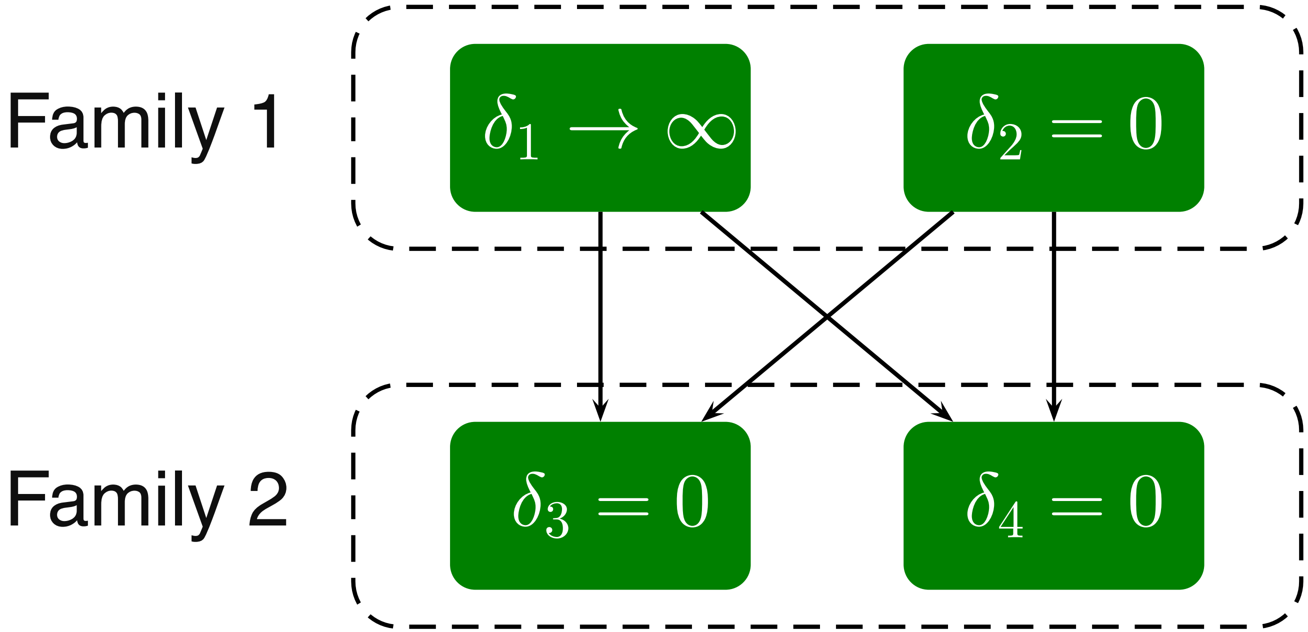
Configuration

Extremely large effect for Dose 1 and no effect for Dose 2 in general population, i.e., $\delta_1 \to \infty$ and $\delta_2 = 0$

No effect for Doses 1 and 2 in subpopulation, i.e., $\delta_3 = 0$ and $\delta_4 = 0$

Naive gatekeeping procedure

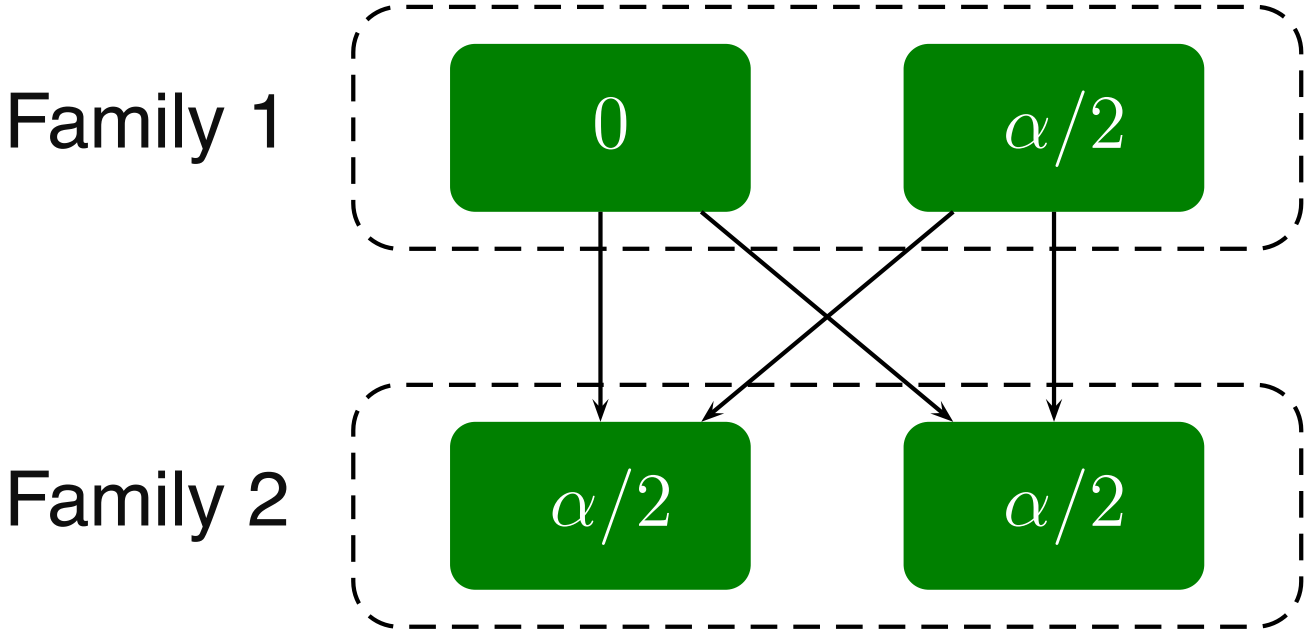
True mean treatment differences



Error rate for this configuration of true and false null hypotheses is probability of incorrectly rejecting at least one null hypothesis (H_2, H_3 or H_4)

Naive gatekeeping procedure

Individual error rates



Gatekeeper is passed virtually 100% of the time since δ_1 is very large and thus p_1 is very small

Naive mixture procedure inflates global FWER ($\simeq \alpha/2 + \alpha/2 + \alpha/2 > \alpha$)

Parallel gatekeeping procedures

Naive approach

Local FWER control does not translate into global FWER control

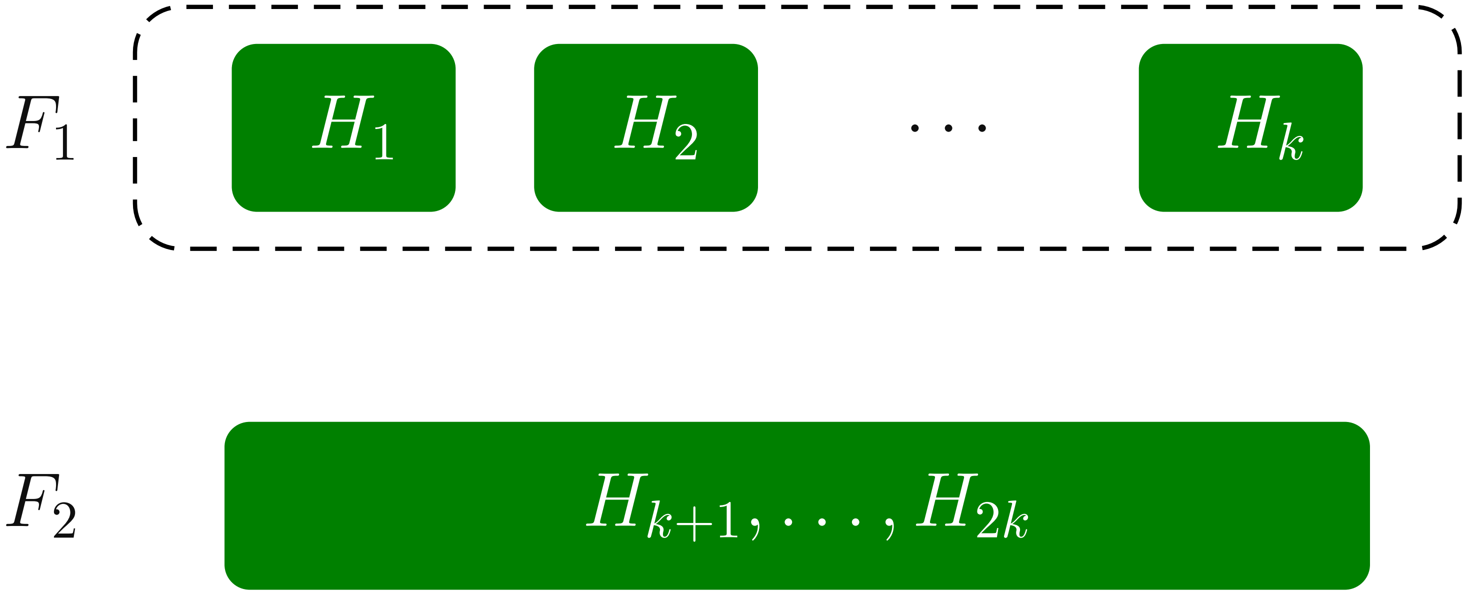
Recommended approaches

Simple chain-based procedures (Bretz et al., 2009)

Advanced procedures (Dmitrienko, Tamhane and Wiens, 2008)

Parallel gatekeeping procedure

Trickle-down principle (α propagation rule)

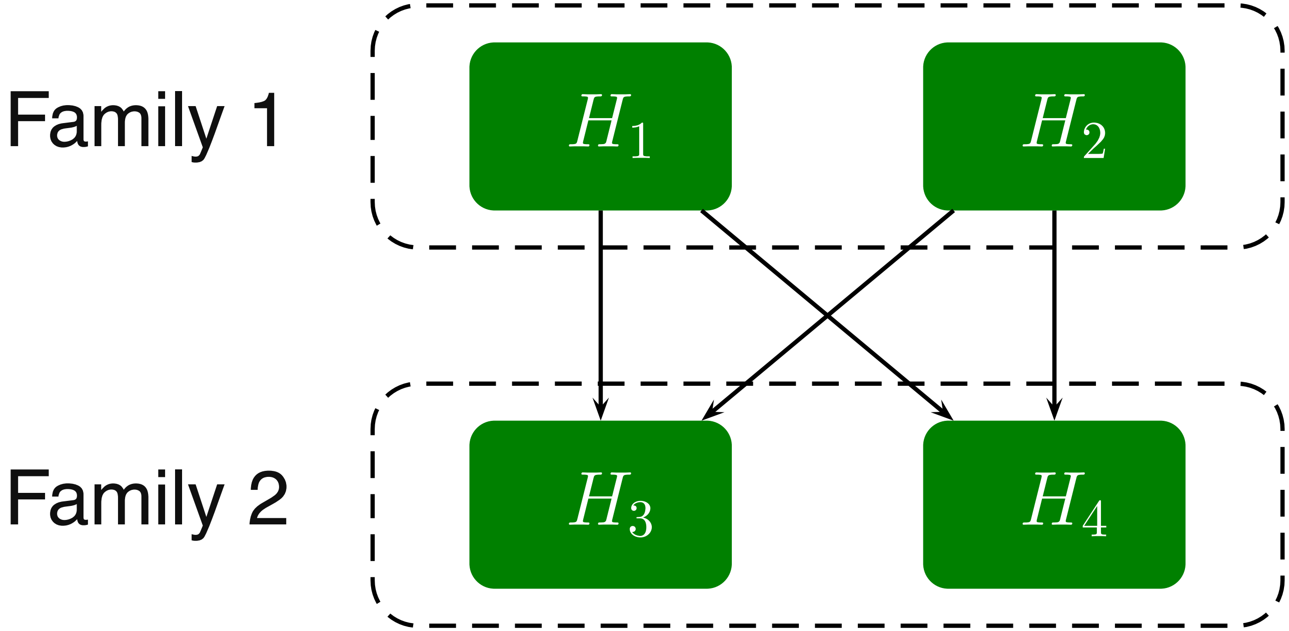


Procedure 1 is carried out in F_1 at α

Procedure 2 is carried out in F_2 at $\alpha_2 \leq \alpha$ (α_2 depends on the number of null hypotheses rejected in F_1)

Example 8: Schizophrenia trial

Decision tree



Family 1: Doses L and H versus Placebo in overall population

Family 2: Doses L and H versus Placebo in subpopulation

Simple parallel gatekeeping procedure

Chain-based procedure

Derived from nonparametric chain procedure

α **allocation rule**

Assign positive weights to primary hypotheses (H_1 and H_2)

Assign zero weights to secondary hypotheses (H_3 and H_4)

α **propagation rule**

Based on parallel gatekeeping relationships

Simple parallel gatekeeping procedure

α allocation rule

Hypothesis weights

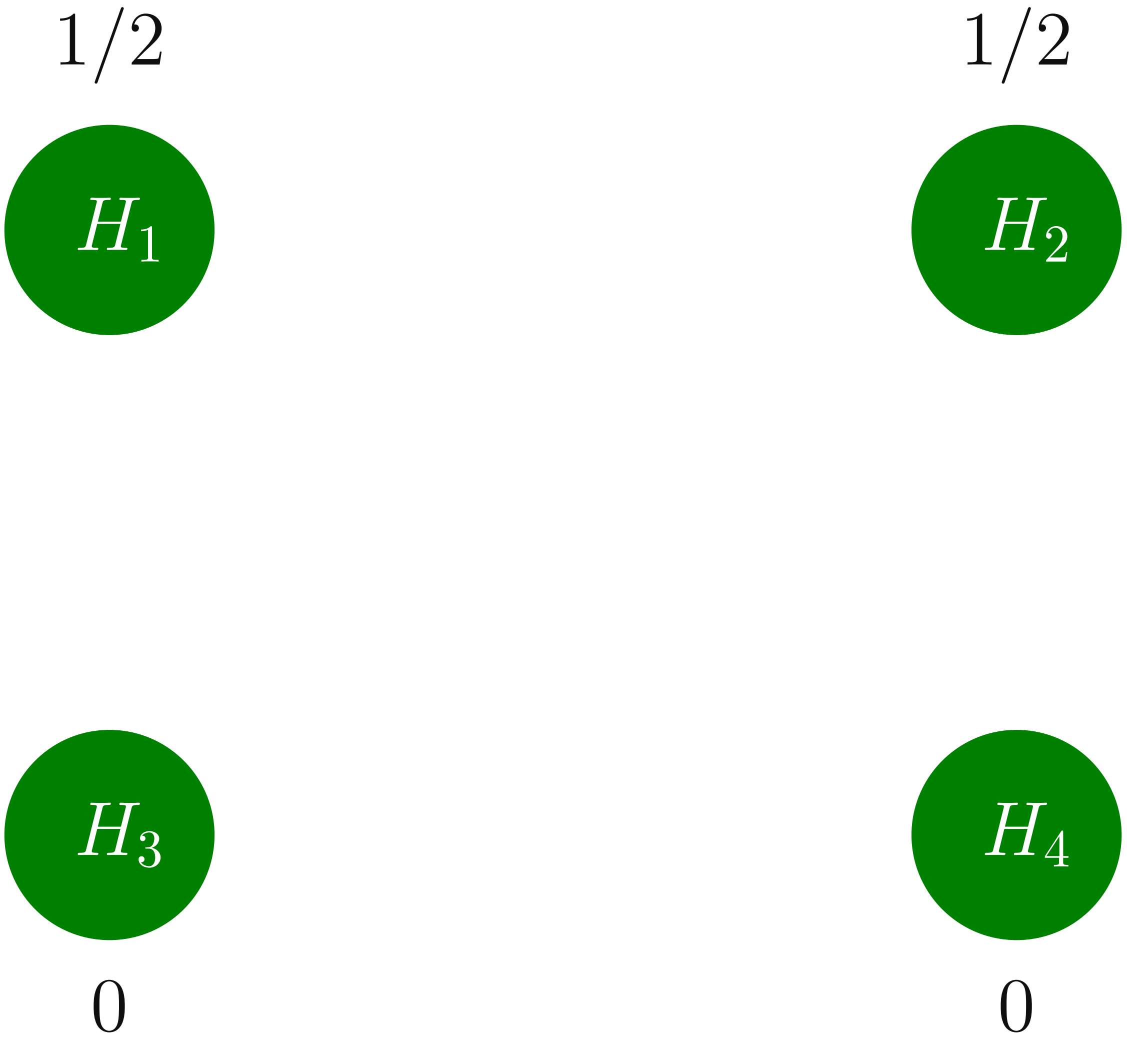
$$W = (1, 1, 0, 0)$$

α propagation rule

Transition parameters

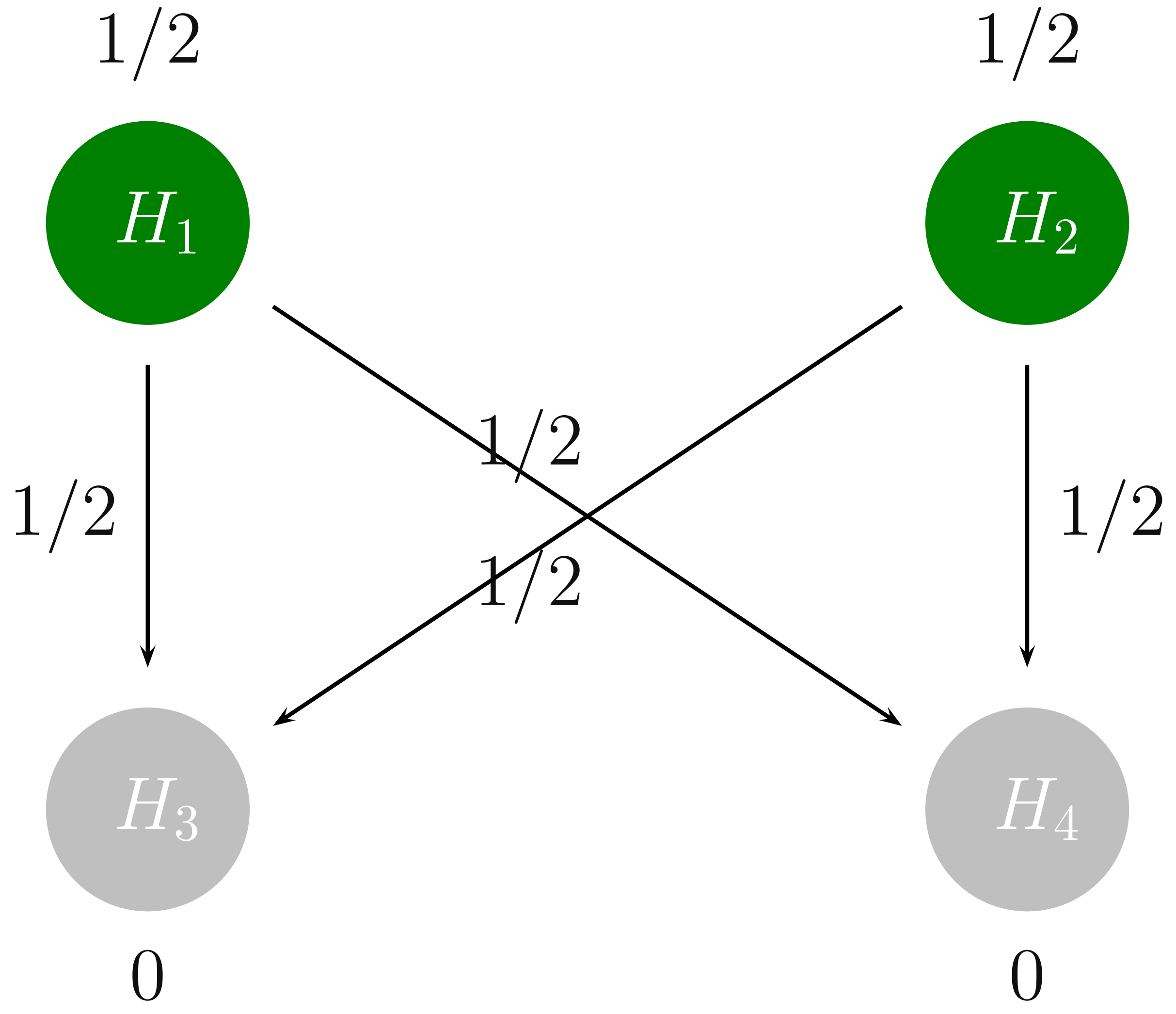
$$G = \begin{bmatrix} 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Simple parallel gatekeeping procedure



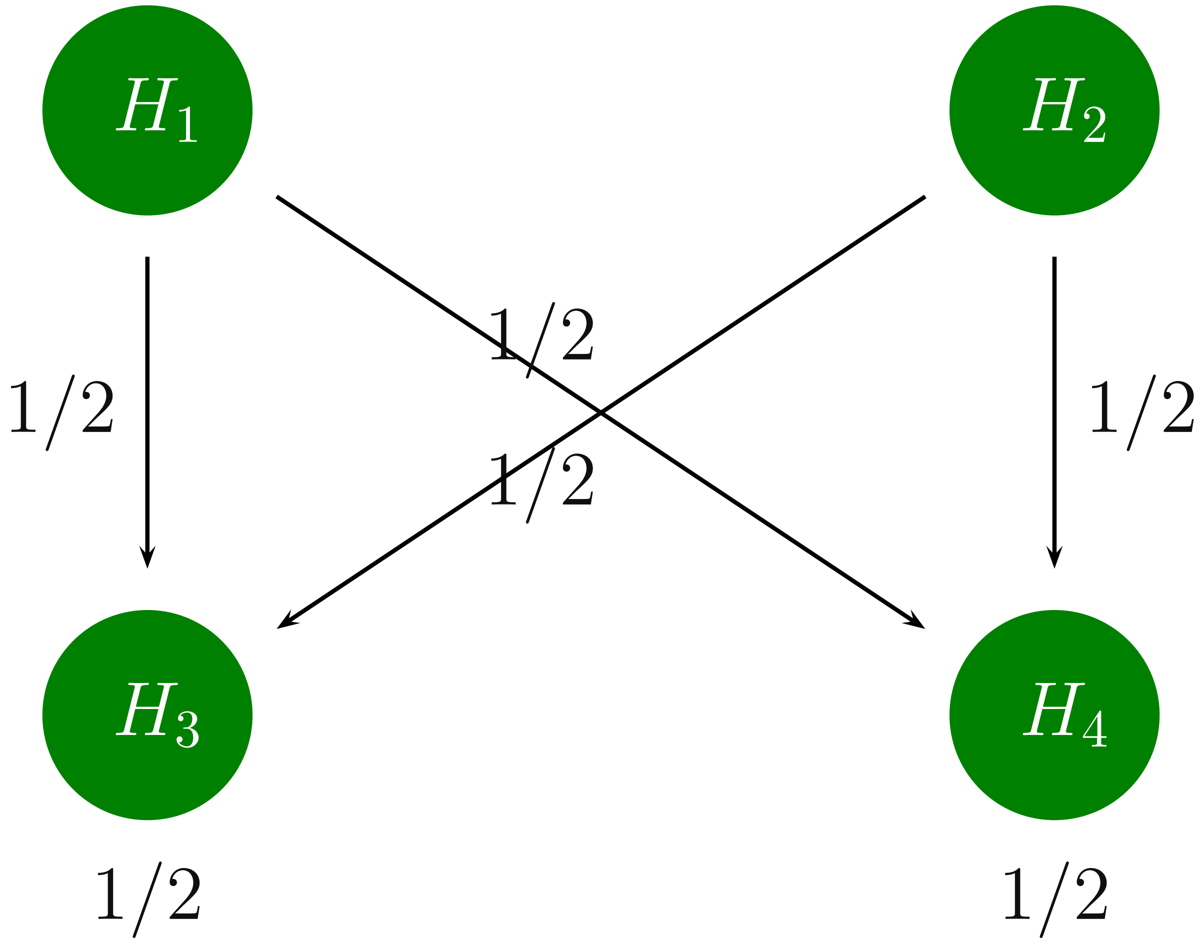
Non-negative weights are assigned to H_1 and H_2

Simple parallel gatekeeping procedure



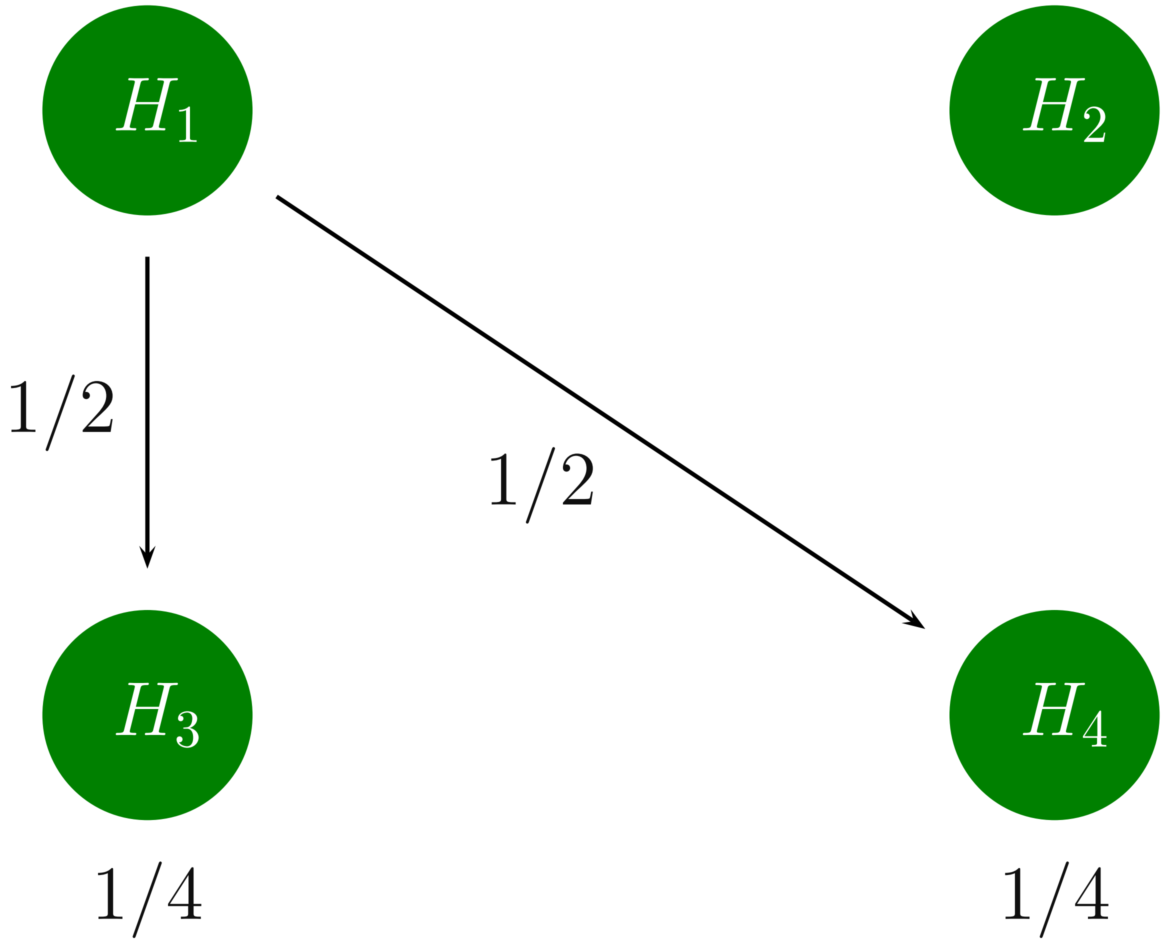
Step 1: Test H_1 and H_2 and update weights for H_3 and H_4

Simple parallel gatekeeping procedure



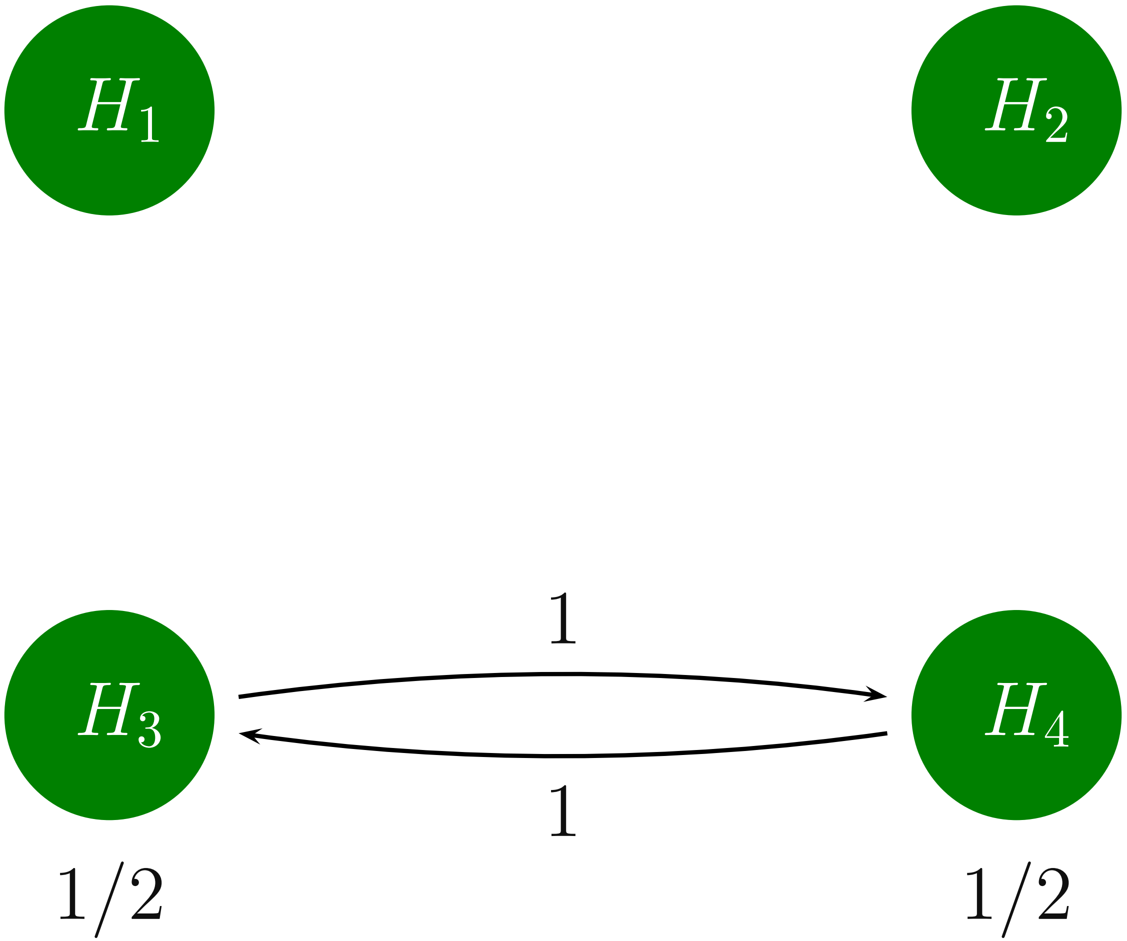
Step 1: H_1 and H_2 are both rejected

Simple parallel gatekeeping procedure



Step 1: Only H_1 is rejected

Simple parallel gatekeeping procedure



Step 2: Test H_3 and H_4 (H_1 and H_2 are both rejected)

Simple parallel gatekeeping procedures

Family 1

Component procedure: Bonferroni procedure

Family 2

Component procedure: Holm procedure

More powerful components

Advisable to use more powerful component procedures to improve power of parallel gatekeeping procedure (semiparametric or parametric components)

Advanced parallel gatekeeping procedures

Family 1

Component procedure: Semiparametric or parametric procedure

Family 2

Component procedure: Semiparametric or parametric procedure

Advanced parallel gatekeeping procedures

Family 1

$F_1 = \{H_1, \ldots, H_k\}$, null hypotheses

$N_1 = \{1, \ldots, k\}$, index set

Component procedure (Procedure 1): Separable procedure with local FWER control

Family 2

$F_2 = \{H_{k+1}, \ldots, H_{2k}\}$, null hypotheses

$N_2 = \{k+1, \ldots, 2k\}$, index set

Component procedure (Procedure 2): Any procedure with local FWER control

α propagation rule

Family 1

Procedure 1 at $\alpha_1 = \alpha$ level

$A_1 \subseteq N_1$, index set of null hypotheses accepted in F_1

Family 2

Procedure 2 at α_2 level

$\alpha_2 = \alpha_1 - e_1(A_1)$, where $e_1(I)$ is error rate function of Procedure 1

Global FWER control

Global FWER is protected at α

Error rate function

Definition

Assume that all null hypotheses H_i, $i \in I$, are true

Error rate function is probability of rejecting at least one true null hypothesis

$$e_1(I) = P\left\{ \bigcup_{i \in I} (\text{Reject } H_i) \,\middle|\, H_I \right\}, \quad I \subseteq N_1$$

Error rate function

Properties

$e_1(\emptyset) = 0$ (there are no null hypotheses to reject)

$e_1(N_1) = \alpha$ (error rate is α when all null hypotheses are true), may need to be enforced

Example

Error rate function of Bonferroni procedure is $e_1(I) = \alpha|I|/k$, where $|I|$ is number of elements in index set I

Properties of gatekeeping procedures

All null hypotheses are rejected in F_1

A_1 is empty

$\alpha_2 = \alpha_1 - e_1(\emptyset) = \alpha_1 - 0 = \alpha$

Null hypotheses in F_2 are tested at full α level

No null hypotheses are rejected in F_1

$A_1 = N_1$

$\alpha_2 = \alpha_1 - e_1(N_1) = \alpha_1 - \alpha_1 = 0$

Null hypotheses in F_2 are not tested

Separable procedures

Separability condition

Procedure 1 is separable if $e_1(I) < \alpha$ provided I is a proper subset of N_1

Implication

If a separable procedure is used in F_1, a fraction of α can be carried over to F_2 if one or more null hypotheses are rejected in F_1

Bonferroni procedure

Problem with three null hypotheses

Index set I	Error rate function $e_1(I)$
$\{1, 2, 3\}$	α
$\{1, 2\}$, $\{1, 3\}$, $\{2, 3\}$	$2\alpha/3$
$\{1\}$, $\{2\}$, $\{3\}$	$\alpha/3$
Empty	0

Error rate function of Bonferroni procedure is $e_1(I) = \alpha|I|/k$

Bonferroni procedure is separable because $e_1(I) < \alpha$ if I is a proper subset of N_1

Holm procedure

Problem with three null hypotheses

Index set I	Error rate function $e_1(I)$
$\{1,2,3\}$	α
$\{1,2\}$, $\{1,3\}$, $\{2,3\}$	α
$\{1\}$, $\{2\}$, $\{3\}$	α
Empty	0

Error rate function of Holm procedure is $e_1(I) = \alpha$ unless I is empty

Holm procedure is not separable

Separability condition

Separability procedures

Most popular procedures (Holm, fallback, Hochberg and Hommel procedures) do not satisfy the separability condition

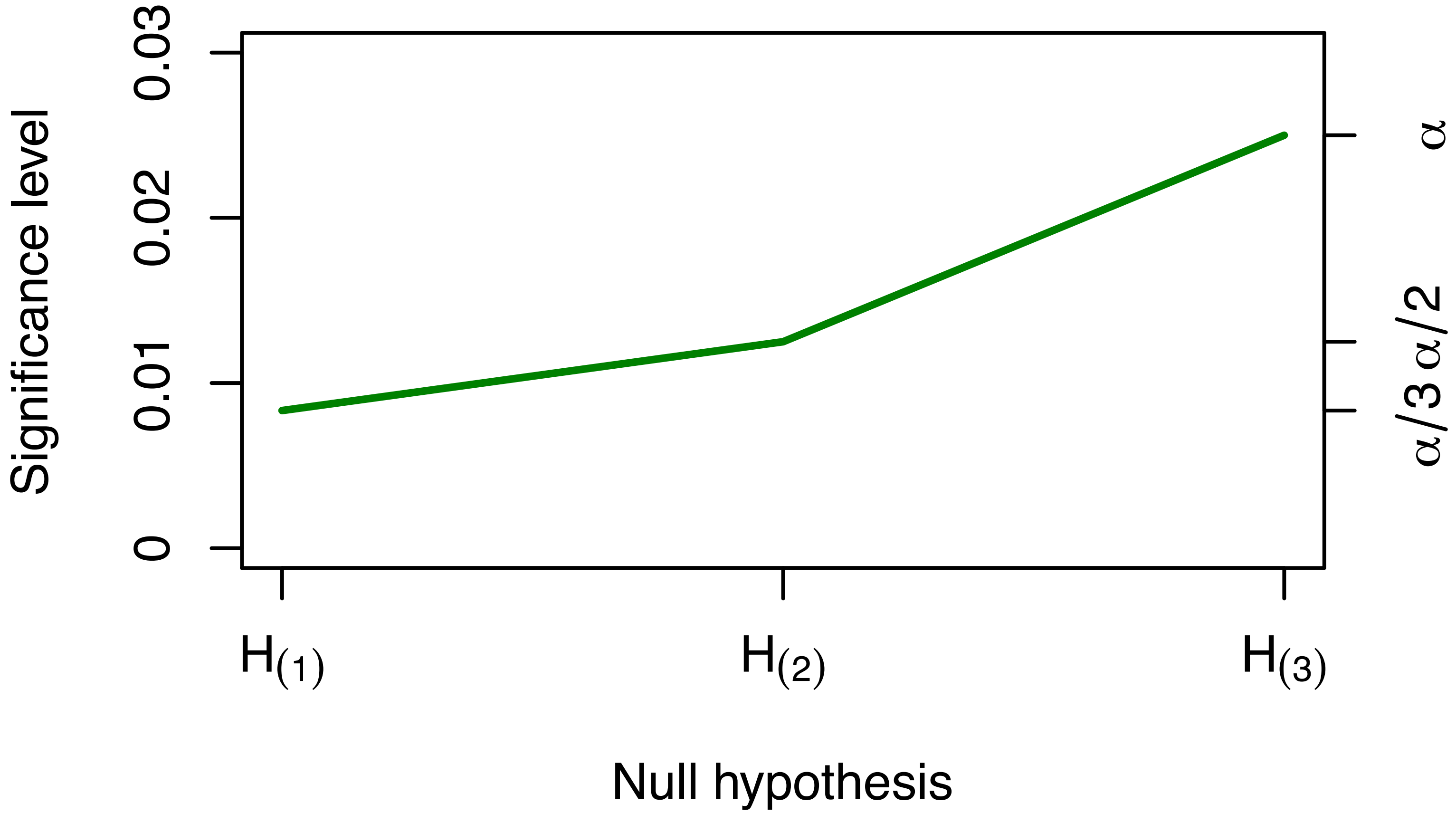
Truncated procedures

Truncated procedure is based on a convex combination between a multiple procedure and Bonferroni procedure

Truncated procedure is separable

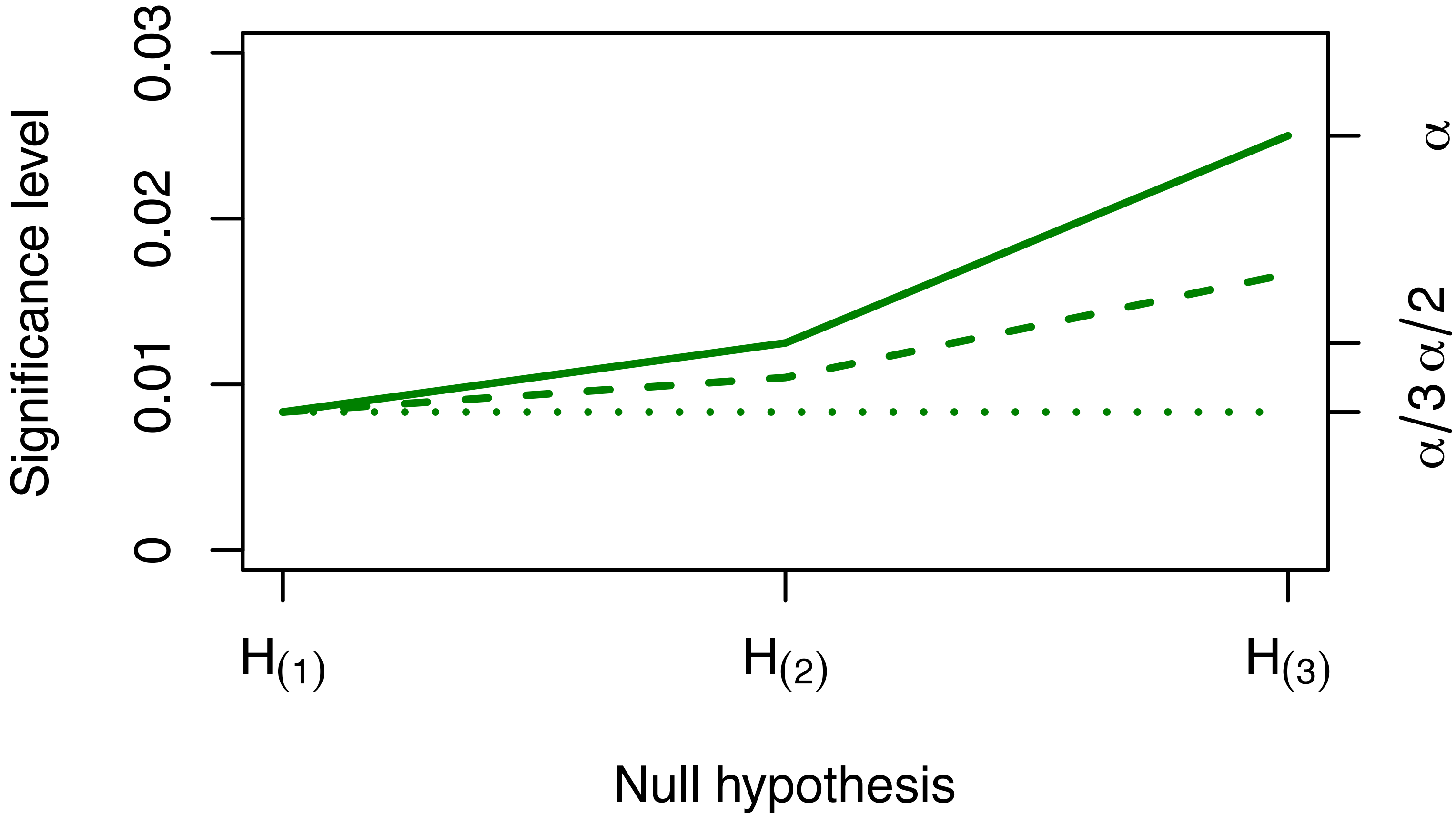
Regular Holm procedure

Problem with three null hypotheses ($\alpha = 0.025$)



Truncated Holm procedure

Problem with three null hypotheses ($\alpha = 0.025$)



—, Regular Holm procedure; - - -, Truncated Holm procedure; $\cdots$, Bonferroni procedure

Truncated Holm procedure

Significance levels for $H_{(1)}$, $H_{(2)}$ and $H_{(3)}$

Regular Holm procedure	Truncated Holm procedure	Bonferroni procedure
$\alpha/3$	$\gamma\alpha/3 + (1-\gamma)\alpha/3$	$\alpha/3$
$\alpha/2$	$\gamma\alpha/2 + (1-\gamma)\alpha/3$	$\alpha/3$
α	$\gamma\alpha + (1-\gamma)\alpha/3$	$\alpha/3$

$0 \leq \gamma \leq 1$, truncation parameter

Power of truncated Holm procedure is a monotonically increasing function of γ

Truncated Holm procedure simplifies to Bonferroni procedure if $\gamma = 0$ and regular Holm procedure if $\gamma = 1$

Truncated Holm procedure

Problem with three null hypotheses

Index set I	Error rate function $e_1(I)$
$\{1, 2, 3\}$	$[\gamma + 3(1-\gamma)/3]\alpha = \alpha$
$\{1, 2\}$, $\{1, 3\}$, $\{2, 3\}$	$[\gamma + 2(1-\gamma)/3]\alpha$
$\{1\}$, $\{2\}$, $\{3\}$	$[\gamma + (1-\gamma)/3]\alpha$
Empty	0

Error rate function for truncated Holm procedure is $e_1(I) = (\gamma + (1-\gamma)|I|/k)\alpha$ if $|I| > 0$ and $e_1(I) = 0$ if I is empty

Truncated Holm procedure is separable if $0 \leq \gamma < 1$

Truncated procedures

Other truncated procedures

Truncated p-value-based procedures: Truncated Holm, fallback and Hochberg procedures

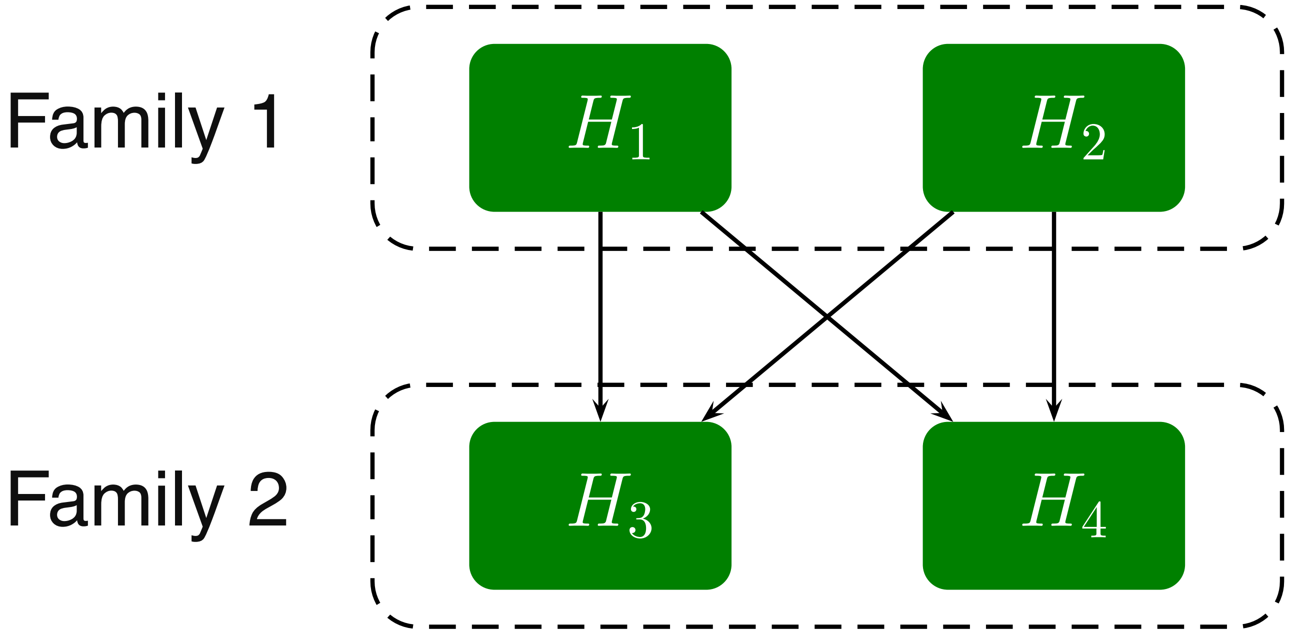
Truncated parametric procedures: Truncated step-down Dunnett procedure

Gatekeeping procedures

Wide variety of parallel gatekeeping procedures can be built based on these truncated procedures

Example 8: Schizophrenia trial

Decision tree



Family 1: Doses L and H versus Placebo in overall population

Family 2: Doses L and H versus Placebo in subpopulation

Advanced parallel gatekeeping procedure

Family 1

Component procedure: Semiparametric procedure (Truncated Hochberg)

Local FWER is controlled since test statistics are positively correlated

Family 2

Component procedure: Semiparametric procedure (Regular Hochberg)

Local FWER is controlled since test statistics are positively correlated

Advanced parallel gatekeeping procedure

Family 1

Truncated Hochberg procedure with $0 \leq \gamma < 1$ at $\alpha_1 = \alpha$

Family 2

Regular Hochberg procedure at $\alpha_2 = \alpha_1 - e_1(A_1)$

$e_1(I)$, error rate function of truncated Hochberg procedure

Advanced parallel gatekeeping procedure

Error rate function of truncated Hochberg procedure

Index set I	Error rate function $e_1(I)$
$\{1,2\}$	α
$\{1\}$, $\{2\}$	$(\gamma + (1-\gamma)/2)\alpha$
Empty	0

α propagation rule

No null hypotheses are rejected in Family 1

$A_1 = \{1, 2\}$

$\alpha_2 = \alpha - e_1(A_1) = \alpha - \alpha = 0$

One null hypothesis is rejected in Family 1

$A_1 = \{1\}$ or $\{2\}$

$\alpha_2 = \alpha - e_1(A_1) = \alpha - (\gamma + (1 - \gamma)/2)\alpha = (1 - \gamma)\alpha/2$

Two null hypotheses are rejected in Family 1

A_1 is empty

$\alpha_2 = \alpha - e_1(A_1) = \alpha - 0 = \alpha$

Example 8: Schizophrenia trial

Four hypothesis tests

Family	Hypothesis	Raw p-value
Family 1	H_1	0.0082
	H_2	0.0174
Family 2	H_3	0.0042
	H_4	0.0180

$\alpha = 0.025$, Global familywise error rate

Example 8: Schizophrenia trial

Scenario 1 (simple gatekeeping procedure)

Bonferroni procedure (truncated Hochberg procedure with $\gamma = 0$) in Family 1

Holm procedure in Family 2

Scenario 2 (advanced gatekeeping procedure)

Truncated Hochberg procedure with $\gamma = 0.5$ in Family 1

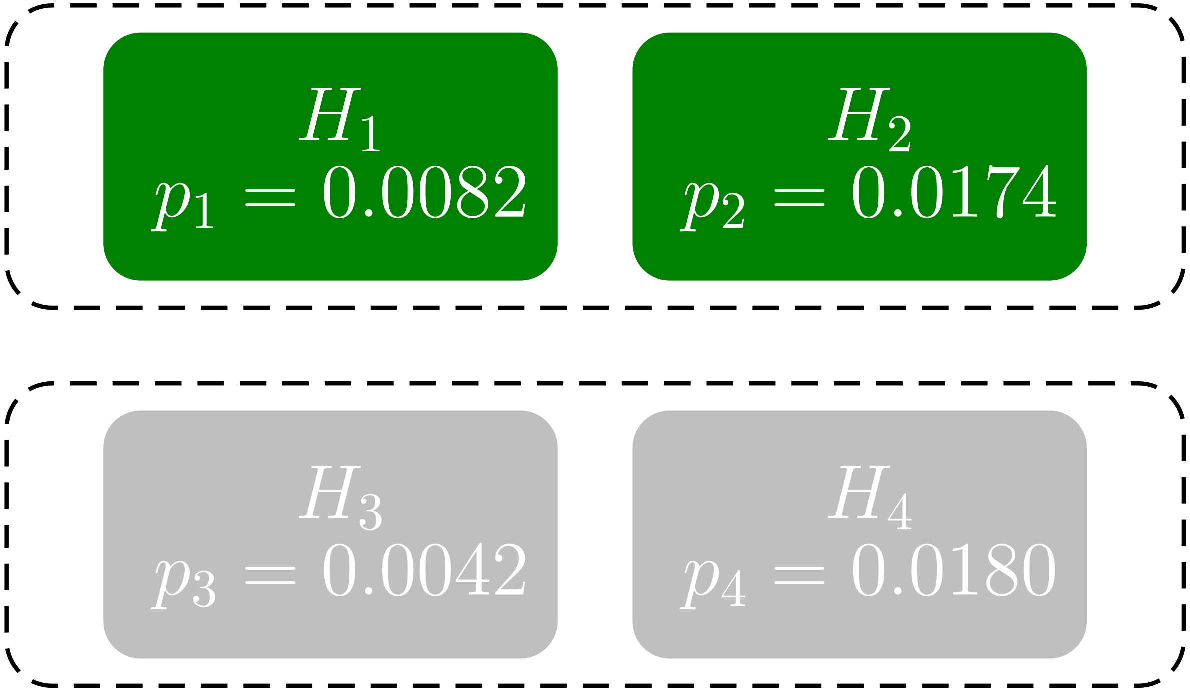
Hochberg procedure in Family 2

Scenario 1 (simple gatekeeping procedure)

Procedure 1 (Bonferroni procedure)

Procedure 1
at $\alpha_1 = 0.025$

Procedure 2



H_1
$p_1 = 0.0082$

H_2
$p_2 = 0.0174$

H_3
$p_3 = 0.0042$

H_4
$p_4 = 0.0180$

Test H_1 and H_2 at $\alpha_1/2 = 0.0125$

H_1 is rejected and H_2 is accepted

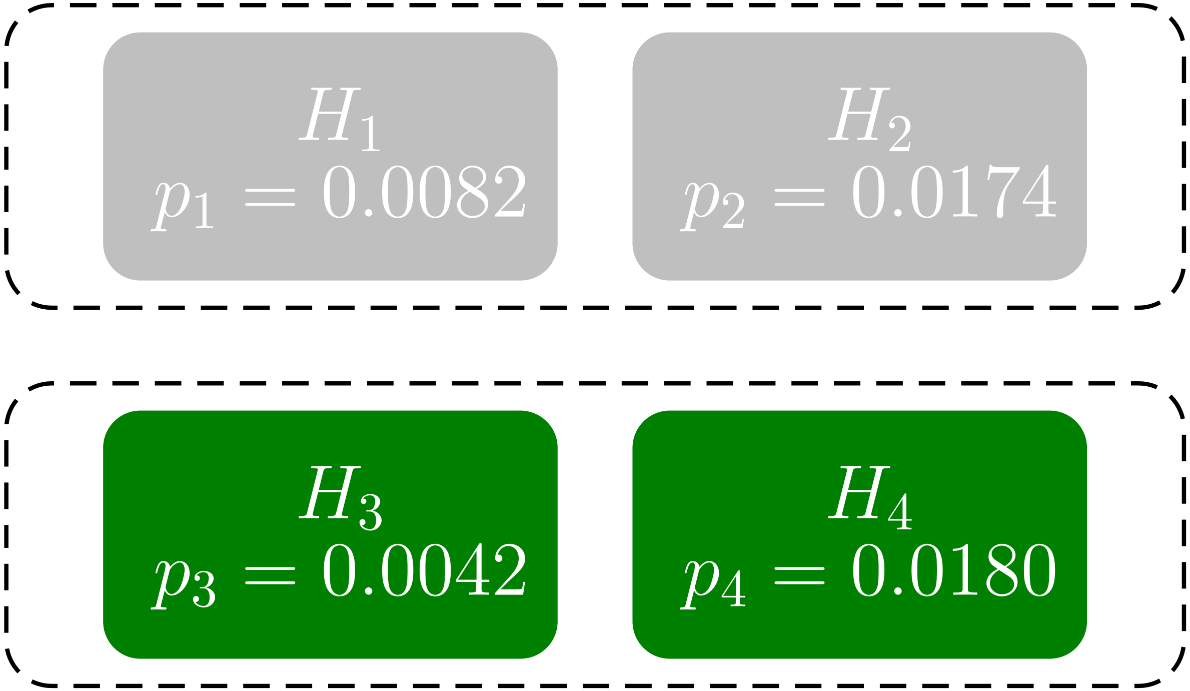
$\alpha_2 = \alpha_1/2 = 0.0125$ is carried over to Family 2

Scenario 1 (simple gatekeeping procedure)

Procedure 2 (Holm procedure)

Procedure 1

Procedure 2
at $\alpha_2 = 0.0125$



H_1
$p_1 = 0.0082$

H_2
$p_2 = 0.0174$

H_3
$p_3 = 0.0042$

H_4
$p_4 = 0.0180$

Test H_3 at $\alpha_2/2 = 0.00625$

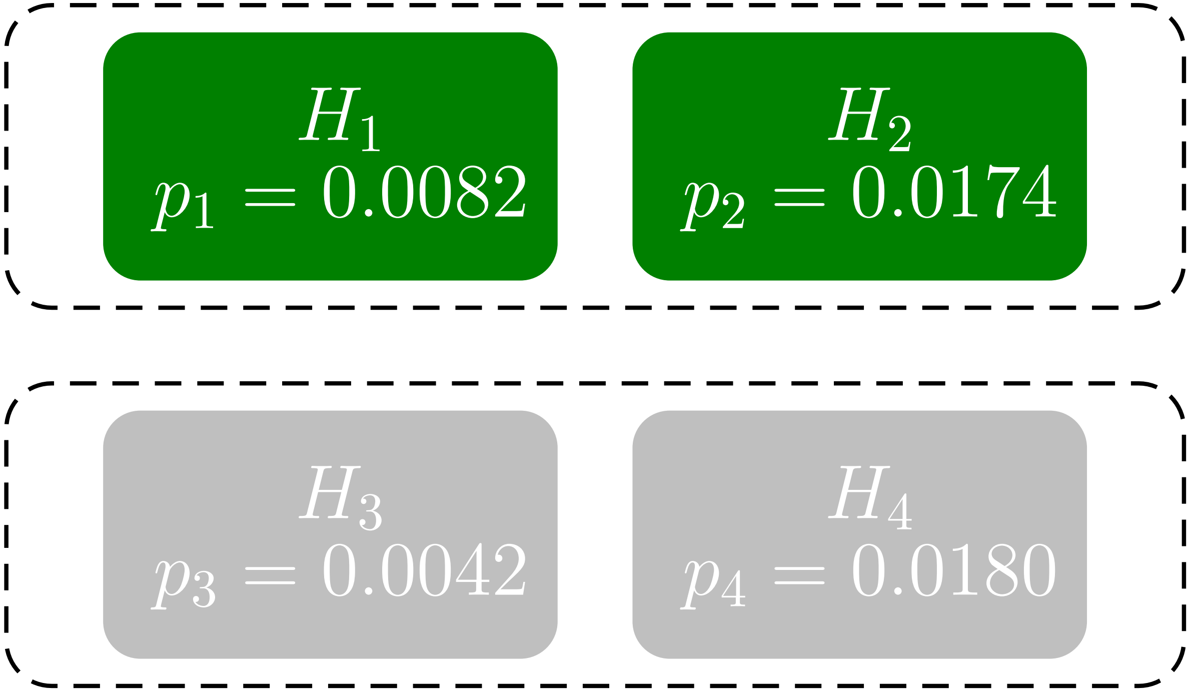
Test H_4 at $\alpha_2 = 0.0125$ if H_3 is rejected

H_3 is rejected and H_4 is accepted

Scenario 2 (advanced gatekeeping procedure)

Procedure 1 (truncated Hochberg procedure)

Procedure 1
at $\alpha_1 = 0.025$



H_1
$p_1 = 0.0082$

H_2
$p_2 = 0.0174$

Procedure 2

H_3
$p_3 = 0.0042$

H_4
$p_4 = 0.0180$

Test H_1 and H_2 at $\gamma\alpha_1 + (1 - \gamma)\alpha_1/2 = 0.01875$

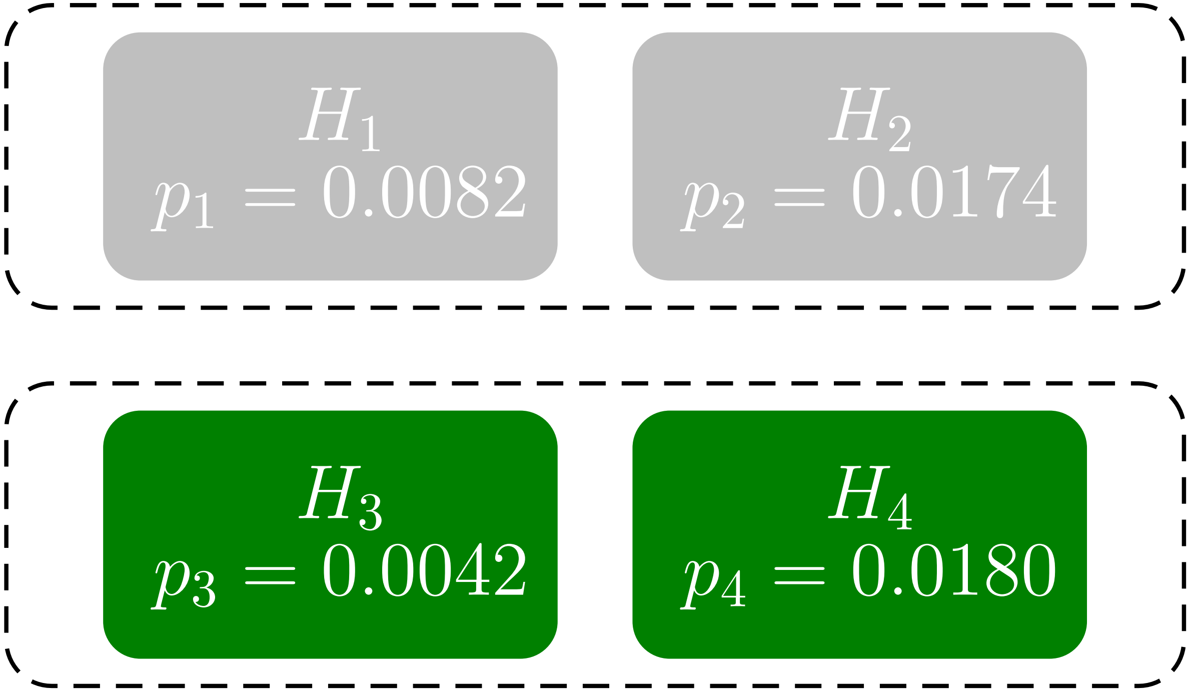
H_1 and H_2 are rejected

$\alpha_2 = \alpha_1 = 0.025$ is carried over to Family 2

Scenario 2 (advanced gatekeeping procedure)

Procedure 2 (Hochberg procedure)

Procedure 1



H_1

$p_1 = 0.0082$

H_2

$p_2 = 0.0174$

Procedure 2
at $\alpha_2 = 0.025$

H_3

$p_3 = 0.0042$

H_4

$p_4 = 0.0180$

Test H_3 and H_4 at $\alpha_2 = 0.025$

H_3 and H_4 are rejected

Role of truncation parameter γ

Example 8: Schizophrenia trial

Simple procedure ($\gamma = 0$): H_1 and H_3 are rejected

Advanced procedure ($\gamma = 0.5$): All null hypotheses are rejected

Greater value of γ improves power of Procedures 1 and 2 in this example

Selection of truncation parameter γ

In general, optimal value of γ depends on effect sizes in Families 1 and 2

Example 8: Schizophrenia trial

Adjusted p-values as a function of γ

Hypothesis	Adjusted p-values		
	$\gamma = 0$	$\gamma = 0.25$	$\gamma = 0.5$
H_1	0.0164	0.0164	0.0164
H_2	0.0348	0.0278	0.0232
H_3	0.0168	0.0224	0.0232
H_4	0.0348	0.0278	0.0232

Number of null hypotheses rejected at $\alpha = 0.025$ increases with increasing truncation parameter γ in Family 1

Optimality considerations

Selection of truncation parameter γ

Select γ based on a suitable optimality criterion

Optimality criteria

Exceedence criterion

Expectation criterion

Optimality criteria

Notation

R_i, Event corresponding to rejection of null hypothesis H_i, $i = 1, 2, 3, 4$, i.e., R_i occurs when H_i is rejected

A_i, Complement of R_i, i.e., A_i occurs when H_i is accepted

Exceedence criterion

Power function

Probability of rejecting at least one null hypothesis in Family 1 and at least one null hypothesis in Family 2

Power function $= P([R_1 \text{ or } R_2] \text{ and } [R_3 \text{ or } R_4])$

Properties

Exceedence criterion lacks sensitivity

It does not differentiate between rejection of one or two primary null hypotheses

Expectation criterion

Idea

Partition into three mutually exclusive events

$$
\begin{aligned}
&[R_1 \text{ or } R_2] \text{ and } [R_3 \text{ or } R_4] \\
&\quad = R_1 \text{ and } R_2 \text{ and } [R_3 \text{ or } R_4] \\
&\quad + R_1 \text{ and } A_2 \text{ and } [R_3 \text{ or } R_4] \\
&\quad + A_1 \text{ and } R_2 \text{ and } [R_3 \text{ or } R_4]
\end{aligned}
$$

Weights

w_1, w_2, w_3, relative importance of three events ($w_1 + w_2 + w_3 = 1$)

Expectation criterion

Power function

Power function

$$= w_1 P(R_1 \text{ and } R_2 \text{ and } [R_3 \text{ or } R_4])$$
$$+ w_2 P(R_1 \text{ and } A_2 \text{ and } [R_3 \text{ or } R_4])$$
$$+ w_3 P(A_1 \text{ and } R_2 \text{ and } [R_3 \text{ or } R_4])$$

Properties

Expectation criterion is more sensitive

It accounts for relative importance of rejecting one or two primary null hypotheses

Example 8: Schizophrenia trial

Notation

$t_1, \ldots, t_4$, Test statistics for four null hypotheses (follow a four-dimensional normal distribution)

$e_1, \ldots, e_4$, Effect sizes for four null hypotheses (mean difference/common standard deviation)

n_0, Sample size per treatment group in general population

n_+, Sample size per treatment group in subpopulation

Example 8: Schizophrenia trial

Distribution of test statistics

Means

Dose-placebo comparisons in general population ($i = 1, 2$)

$$\mu_i = E(t_i) = e_i/\sqrt{2/n_0}$$

Dose-placebo comparisons in subpopulation ($i = 3, 4$)

$$\mu_i = E(t_i) = e_i/\sqrt{2/n_+}$$

Example 8: Schizophrenia trial

Covariance matrix

$$\Sigma = \begin{bmatrix} 1 & 1/2 & \rho & \rho/2 \\ 1/2 & 1 & \rho/2 & \rho \\ \rho & \rho/2 & 1 & 1/2 \\ \rho/2 & \rho & 1/2 & 1 \end{bmatrix}$$

where $\rho = \sqrt{n_+/n_0}$

Example 8: Schizophrenia trial

Assumptions

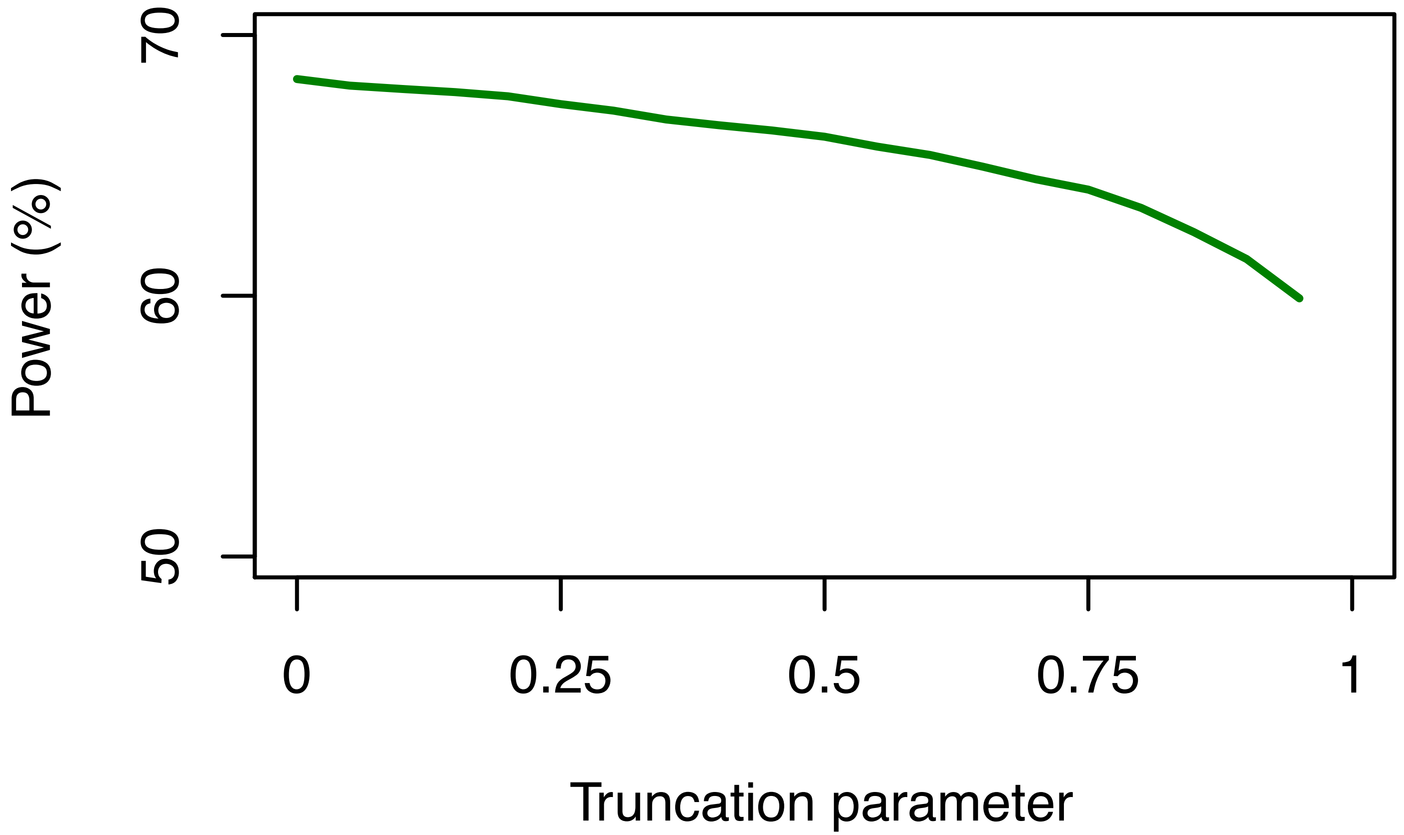
Hypothesis	Sample size	Effect size
H_1	$n_0 = 300$	$e_1 = 0.25$
H_2	$n_0 = 300$	$e_2 = 0.20$
H_3	$n_+ = 100$	$e_3 = 0.40$
H_4	$n_+ = 100$	$e_4 = 0.30$

Correlation, $\rho = 0.577$

Weights in expectation criterion, $w_1 = 0.6$, $w_2 = 0.25$ and $w_3 = 0.15$

Example 8: Schizophrenia trial

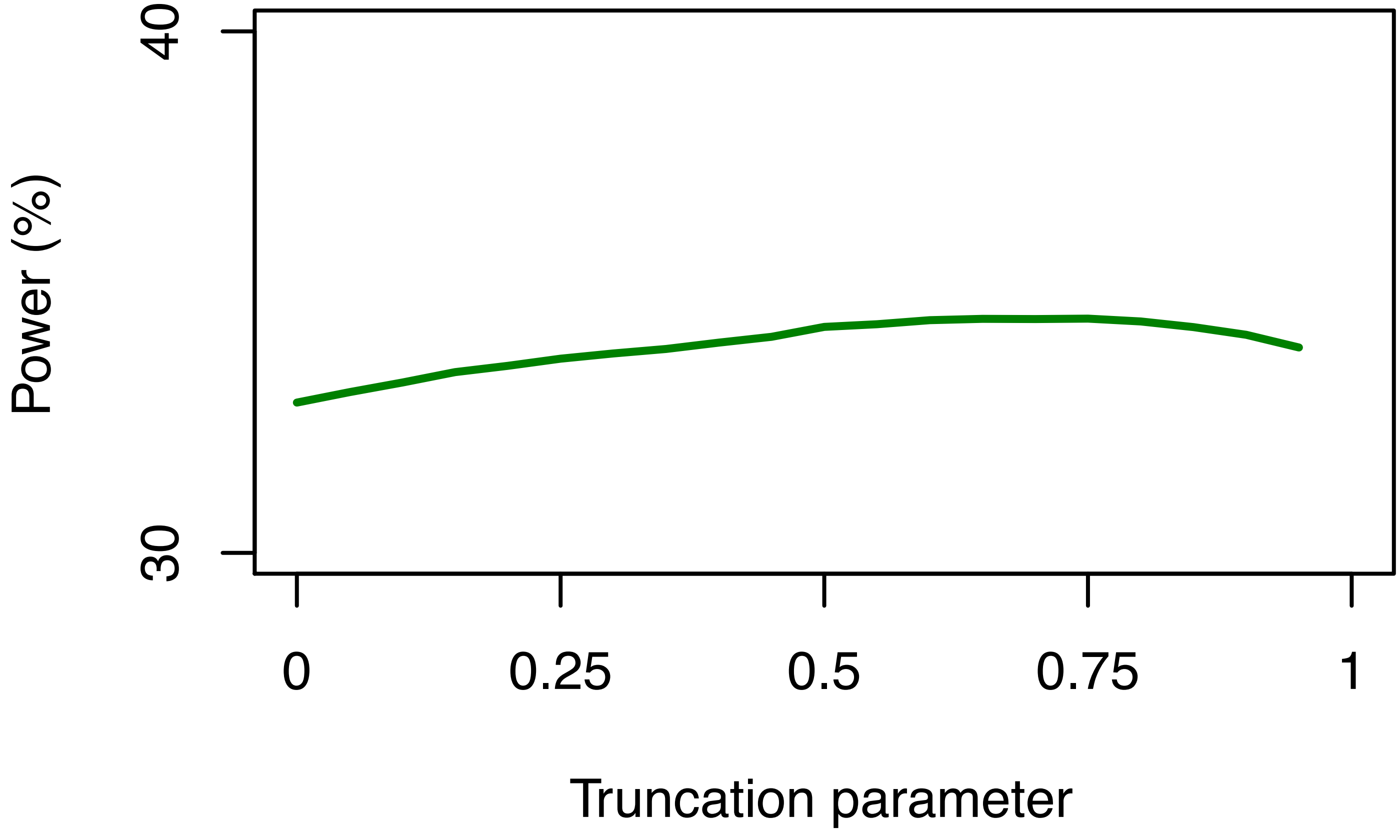
Exceedence criterion



$P(R_1$ or $R_2)$ is virtually independent of γ while $P(R_3$ or $R_4)$ decreases with γ

Example 8: Schizophrenia trial

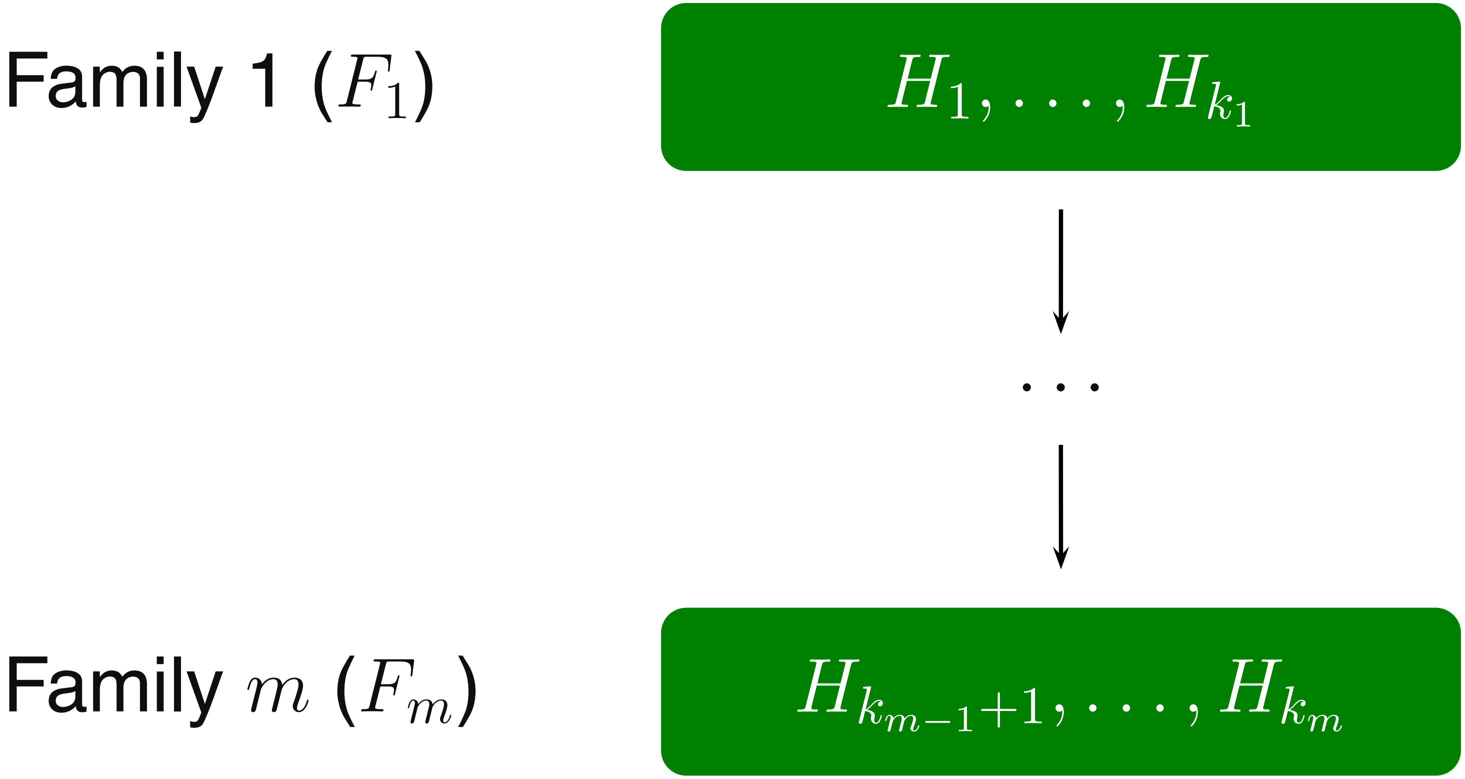
Expectation criterion



Larger values of truncation parameter γ are recommended

General setting

Multiple families of null hypotheses



Family 1 (F_1)

$H_1, \ldots, H_{k_1}$

$\ldots$

Family m (F_m)

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

Advanced parallel gatekeeping procedures

Family 1

Separable FWER-controlling procedure at $\alpha_1 = \alpha$

Family i **(**$i = 2, \ldots, m-1$**)**

Separable FWER-controlling procedure at $\alpha_i = \alpha_{i-1} - e_{i-1}(A_{i-1})$

A_{i-1}, index set of null hypotheses accepted in F_{i-1}

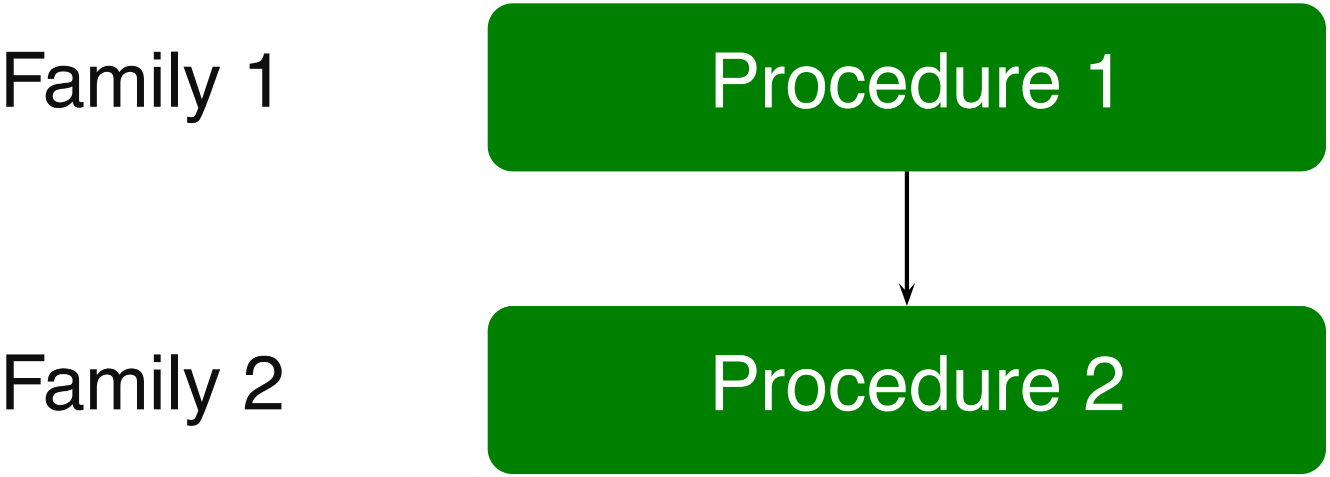
Family m

Any FWER-controlling procedure at $\alpha_m = \alpha_{m-1} - e_{m-1}(A_{m-1})$

3. Sequential testing with re-testing

Parallel gatekeeping procedures

Sequential testing



Family 1

Procedure 1

Family 2

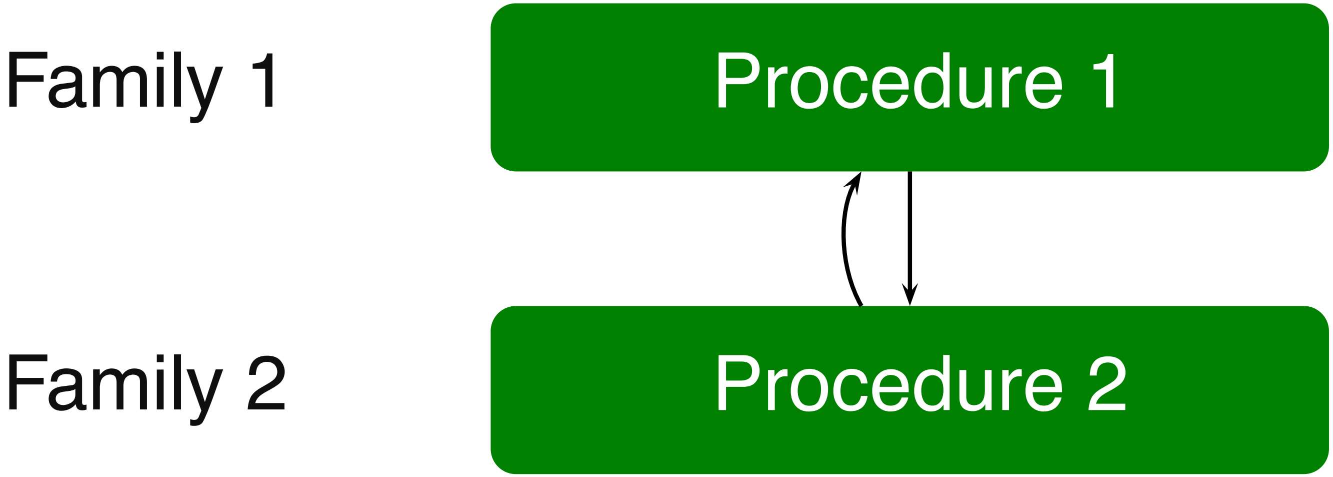
Procedure 2

Step 1: Family 1 is tested

Step 2: If at least one hypothesis is rejected in Family 1, Family 2 is tested

Parallel gatekeeping procedures

Sequential testing with re-testing

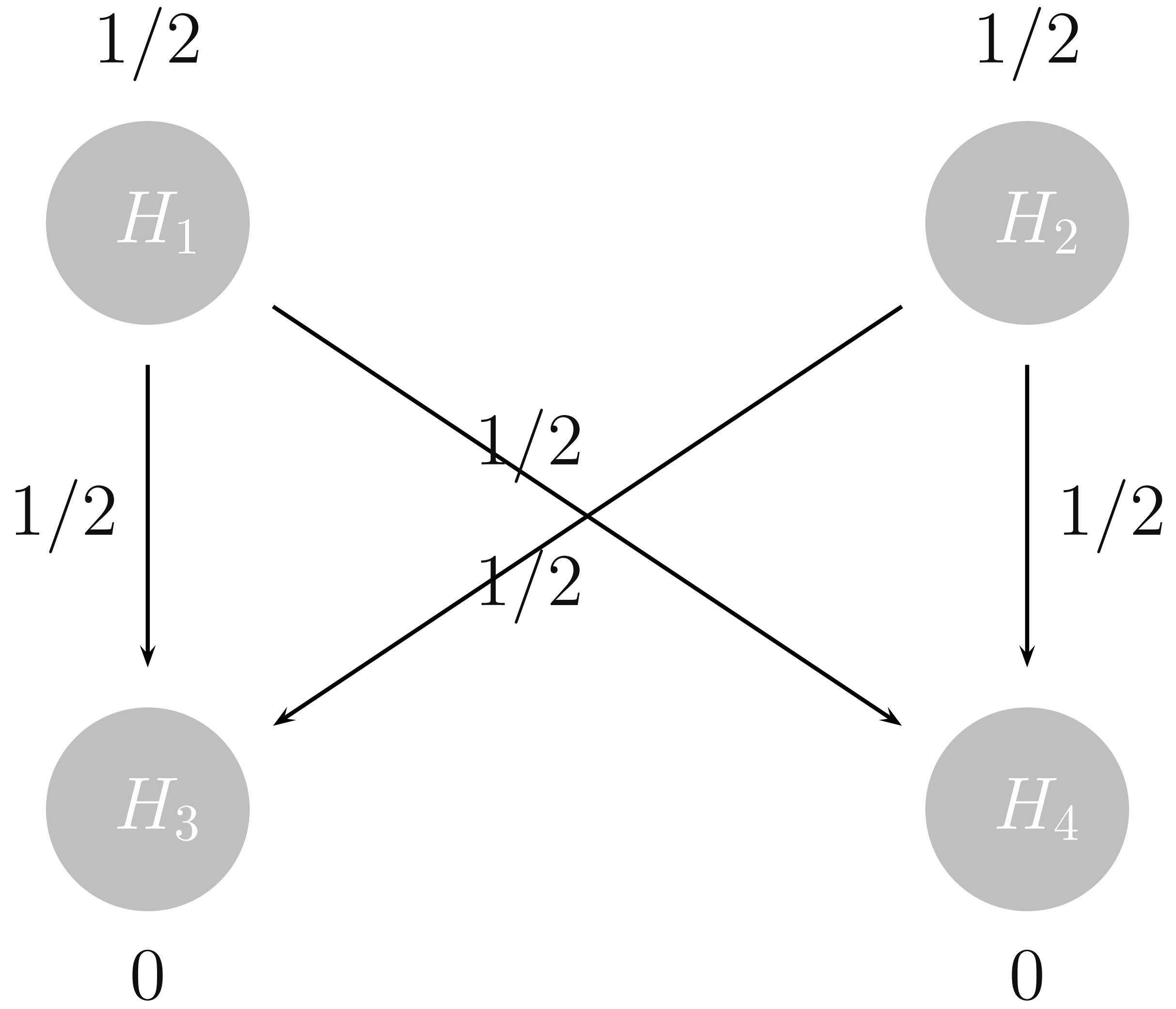


Step 1: Family 1 is tested

Step 2: If at least one hypothesis is rejected in Family 1, Family 2 is tested

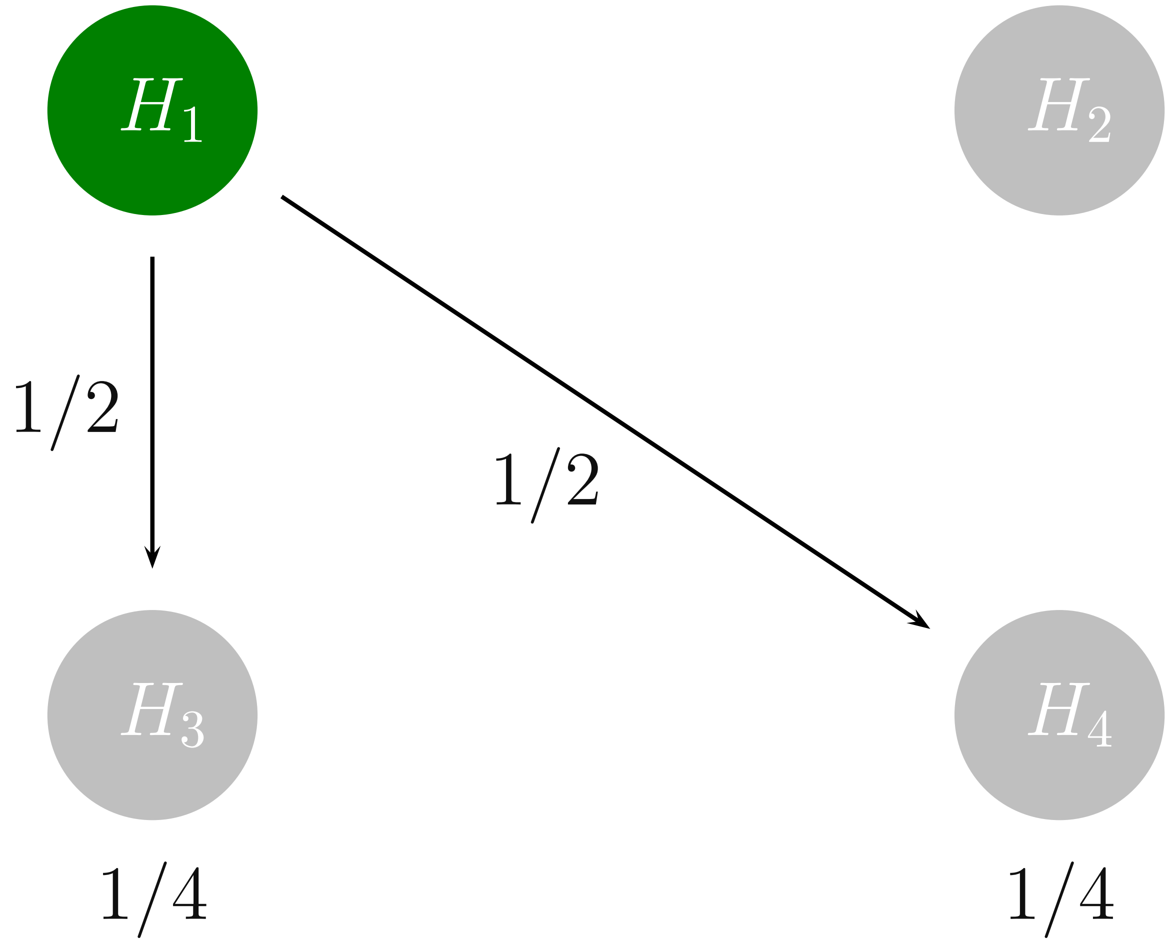
Step 3: If at least one hypothesis is rejected in Family 2 and there are non-rejected hypotheses in Family 1, Family 1 is re-tested

Simple parallel gatekeeping procedure



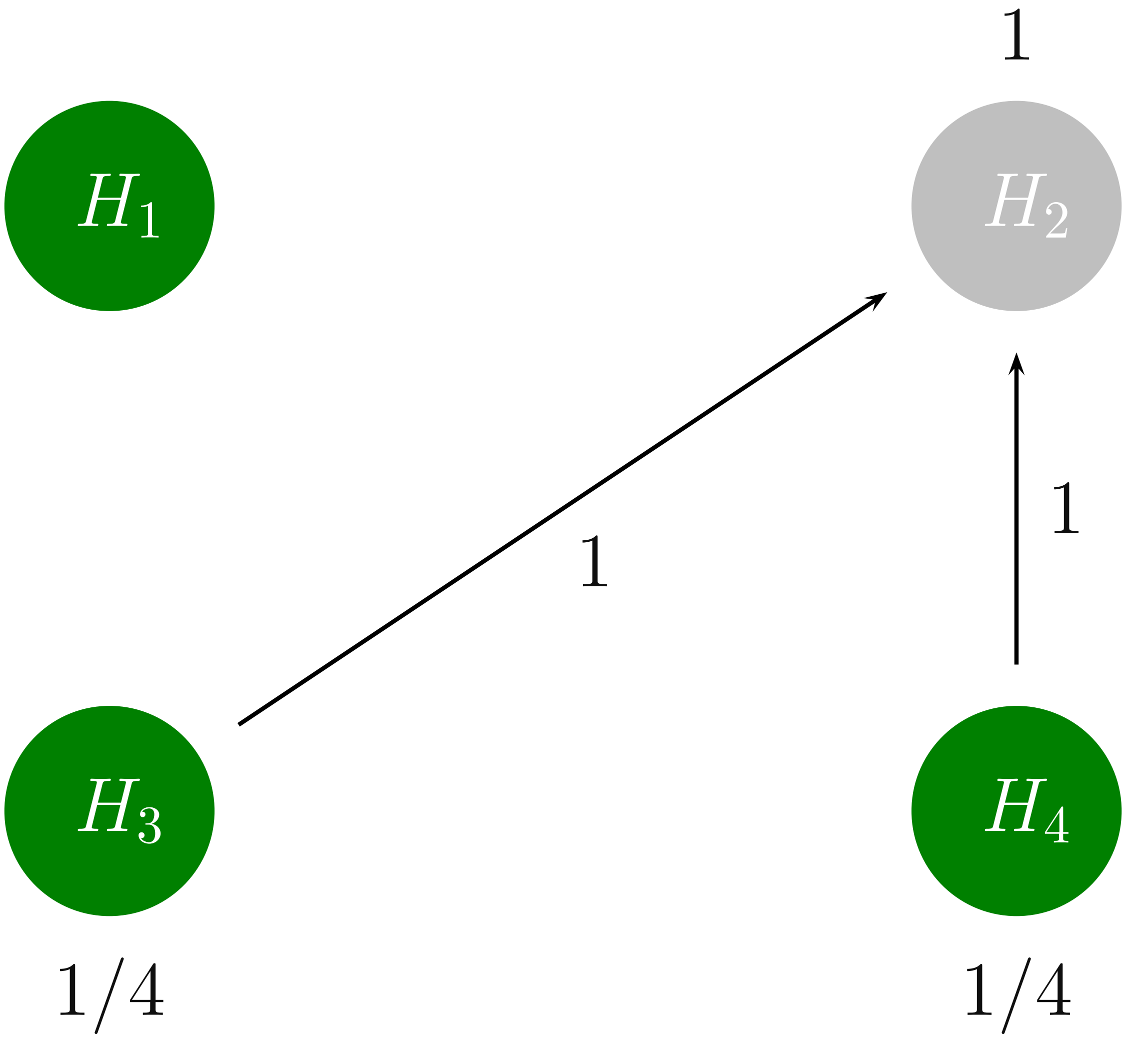
Step 1: Test H_1 and H_2 and update weights for H_3 and H_4

Simple parallel gatekeeping procedure



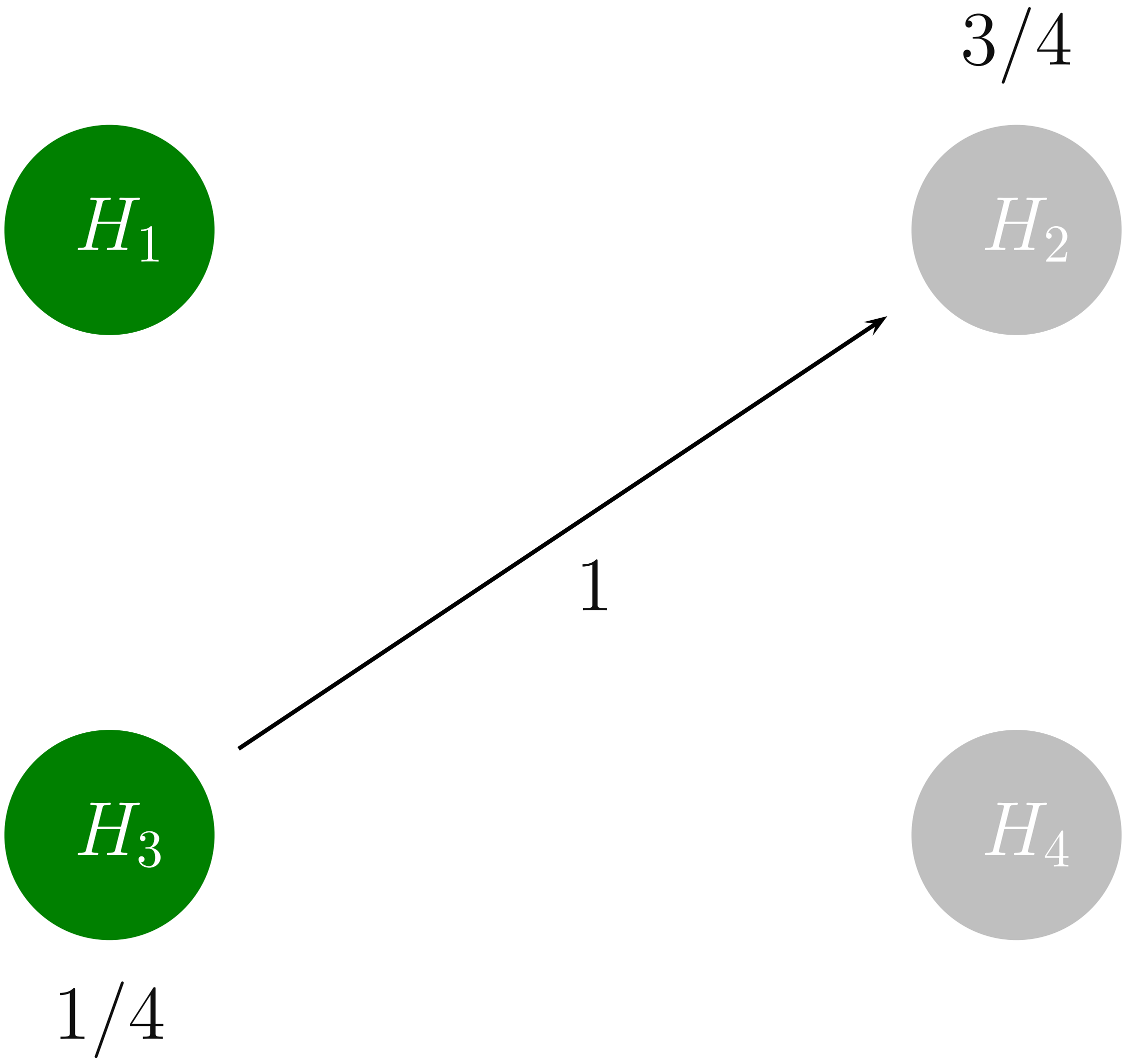
Step 1: Suppose that H_1 is rejected and H_2 is accepted

Simple parallel gatekeeping procedure



Steps 2 and 3: Suppose that H_3 and H_4 are both rejected

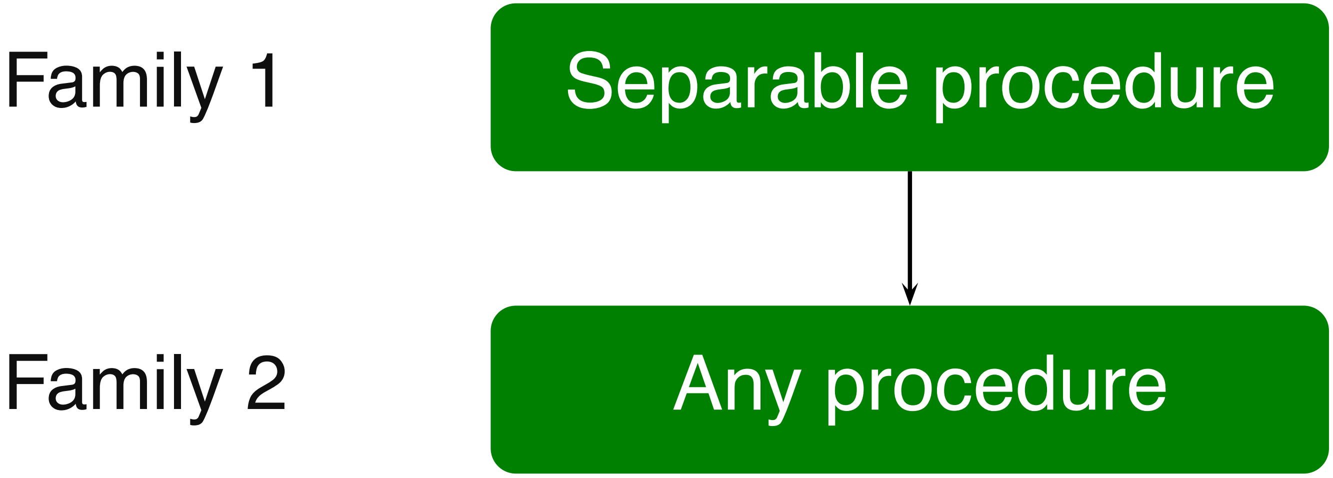
Simple parallel gatekeeping procedure



Steps 2 and 3: Suppose that only H_3 is rejected

Advanced parallel gatekeeping procedures

Sequential testing



Step1: Family 1 is tested using a separable procedure

Step2: If at least one hypothesis is rejected in Family 1, Family 2 is tested using any procedure

Hochberg-based gatekeeping procedure

Step 1: Family 1

Truncated Hochberg procedure with $0 \leq \gamma < 1$ at $\alpha_1 = \alpha$

Step 2: Family 2

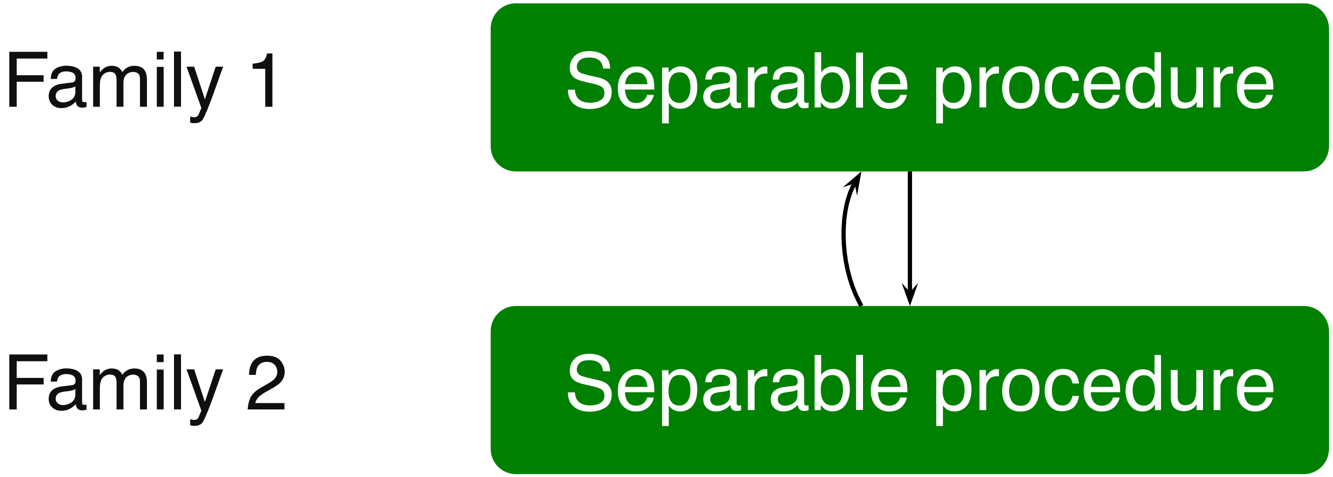
Regular Hochberg procedure at α_2

$\alpha_2 = \alpha$ if both hypotheses are rejected in Family 1

$\alpha_2 = (1 - \gamma)\alpha/2$ if only one hypothesis is rejected in Family 1

General parallel gatekeeping procedures

Sequential testing with re-testing



Family 1

Separable procedure

Family 2

Separable procedure

Step1: Family 1 is tested using a separable procedure

Step 2: If at least one hypothesis is rejected in Family 1, Family 2 is tested using a separable procedure

Step 3: If at least one hypothesis is rejected in Family 2 and there are non-rejected hypotheses in Family 1, Family 1 is re-tested

Hochberg-based gatekeeping procedure

Step 1: Family 1

Truncated Hochberg procedure with $0 \leq \gamma_1 < 1$ at $\alpha_1 = \alpha$

Step 2: Family 2

Truncated Hochberg procedure with $0 \leq \gamma_2 < 1$ at α_2

$\alpha_2 = \alpha$ if both hypotheses are rejected in Family 1

$\alpha_2 = (1 - \gamma_1)\alpha/2$ if only one hypothesis is rejected in Family 1

Hochberg-based gatekeeping procedure

Step 3: Family 1

Truncated Hochberg procedure with γ_3 at $\alpha_3 = \alpha$

$\gamma_3 = 1$ if both hypotheses are rejected in Family 2

$$\gamma_3 = \gamma_1 + \frac{(1-\gamma_1)(1-\gamma_2)}{2}$$

if only one hypothesis is rejected in Family 2

Hochberg-based gatekeeping procedure

Step 2: Family 2

The more hypotheses are rejected in Family 1, the higher the significance level

α pumping: α is increased and truncation parameter is constant

Step 3: Family 1

The more hypotheses are rejected in Family 2, the higher the truncation parameter

Power pumping: α is constant and truncation parameter is increased

Example

Step 1: Family 1

Truncated Hochberg procedure with $\gamma_1 = 0.5$ at $\alpha_1 = 0.025$

Suppose that only H_1 is rejected

Step 2: Family 2

Truncated Hochberg procedure with $\gamma_2 = 0.5$ at $\alpha_2 = (1 - \gamma_1)\alpha/2 = 0.00625$

Suppose that only H_3 is rejected

Example

Step 3: Family 1

Truncated Hochberg procedure with $\gamma_3 = \gamma_1 + (1 - \gamma_1)(1 - \gamma_2)/2 = 0.625$ at $\alpha_3 = 0.025$

More powerful than original procedure in Family 1 since $\gamma_3 > \gamma_1$

H_1 is guaranteed to be rejected

H_2 may be rejected in Step 3 even though it was not rejected in Step 1

4. Simulation study

Clinical trial

Design

Three arms: Two doses of experimental treatment (Doses L and H) versus placebo

Sample size: 280 patients per arm

Endpoints

Primary endpoint (Endpoint P) and key secondary endpoint (Endpoint S)

Correlation is 0.5

Two-family problem

Family 1

Primary endpoint: Dose H versus placebo (H_1) and Dose L versus placebo (H_2)

Family 1 is a parallel gatekeeper for Family 2

Family 2

Secondary endpoint: Dose H versus placebo (H_3) and Dose L versus placebo (H_4)

Gatekeeping procedures

Procedure Bonf

Simple Bonferroni-based parallel gatekeeping procedure

Two steps

Step 1: Bonferroni procedure in Family 1

Step 2: Holm procedure in Family 2 if at least hypothesis is rejected in Family 1

Gatekeeping procedures

Procedure BonfRetest

Simple Bonferroni-based parallel gatekeeping procedure with re-testing

Three steps

Step 1: Bonferroni procedure in Family 1

Step 2: Holm procedure in Family 2 if at least hypothesis is rejected in Family 1

Step 3: Holm procedure in Family 1 if there is a non-rejected hypothesis in Family 1 and both hypotheses are rejected in Family 2

Gatekeeping procedures

Procedure Hoch

Advanced Hochberg-based parallel gatekeeping procedure

Two steps

Step 1: Truncated Hochberg procedure with $\gamma = 0.8$ in Family 1

Step 2: Hochberg procedure in Family 2 if at least hypothesis is rejected in Family 1

Gatekeeping procedures

Procedure HochRetest

Advanced Hochberg-based parallel gatekeeping procedure with re-testing

Two steps

Step 1: Truncated Hochberg procedure with $\gamma = 0.8$ in Family 1

Step 2: Hochberg procedure in Family 2 if at least hypothesis is rejected in Family 1

Step 3: Hochberg procedure in Family 1 if there is a non-rejected hypothesis in Family 1 and both hypotheses are rejected in Family 2

Power comparison

BonfRetest versus Bonf

Does re-testing help improve power of Bonferroni-based procedures?

HochRetest versus Hoch

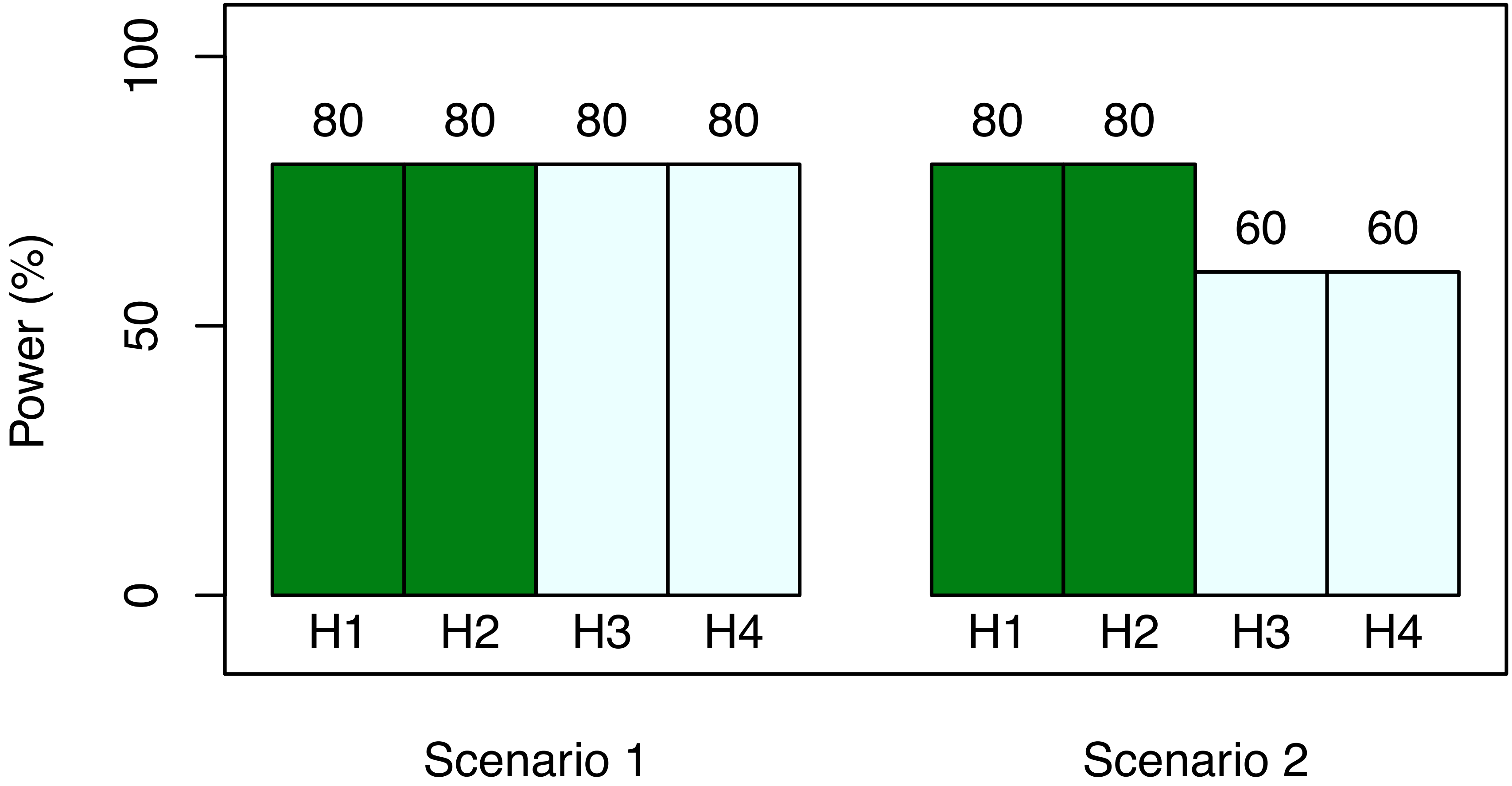
Does re-testing help improve power of Hochberg-based procedures?

Hoch versus Bonf and Hoch versus BonfRetest

Are Hochberg-based procedures more powerful than Bonferroni-based procedures?

Evaluation scenarios

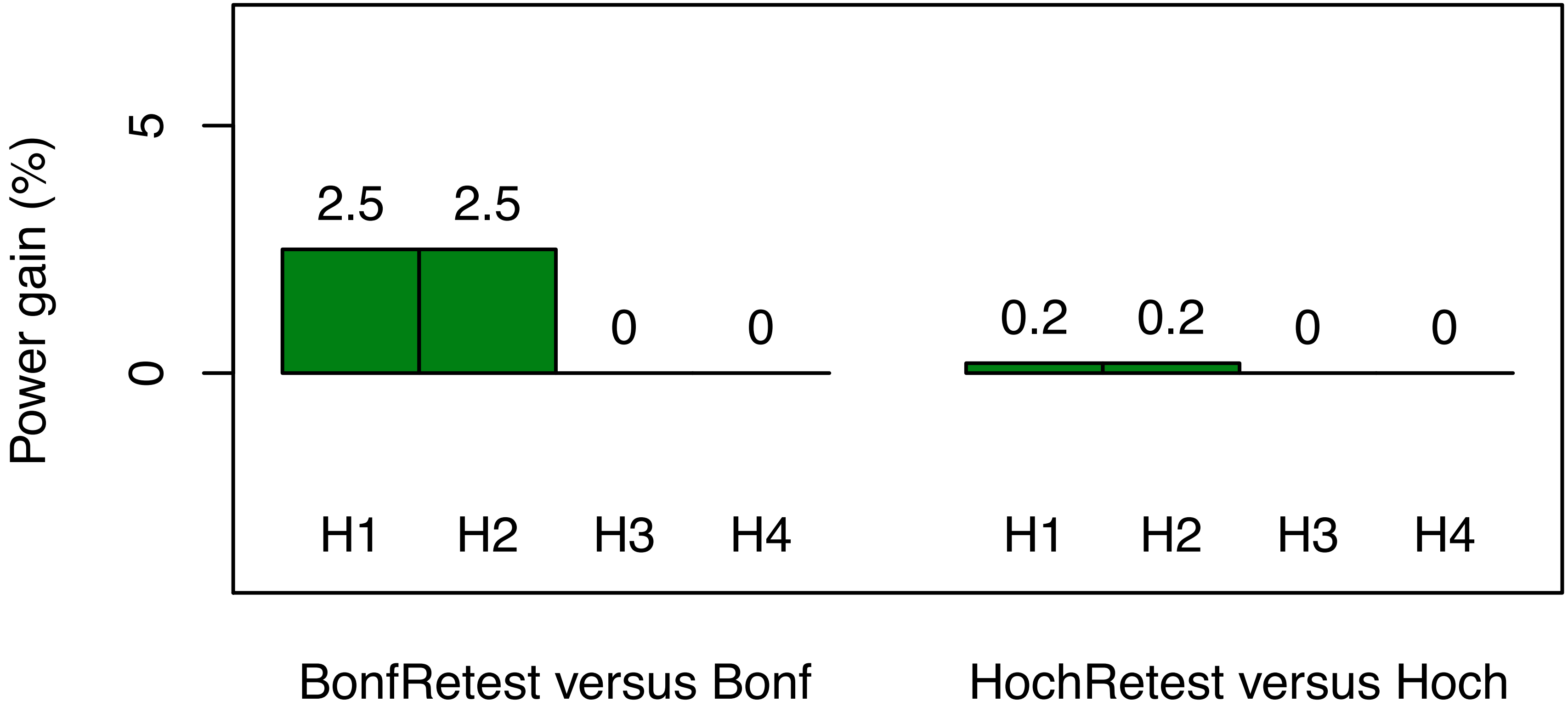
Marginal power of individual tests



Scenario 1: All tests are adequately powered,
Scenario 2: Secondary tests are underpowered

Scenario 1

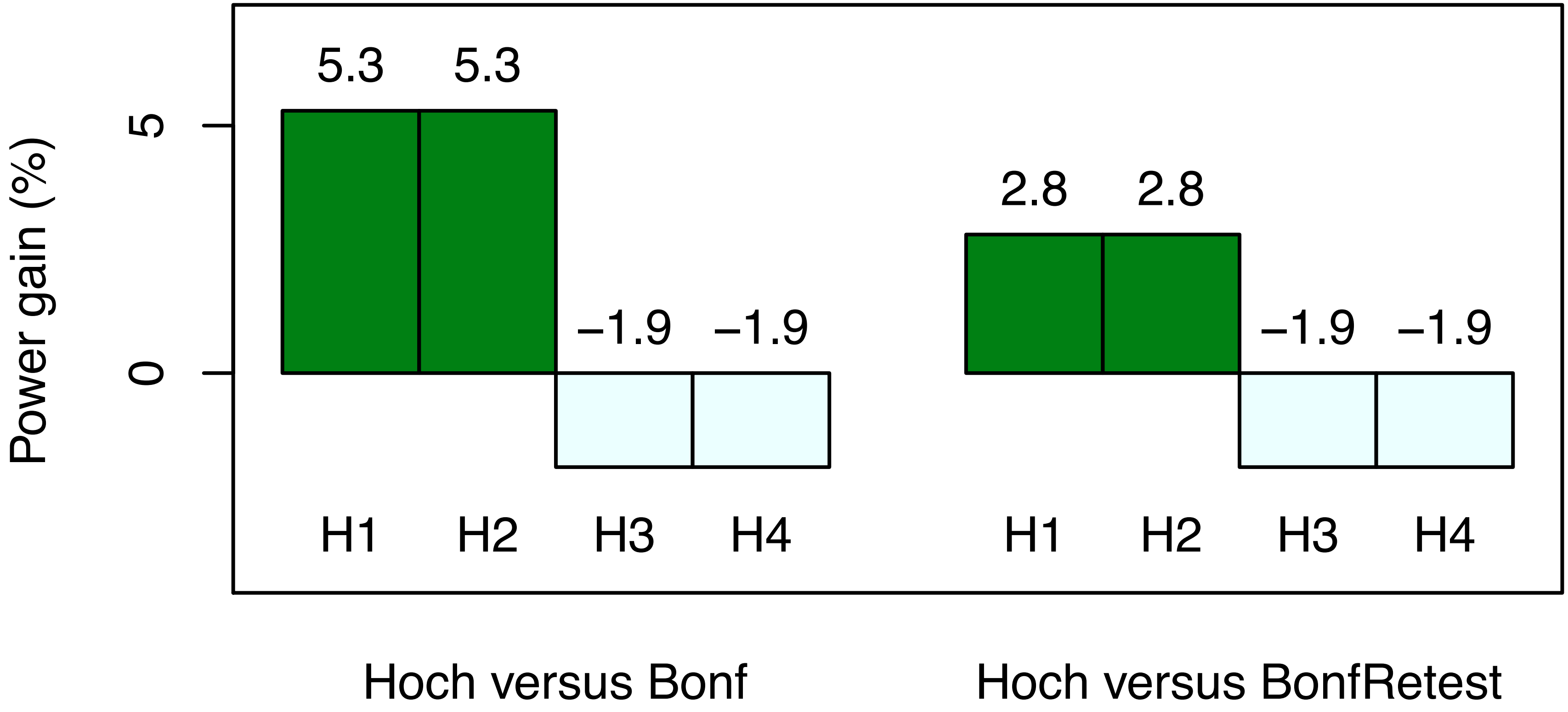
All tests are adequately powered



Re-testing helps improve power of primary tests for Bonferroni-based procedures but not Hochberg-based procedures

Scenario 1

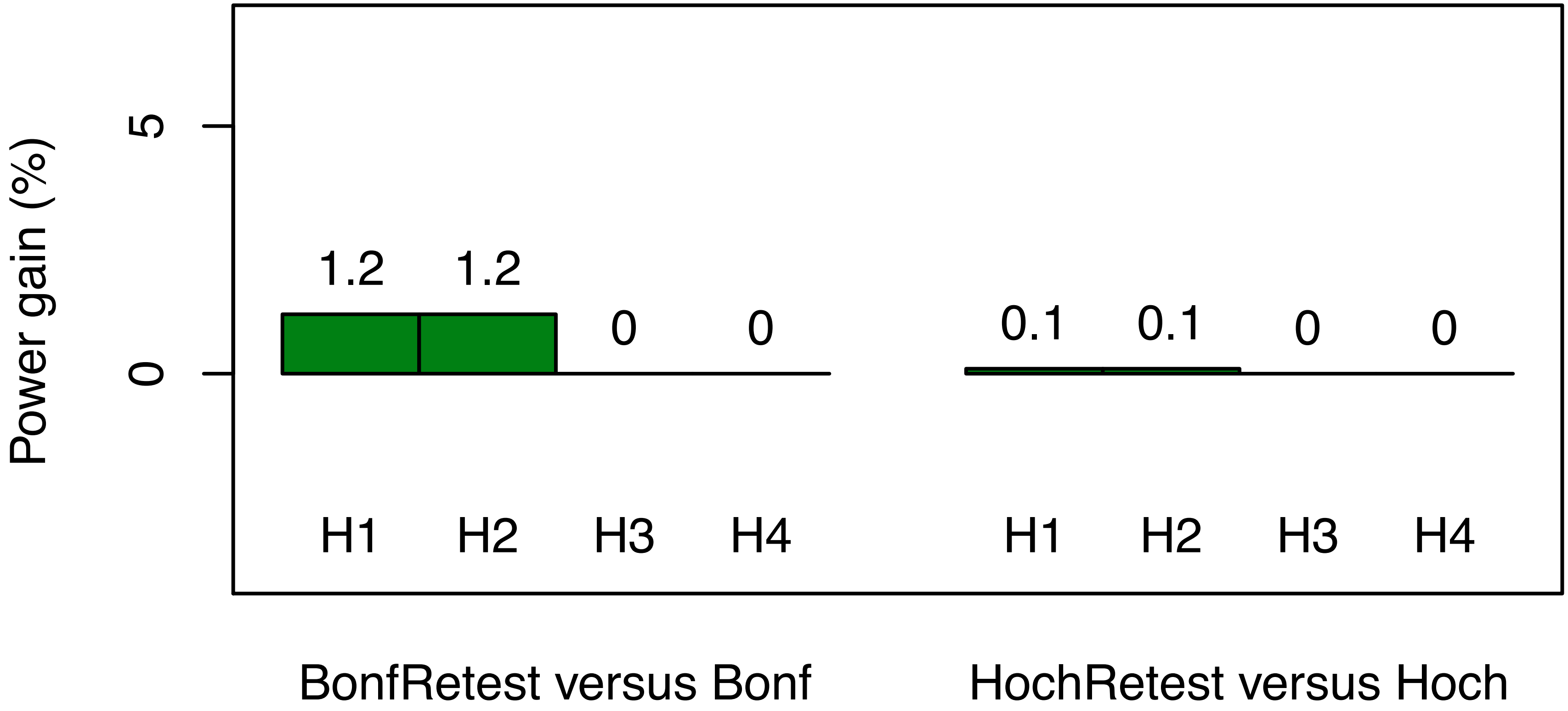
All tests are adequately powered



Hochberg-based procedures improve power of primary tests and reduce power of secondary tests compared to Bonferroni-based procedures

Scenario 2

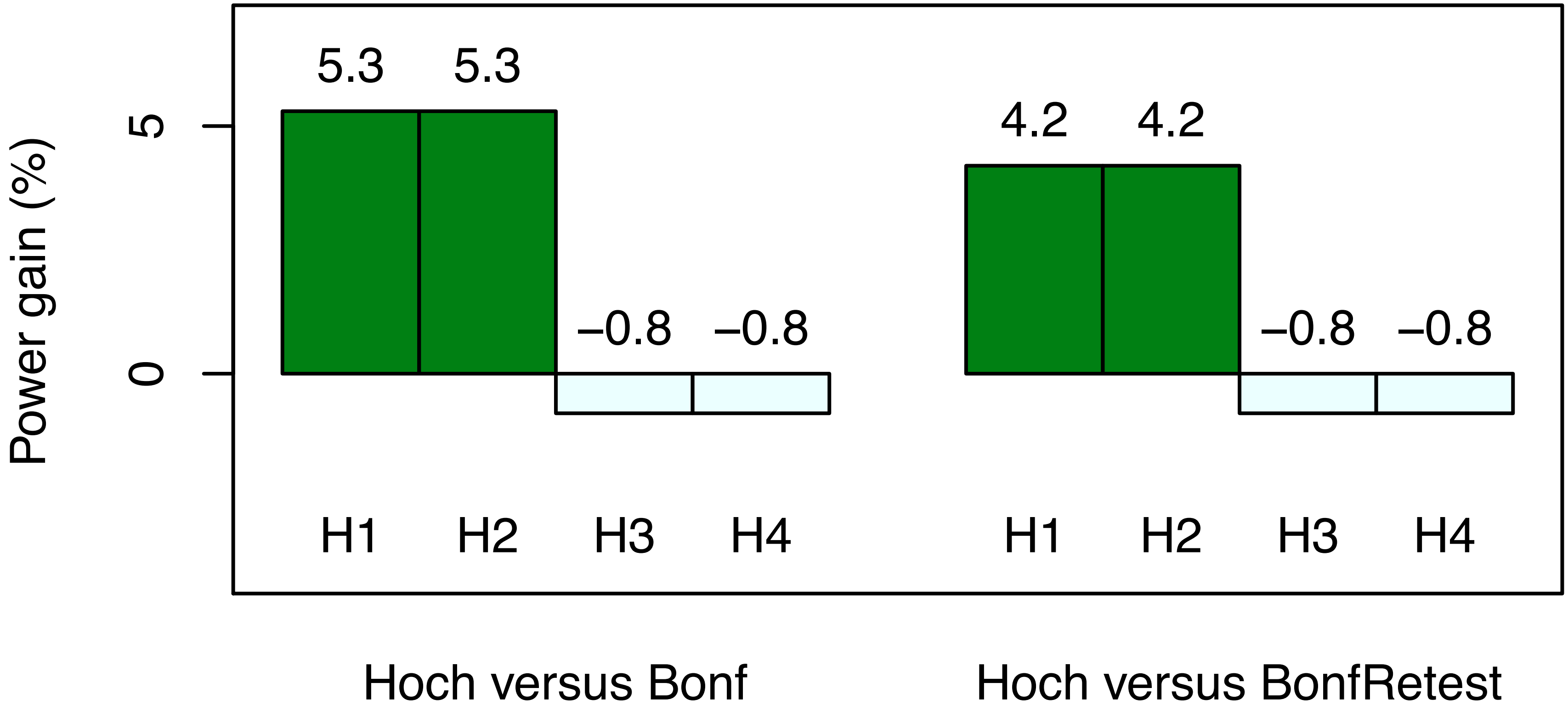
Secondary tests are underpowered



Re-testing helps improve power of primary tests for Bonferroni-based procedures but not Hochberg-based procedures

Scenario 2

Secondary tests are underpowered



Hochberg-based procedures improve power of primary tests and reduce power of secondary tests compared to Bonferroni-based procedures

Power comparison

Simple Bonferroni-based gatekeeping procedures

Re-testing helps improve power of primary tests

Primary tests tend to lose power after conservative Bonferroni adjustment

Advanced Hochberg-based gatekeeping procedures

Re-testing has virtually no impact on power of primary tests

Primary tests lose little power after efficient Hochberg adjustment

Power comparison

Bonferroni-based procedures versus Hochberg-based procedures

Hochberg-based procedure shift power balance from secondary tests to primary tests compared to Bonferroni-based procedure with or without re-testing

Power gain for primary tests outweighs power loss for secondary tests, especially when secondary tests are underpowered

5. Software implementation

Software implementation

Custom SAS macro

ParGate macro: Advanced parallel gatekeeping procedures

MultXpert R package

ParGate function: Advanced parallel gatekeeping procedures

Summary

Serial gatekeepers

Serial gatekeeping procedures

Serial gatekeepers: Reject all null hypotheses in current family to proceed to next family

Serial gatekeeping procedures are built from all-or-none components using a multistage algorithm similar to fixed-sequence procedure

Parallel gatekeepers

Parallel gatekeeping procedures

Parallel gatekeepers: Reject one or more null hypotheses in current family to proceed to next family

Parallel gatekeeping procedures can be built from powerful semiparametric or parametric components using a multistage algorithm with α propagation rules

Module I
Problems with General Gatekeepers

Outline

1. General gatekeeping procedures

Mixture-based method for constructing gatekeeping procedures with general logical relationships and general distributional relationships

2. Software implementation

Software implementation of general gatekeeping procedures in SAS

1. General gatekeeping procedures

Two families of null hypotheses

General gatekeeping

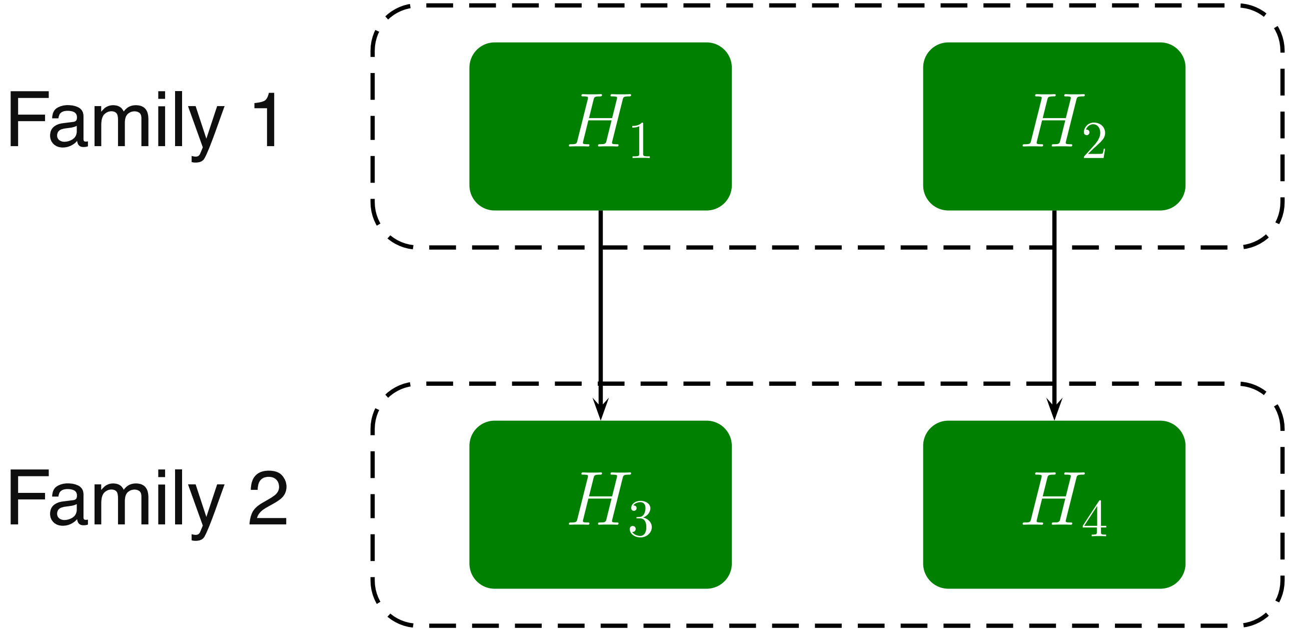
Family 1 $H_1, \ldots, H_k$

Family 2 $H_{k+1}, \ldots, H_{2k}$

Family 1 is not a parallel gatekeeper for Family 2

Example 9: Hypertension trial

Decision tree



Family 1

H_1 H_2

Family 2

H_3 H_4

Family 1: Doses L and H versus Active control (Noninferiority test)

Family 2: Doses L and H versus Active control (Superiority test)

General gatekeeping procedures

Mixture methodology

General method for constructing gatekeeping procedures (Dmitrienko and Tamhane, 2011; Dmitrienko and Tamhane, 2013)

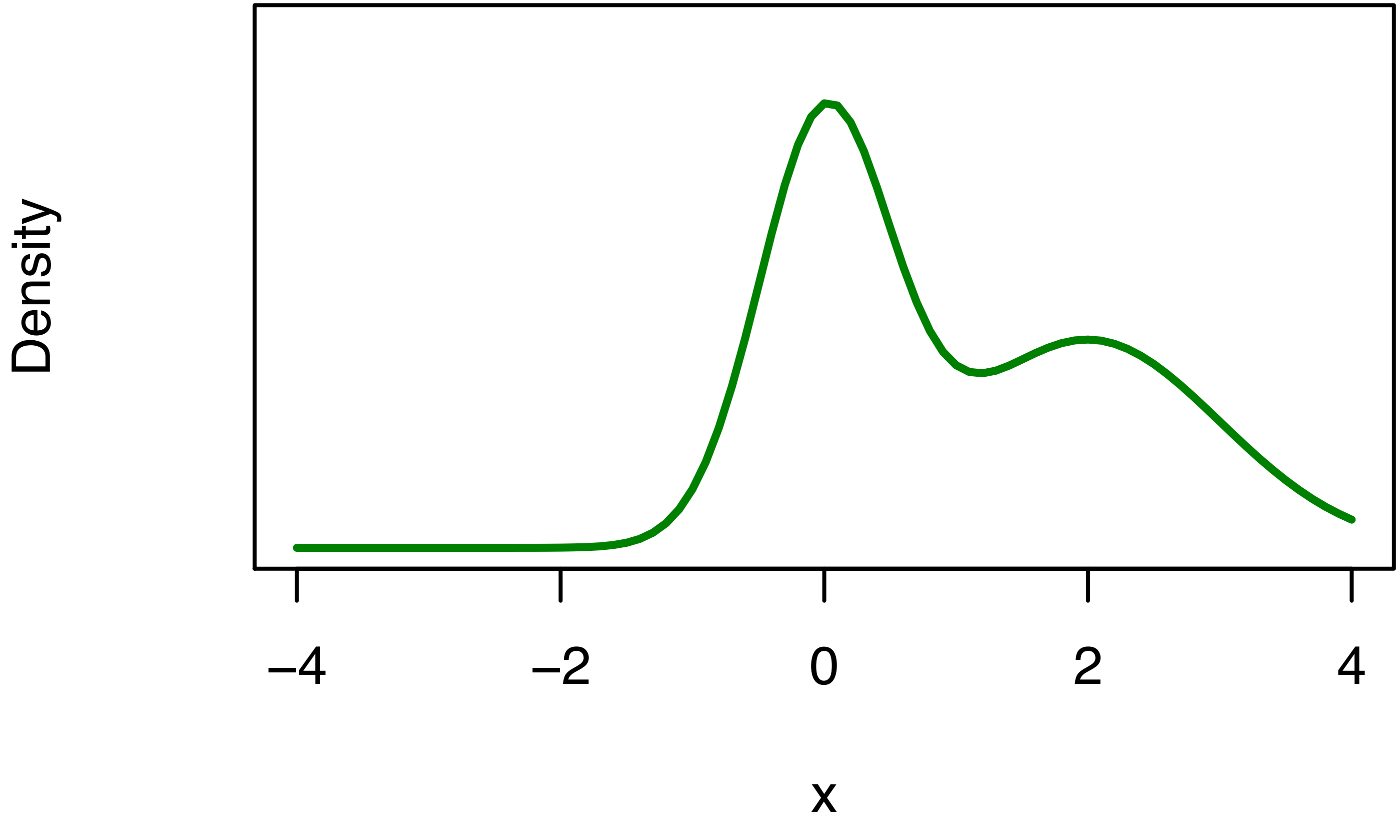
Gatekeeping procedure is a mixture of component procedures used within each family of null hypotheses

Mixture distributions

Similar to method for defining mixtures of multiple distributions

Mixture distributions

Mixture of two normal distributions



Mixture distributions

Three ingredients

Component distributions:

Normal distribution $N(0, 1/2)$

Normal distribution $N(2, 1)$

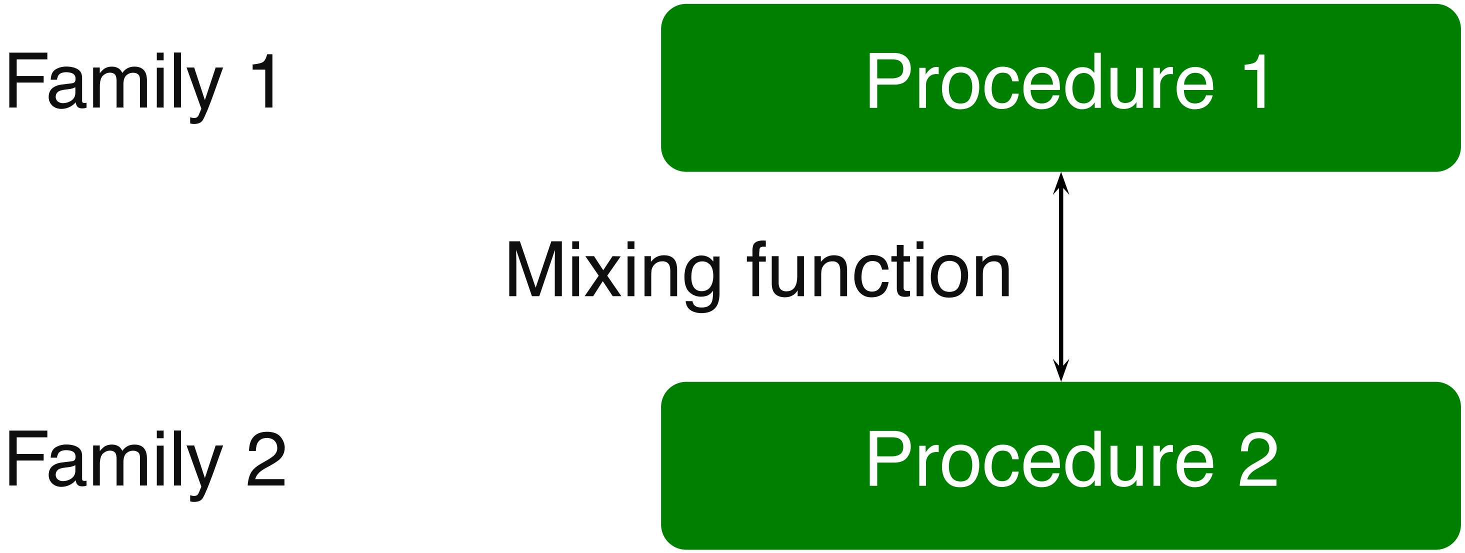
Mixing proportions, $1/2$ and $1/2$

Mixture distribution

Mixture $= N(0, 1/2)/2 + N(2, 1)/2$

General gatekeeping procedures

Mixture procedures



Family 1

Procedure 1

Mixing function

Family 2

Procedure 2

Procedure 1 controls local FWER within Family 1

Procedure 2 controls local FWER within Family 2

Mixture procedure controls global FWER in Families 1 and 2

General gatekeeping procedures

General logical relationships

Any logical relationships among null hypotheses can be taken into account

General distributional relationships

Within- and between-family correlations can be taken into account when appropriate

Example 9: Hypertension trial

Logical relationships

Rejected null hypotheses	Null hypothesis	Testable
H_1, H_2	H_3	Yes
H_1	H_3	Yes
H_2	H_3	No
Empty	H_3	No
H_1, H_2	H_4	Yes
H_1	H_4	No
H_2	H_4	Yes
Empty	H_4	No

Distributional relationships

Example 9: Hypertension trial

$t_1, \ldots, t_4$, test statistics for H_i, $i = 1, 2, 3, 4$

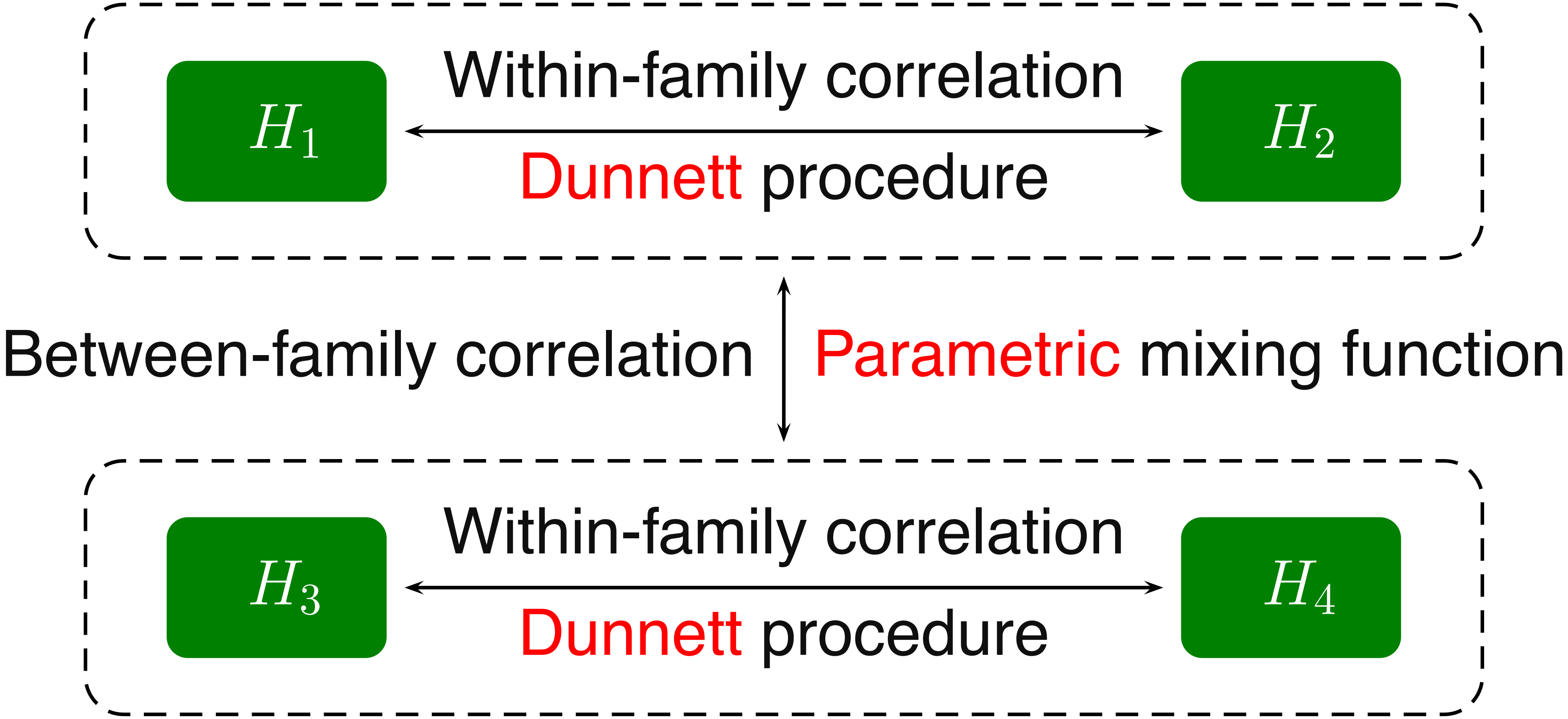
Assume a balanced design (n patients per treatment group)

Under global null hypothesis, $t_1, \ldots, t_4$ follow a standard t distribution with correlation matrix

$$\begin{bmatrix} 1 & 0.5 & 1 & 0.5 \\ 0.5 & 1 & 0.5 & 1 \\ 1 & 0.5 & 1 & 0.5 \\ 0.5 & 1 & 0.5 & 1 \end{bmatrix}$$

Example 9: Hypertension trial

Mixture-based procedure



H_1
Within-family correlation
Dunnett procedure
H_2
Between-family correlation
Parametric mixing function
H_3
Within-family correlation
Dunnett procedure
H_4

General gatekeeping procedures account for within- and between-family correlations

General gatekeeping procedures

Derivation

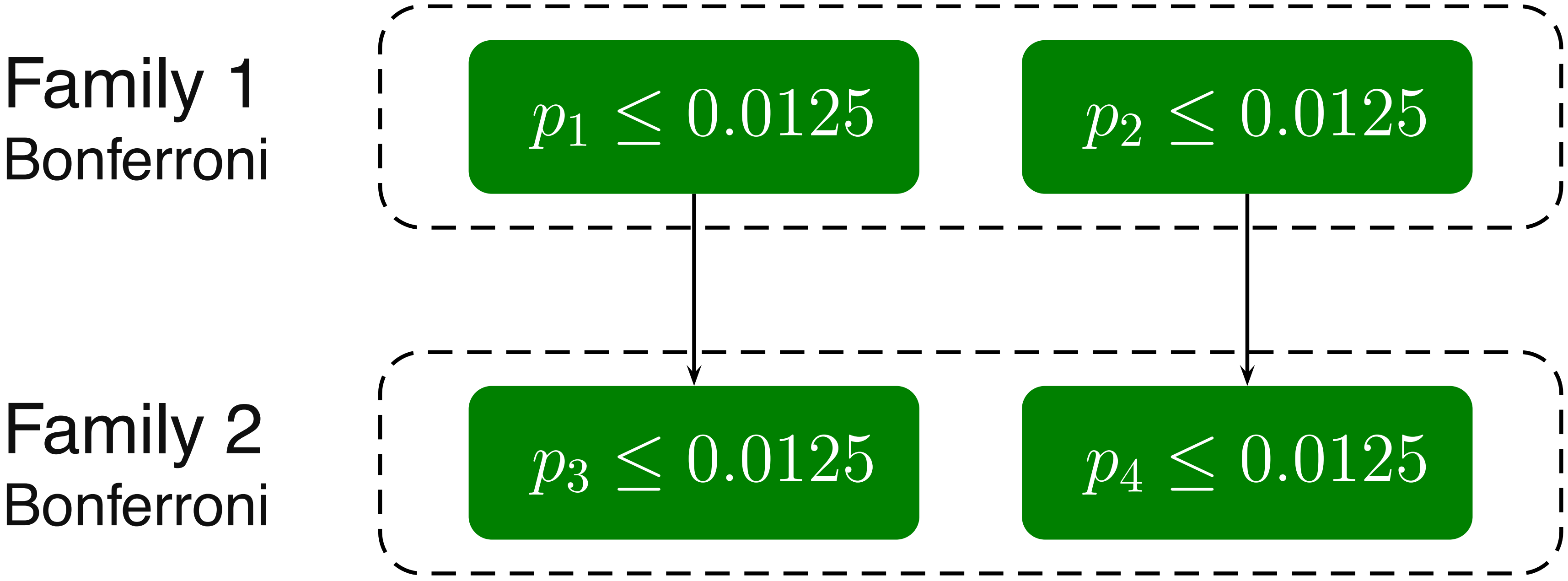
Based on the closure principle and more complex to derive compared to multistage parallel gatekeeping procedures

Decision rules

Underlying decision rules are generally very simple

Example 9: Hypertension trial

Bonferroni-based procedure

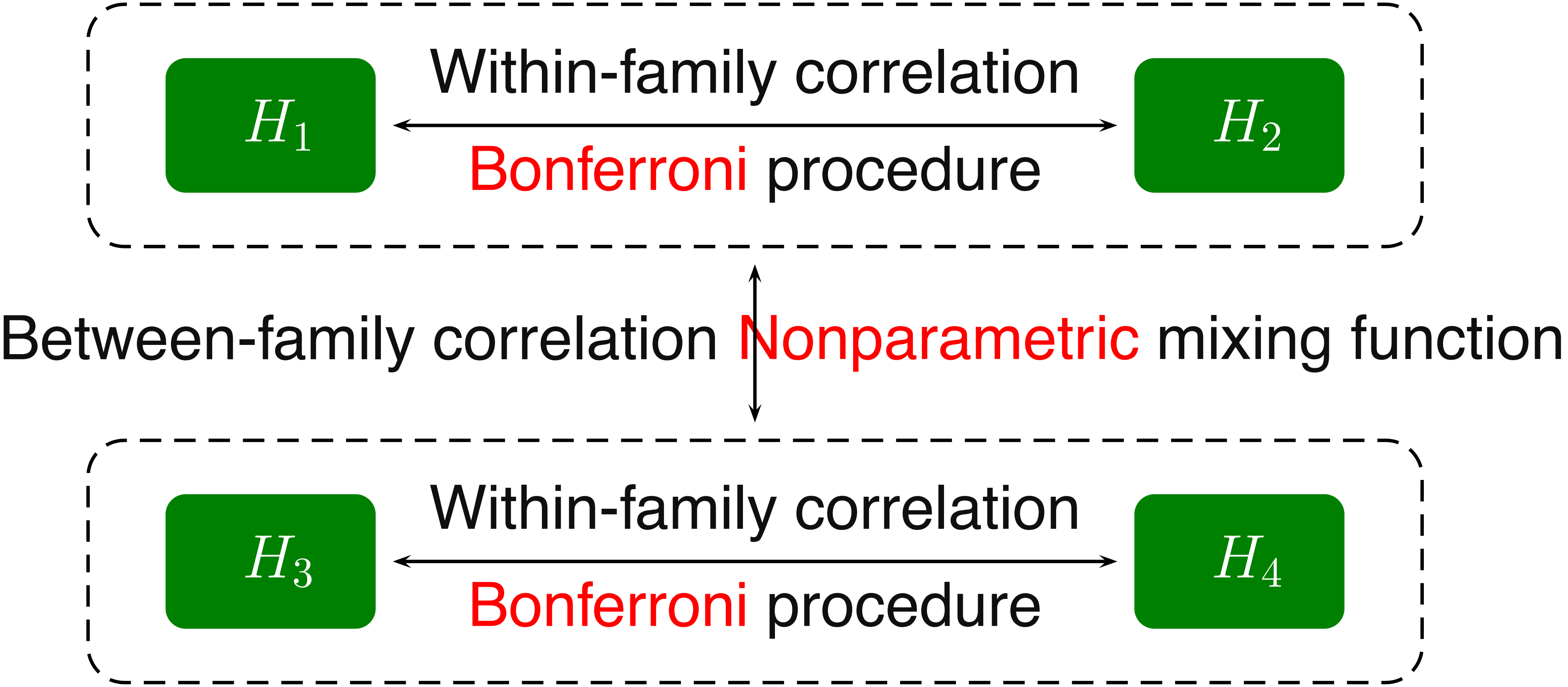


Within-family correlations are ignored (Bonferroni procedure is used)

Between-family correlations are ignored (nonparametric mixing function is used)

Example 9: Hypertension trial

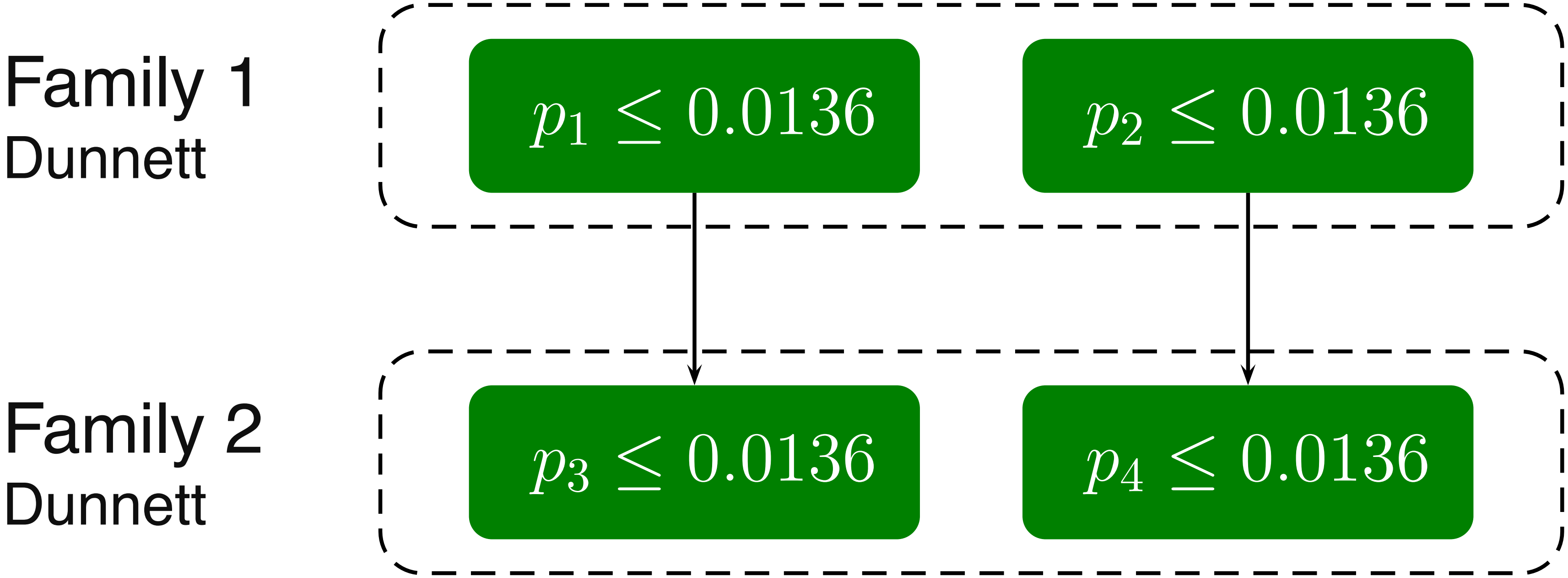
Bonferroni-based procedure



H_1
Within-family correlation
Bonferroni procedure
H_2
Between-family correlation
Nonparametric mixing function
H_3
Within-family correlation
Bonferroni procedure
H_4

Example 9: Hypertension trial

Mixture-based procedure



Family 1
Dunnett

$p_1 \leq 0.0136$

$p_2 \leq 0.0136$

Family 2
Dunnett

$p_3 \leq 0.0136$

$p_4 \leq 0.0136$

Within-family correlations are incorporated (Dunnett procedure is used)

Between-family correlations are incorporated (parametric mixing function is used)

Comparison of two gatekeeping procedures

Distributional relationships

Bonferroni-based procedure ignores distributional relationships

Dunnett-based procedure fully takes distributional relationships into account

Key property

Dunnett-based procedure is uniformly more powerful than Bonferroni-based procedure

Example 9: Hypertension trial

Four hypothesis tests

Family	Hypothesis	Raw p-value
Family 1	H_1	0.0290
	H_2	0.0121
Family 2	H_3	0.0310
	H_4	0.0131

$\alpha = 0.025$, global familywise error rate

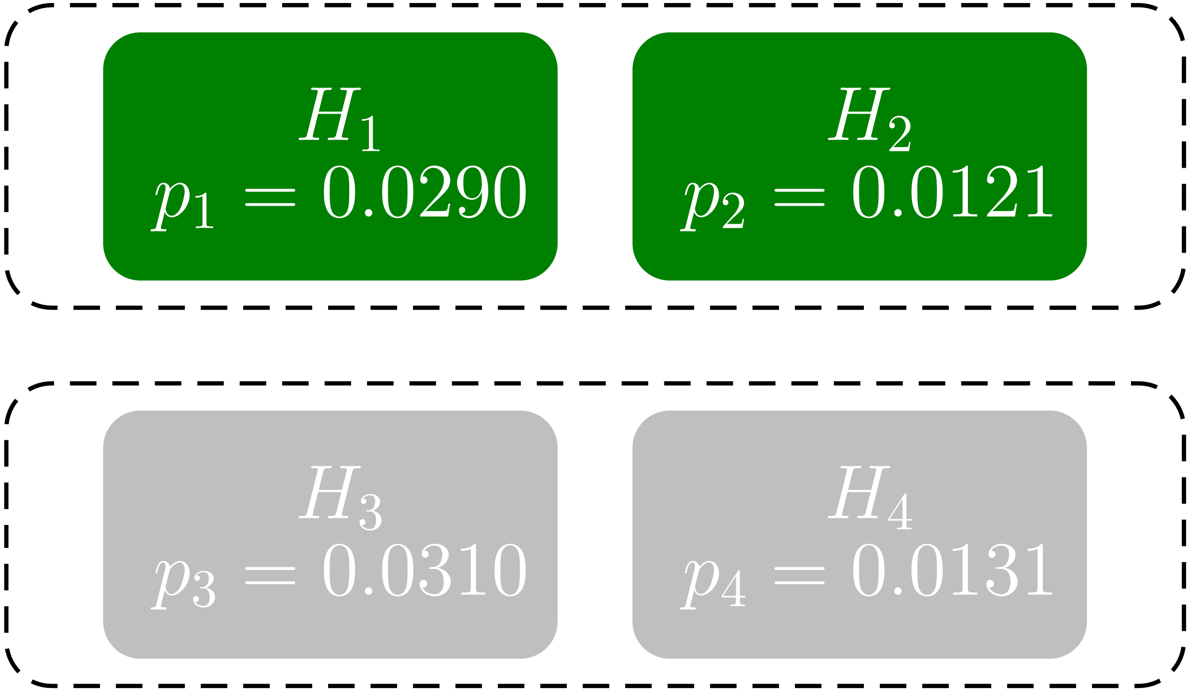
$n = 200$, number of patients per treatment arm

Mixture-based gatekeeping procedure

Procedure 1 (Dunnett procedure)

Procedure 1
at $\alpha_1 = 0.025$

Procedure 2



H_1
$p_1 = 0.0290$

H_2
$p_2 = 0.0121$

H_3
$p_3 = 0.0310$

H_4
$p_4 = 0.0131$

Test H_1 and H_2 at 0.0136

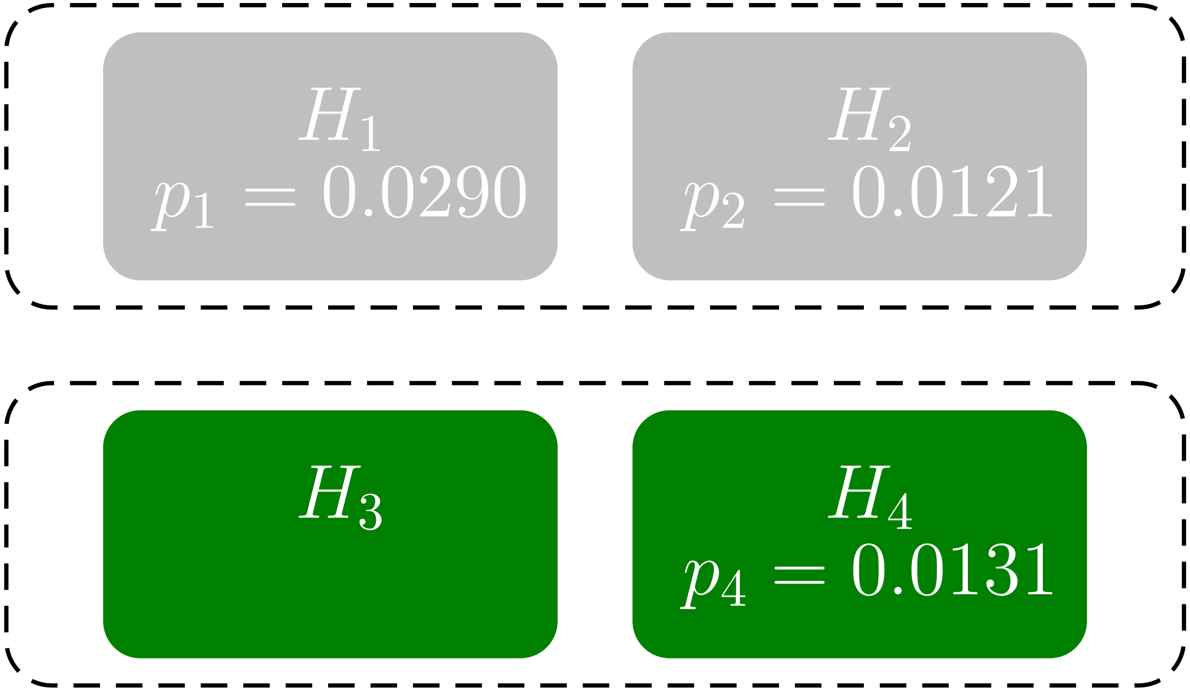
H_1 is accepted and H_2 is rejected

$\alpha_2 = 0.0136$ is carried over to Family 2

Mixture-based gatekeeping procedure

Procedure 2 (Univariate procedure)

Procedure 1



H_1

$p_1 = 0.0290$

H_2

$p_2 = 0.0121$

Procedure 2
at $\alpha_2 = 0.0136$

H_3

H_4

$p_4 = 0.0131$

Test H_4 at 0.0136 (H_3 is not testable and is automatically accepted)

H_4 is rejected

Computation of adjusted p-values

General definition

Adjusted p-value for a null hypothesis is the lowest significance level at which gatekeeping procedure rejects this null hypothesis

Direct-calculation algorithm

Adjusted p-value for H_i, $i = 1, 2, 3, 4$, is found using an iterative algorithm

Example 9: Hypertension trial

Comparison of two gatekeeping procedures

Hypothesis	Adjusted p-values	
	Procedure B	Procedure D
H_1	0.0580	0.0521
H_2	0.0242	0.0223
H_3	0.0620	0.0555
H_4	0.0262	0.0241

Procedure B: Bonferroni-based gatekeeping procedure

Procedure D: Mixture-based gatekeeping procedure

Procedure B rejects one null hypothesis and Procedure D rejects two null hypotheses at $\alpha = 0.025$

2. Software implementation

Software implementation

Custom SAS macro

TreeGate macro: Mixture-based general gatekeeping procedures

http://multxpert.com/wiki/Software

Summary

Summary

Gatekeeping procedures

Multiple testing procedures for problems with several families of hypotheses

Control global familywise error rate

Account for logical and distributional relationships among families

Enable clinical trial sponsors to enrich product labels by including key secondary findings

Modules G through I

Further reading

Multiple Testing Problems in Pharmaceutical Statistics (edited by Alex Dmitrienko, Ajit Tamhane and Frank Bretz)

Chapter 5 (Gatekeeping Procedures in Clinical Trials)

Multiplicity Expert web site

Recent publications on gatekeeping procedures

http://multxpert.com/wiki/Gatekeeping_Papers

Thank you!

Contact information

alex.dmitrienko@quintiles.com

Multiplicity Expert web site

http://multxpert.com/