Thorsten Dickhaus

Simultaneous Statistical Inference

With Applications in the Life Sciences

Springer

Simultaneous Statistical Inference

Thorsten Dickhaus

Simultaneous Statistical Inference

With Applications in the Life Sciences

Springer

Thorsten Dickhaus
Research Group "Stochastic Algorithms
and Nonparametric Statistics"
Weierstrass Institute for Applied Analysis
and Stochastics
Berlin
Germany

ISBN 978-3-642-45181-2 ISBN 978-3-642-45182-9 (eBook)
DOI 10.1007/978-3-642-45182-9
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013955256

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

It can't be all coincidence
Too many things are evident

(Iron Maiden, Infinite Dreams, 1988)

Preface

The more questions you ask, the more wrong answers you are expected to receive—even if every single source of your information is quite trustworthy. In this work, the sources of information are data, and the questions are formalized by statistical hypothesis-alternative pairs. From the mathematical point of view, this leads to multiple test problems. We will discuss criteria and methods (in particular multiple tests) which ensure that with high probability not too many wrong decisions are made, even if many hypotheses are of interest under the scope of one and the same statistical model, i.e., regarding one and the same dataset.

High-throughput technologies in different fields of modern life sciences have led to massive multiplicity and given rise to multiple test problems with more hypotheses than observations. Driven by these developments, also new statistical paradigms have arisen. It is fair to say that a new era of multiple testing began when Yoav Benjamini and Yosef Hochberg formally introduced the false discovery rate (FDR) and the linear step-up test for FDR control in 1995. In this book, apart from classical methods controlling the family-wise error rate (FWER), theory and important life science applications of the FDR are presented in a systematic way, presumably for the first time in this depth in a monograph. In this, focus is on frequentist approaches aiming at FDR control at a fixed level. Other type I and type II error rates are mentioned and discussed where appropriate, but focus is on FWER and FDR. Chapters 6 and 7 broaden the view and show how multiple testing methodology can be used in the context of binary classification and model selection, respectively, with life science applications provided in Parts II and III. Further relationships between multiple testing and other simultaneous statistical inference problems are discussed in Chap. 1 and at respective occasions.

The book is primarily meant to be a research monograph and an introduction to simultaneous inference for applied statisticians and practitioners from the life sciences. To this end, presentation is with emphasis on applicability and we provide a couple of hints concerning which multiple test to use for which type of data. Furthermore, Chap. 8 deals with software implementing the theoretically treated procedures. However, the mainly theoretical Part I of the book may also serve as the basis for a graduate course on simultaneous statistical inference with emphasis on multiple testing for mathematical statisticians. I used parts of Chaps. 2, 4 and 5 for such a course at Humboldt-University Berlin and a couple of diploma theses in mathematics originated from this teaching.

The material for this book originated from joint work with many colleagues. I acknowledge the respective co-workers at the end of each chapter. Apart from his scientific contributions, I am especially grateful to Taras Bodnar who critically read every chapter and provided many constructive comments which helped to improve the presentation.

My deepest gratitude, however, is due to the Thai branch of my family for their enduring support in hard times. Therefore, I dedicate this work to Prayun, Duangchan, Dako, Pipat, Janyarak and her children, and my wife Supansa.

Berlin, January 2014 Thorsten Dickhaus

Contents

1 The Problem of Simultaneous Inference 1
1.1 Sources of Multiplicity 3
1.2 Multiple Hypotheses Testing 4
1.2.1 Measuring and Controlling Errors. 4
1.2.2 Structured Systems of Hypotheses 8
1.3 Relationships to Other Simultaneous Statistical Inference Problems 9
1.4 Contributions of this Work 11
References 12

Part I General Theory

2 Some Theory of *p*-values. 17
2.1 Randomized *p*-values 20
2.1.1 Randomized *p*-values in Discrete Models 20
2.1.2 Randomized *p*-values for Testing Composite Null Hypotheses. 21
2.2 *p*-value Models 22
2.2.1 The iid.-Uniform Model 22
2.2.2 Dirac-Uniform Configurations 24
2.2.3 Two-Class Mixture Models 25
2.2.4 Copula Models Under Fixed Margins 26
2.2.5 Further Joint Models. 26
References 27

3 Classes of Multiple Test Procedures 29
3.1 Margin-Based Multiple Test Procedures 30
3.1.1 Single-Step Procedures 30
3.1.2 Stepwise Rejective Multiple Tests 32
3.1.3 Data-Adaptive Procedures 35
3.2 Multivariate Multiple Test Procedures. 37
3.2.1 Resampling-Based Methods 37
3.2.2 Methods Based on Central Limit Theorems 38

3.2.3 Copula-Based Methods 38
3.3 Closed Test Procedures 40
References 43

4 Simultaneous Test Procedures 47
4.1 Three Important Families of Multivariate Probability Distributions 50
4.1.1 Multivariate Normal Distributions 50
4.1.2 Multivariate *t*-distributions 51
4.1.3 Multivariate Chi-Square Distributions 51
4.2 Projection Methods Under Asymptotic Normality 52
4.3 Probability Bounds and Effective Numbers of Tests 56
4.3.1 Sum-Type Probability Bounds 57
4.3.2 Product-Type Probability Bounds 58
4.3.3 Effective Numbers of Tests 61
4.4 Simultaneous Test Procedures in Terms of *p*-value Copulae 62
4.5 Exploiting the Topological Structure of the Sample Space via Random Field Theory 65
References 68

5 Stepwise Rejective Multiple Tests 71
5.1 Some Concepts of Dependency 72
5.2 FWER-Controlling Step-Down Tests 74
5.3 FWER-Controlling Step-Up Tests 76
5.4 FDR-Controlling Step-Up Tests 80
5.5 FDR-Controlling Step-Up-Down Tests 82
References 89

6 Multiple Testing and Binary Classification 91
6.1 Binary Classification Under Sparsity 93
6.2 Binary Classification in Non-Sparse Models 96
6.3 Feature Selection for Binary Classification via Higher Criticism 99
References 101

7 Multiple Testing and Model Selection 103
7.1 Multiple Testing for Model Selection 104
7.2 Multiple Testing and Information Criteria 106
7.3 Multiple Testing After Model Selection 108
7.3.1 Distributions of Regularized Estimators 108
7.3.2 Two-Stage Procedures 111
7.4 Selective Inference 112
References 114

8 Software Solutions for Multiple Hypotheses Testing 117
8.1 The R Package multcomp 118
8.2 The R Package multtest 118
8.3 The R-based μTOSS Software 119
8.3.1 The μTOSS Simulation Tool 120
8.3.2 The μTOSS Graphical User Interface 122
References 124

Part II From Genotype to Phenotype

9 Genetic Association Studies 129
9.1 Statistical Modeling and Test Statistics 130
9.2 Estimation of the Proportion of Informative Loci 133
9.3 Effective Numbers of Tests via Linkage Disequilibrium 134
9.4 Combining Effective Numbers of Tests and Pre-estimation of π_0 137
9.5 Applicability of Margin-Based Methods 138
References 139

10 Gene Expression Analyses 141
10.1 Marginal Models and p-values 141
10.2 Dependency Considerations 143
10.3 Real Data Examples 146
10.3.1 Application of Generic Multiple Tests to Large-Scale Data 146
10.3.2 Copula Calibration for a Block of Correlated Genes 147
10.4 LASSO and Statistical Learning Methods 149
10.5 Gene Set Analyses and Group Structures 150
References 151

11 Functional Magnetic Resonance Imaging 155
11.1 Spatial Modeling 156
11.2 False Discovery Rate Control for Grouped Hypotheses 157
11.2.1 Clusters of Voxels 157
11.2.2 Multiple Endpoints per Location 159
11.3 Exploiting Topological Structure by Random Field Theory . . . 160
11.4 Spatio-Temporal Models via Multivariate Time Series 161
11.4.1 Which of the Specific Factors have a Non-trivial Autocorrelation Structure? 164
11.4.2 Which of the Common Factors have a Lagged Influence on Which X_i? 165
References 165

Part III Further Applications in the Life Sciences

12 Further Life Science Applications 169
12.1 Brain-Computer Interfacing 169
12.2 Gel Electrophoresis-Based Proteome Analysis 172
References 174

Index 177

Acronyms

$\mathscr{X}, \mathscr{F}, \mathscr{P}$	Statistical model
$\mathscr{X}, \mathscr{F}, \mathscr{P}, \mathscr{H}$	Multiple test problem
V_m	Number of type I errors among m tests
R_m	Total number of rejections among m tests
$\overline{\mathbb{R}}$	$\mathbb{R} \cup \{-\infty, +\infty\}$
Φ	Cumulative distribution function of the standard normal law on $\mathbb{R}$
ϕ	Lebesgue density of the standard normal law on $\mathbb{R}$
χ^2_v	Chi-square distribution with v degrees of freedom
t_v	Student's t-distribution with v degrees of freedom
$\text{Beta}(a, b)$	Beta distribution with parameters a and b
$W_m(v, \Sigma)$	Wishart distribution with parameters m, v and Σ
$\mathscr{M}_c(n, p)$	Multinomial distribution with c categories, sample size n and vector of probabilities p
1_A	Indicator function of the set A
$\mathscr{N}(\mu, \sigma^2)$	Normal distribution on $\mathbb{R}$ with mean μ and variance σ^2
$\mathscr{N}_k(\mu, \Sigma)$	Normal distribution on $\mathbb{R}^k$ with mean vector μ and covariance matrix Σ
$\mathscr{L}(X)$	Law (or distribution) of the random variate X
$\Gamma(\cdot)$	Gamma function, $\Gamma\ (x) = \int_0^\infty t^{x-1}\exp(-\text{t})\text{dt}, \text{x} > 0$
$\det(A)$	Determinant of the matrix A
A^+	Moore-Penrose pseudo inverse of the matrix A
$\text{diag}(\cdots)$	Diagonal matrix the diagonal elements of which are given by $\cdots$
$\cap$-closed	Closed under intersection
F_{v_1, v_2}	Fisher's F-distribution with v_1 and v_2 degrees of freedom
$\mathcal{B}(\mathcal{X})$	System of Borel sets of $\mathcal{X}$
$\overline{X}_{i.}$	Empirical mean in group $1 \leq i \leq k$ in k-sample problem
$\xrightarrow{w}$	Weak convergence
$\xrightarrow{d}$	Convergence in distribution
$\overset{d}{=}$	Equality in distribution
ABOS	Asymptotically Bayes optimal under sparsity
ANOVA	Analysis of variance
AORC	Asymptotically optimal rejection curve

BCI	Brain-computer interface
BOLD	Blood oxygen level dependent
BPI	Bonferroni plug-in
cdf	Cumulative distribution function
CRAN	Comprehensive R Archive Network
CSP	Common spatial pattern
DFM	Dynamic factor model
EC	Euler characteristic
ecdf	Empirical cumulative distribution function
EEG	Electroencephalogram
ERD	Event-related desynchronization
FCR	False coverage-statement rate
FDP	False discovery proportion
FDR	False discovery rate
fMRI	Functional magnetic resonance imaging
FWER	Family-wise error rate
GED	Generalized error distribution
GLM	Generalized linear model
HC	Higher criticism
LASSO	Least absolute shrinkage and selection operator
LD	Linkage disequilibrium
LDA	Linear discriminant analysis
LFC	Least favorable (parameter) configuration
iid.	Independent and identically distributed
MCP	Multiple comparison procedure
MLE	Maximum likelihood estimator
MSM_i	Monotonically sub-Markovian of order i
MTP	Multiple test procedure
MTP_2	Multivariate total positivity of order 2
pdf	Probability density function
pFDR	Positive false discovery rate
pFNR	Positive false non-discovery rate
PLOD	Positive lower orthant dependent
pmf	Point mass function
PRDS	Positive regression dependency on subsets
ROI	Region of interest
SD	Step-down
SNP	Single nucleotide polymorphism
SPC	Subset pivotality condition
STP	Simultaneous test procedure
SU	Step-up
SUD	Step-up-down
SVM	Support vector machine
μTOSS	Multiple hypothesis testing in an open software system
UNI$[a, b]$	Uniform distribution on the interval $[a, b]$

Chapter 1
The Problem of Simultaneous Inference

Abstract We introduce the problem of simultaneous statistical inference, with particular emphasis on testing multiple hypotheses. After a historic overview, general notation for the whole work is set up and different sources of multiplicity are distinguished. We define a variety of classical and modern type I and type II error rates in multiple hypotheses testing, analyze some relationships between them, and consider different ways to cope with structured systems of hypotheses. Relationships between multiple testing and other simultaneous statistical inference problems, in particular the construction of confidence regions for multi-dimensional parameters, as well as selection, ranking and partitioning problems, are elucidated. Finally, a general outline of the remainder of the work is given.

Simultaneous statistical inference is concerned with the problem of making several decisions simultaneously based on one and the same dataset. In this work, simultaneous statistical decision problems will mainly be formalized by multiple hypotheses and multiple tests. Not all simultaneous statistical decision problems are given in this formulation in the first place, but they can often be re-formulated in terms of multiple test problems. General relationships between multiple testing and other kinds of simultaneous statistical decision problems will briefly be discussed in Sect. 1.3. Moreover, we will refer to specific connections at respective occasions. For instance, we will elucidate connections between multiple testing and binary classification in Chap. 6 and discuss multiple testing methods in the context of model selection in Chap. 7.

The origins of multiple hypotheses testing can at least be traced back to Bonferroni (1935, 1936). The "Bonferroni correction"(cf. Example 3.1) is a generic method for evaluating several statistical tests simultaneously and ensuring that the probability for *at least one* type I error is bounded by a pre-defined significance level α. The latter criterion is nowadays referred to as (strong) control of the family-wise error rate (FWER) at level α and will be defined formally in Definition 1.2 below. In well-defined model classes, the Bonferroni method can be improved. In the 1950s, especially analysis of variance (ANOVA) models have been studied with respect to multiple comparisons of group-specific means. For instance, Tukey (1953)

T. Dickhaus, *Simultaneous Statistical Inference*,
DOI: 10.1007/978-3-642-45182-9_1,

developed a multiple test for all pairwise comparisons of means in ANOVA models based on the studentized range distribution. Keuls (1952) applied this technique to a ranking problem of ANOVA means in an agricultural context. The works of Dunnett (1955, 1964) treated the problem of multiple comparisons with a control group, while Scheffé (1953) provided a method for testing general linear contrasts simultaneously in the ANOVA context.Concepts from multivariate analysis and probability theory, in particular multivariate dependency concepts, have also been used for multiple testing, cf. for instance the works by Šidák (1967, 1968, 1971, 1973). These concepts allow for establishing probability bounds which in turn can be used for adjusting significance levels for multiplicity. We will provide details in Sect. 4.3. While all the aforementioned historical methods lead to single-step tests (meaning that the same, multiplicity-adjusted critical value is used for all test statistics corresponding to the considered tests), the formal introduction of the closed test principle by Marcus et al. (1976) paved the way for stepwise rejective multiple tests (for a detailed description of these different classes of multiple test procedures, see Chap. 3). These stepwise rejective tests are often improvements of the classical single-step tests with respect to power, meaning that they allow (on average) for more rejections of false hypotheses while controlling the same type I error criterion (namely, the FWER at a given level of significance). Stepwise rejective FWER-controlling multiple tests have been developed in the late 1970s, the 1980s and early 1990s; see, for example, Holm (1977, 1979), Hommel (1988) (based on Simes (1986)), Hochberg (1988), and Rom (1990). Around this time, the theory of FWER control had reached a high level of sophistication and was treated in the monographs by Hochberg and Tamhane (1987) and Hsu (1996).

It is fair to say that a new era of multiple testing began when Benjamini and Hochberg (1995) introduced a new type I error criterion, namely control of the false discovery rate (FDR), see Definition 1.2. Instead of bounding the probability of one or more type I errors, the FDR criterion bounds the expected proportion of false positives among all significant findings, which typically implies to allow for a few type I errors; see also Seeger (1968) and Sorić (1989) for earlier instances of this idea. During the past 20 years, simultaneous statistical inference and, in particular, multiple statistical hypothesis testing has become a major branch of mathematical and applied statistics, cf. Benjamini (2010) for some bibliometric details. Even for experts it is hardly possible to keep track of the exponentially (over time) growing literature in the field. This growing importance is not least due to the data-analytic challenges posed by large-scale experiments in modern life sciences such as, for instance, genetic association studies (cf. Chap. 9), gene expression studies (Chap. 10), functional magnetic resonance imaging (Chap. 11), and brain-computer interfacing (Chap. 12). Hence, the present work is attempting to explain some of the most important theoretical basics of simultaneous statistical inference, together with applications in diverse areas of the life sciences.

1.1 Sources of Multiplicity

The following definition is fundamental for the remainder of this work.

Definition 1.1 (Statistical model). A statistical model is a triple $(\mathscr{X}, \mathscr{F}, \mathscr{P})$. In this, $\mathscr{X}$ denotes the sample space (the set of all possible observations), $\mathscr{F}$ a σ-field on $\mathscr{X}$ (the set of all events that we can assign a probability to) and $\mathscr{P}$ a family of probability measures on the measurable space $(\mathscr{X}, \mathscr{F})$. Often, we will write $\mathscr{P}$ in the form $(\mathbb{P}_\vartheta)_{\vartheta \in \Theta}$, such that the family is indexed by the parameter ϑ of the model which can take values in the parameter space Θ, where Θ may have infinite dimension. Unless stated otherwise, an observation will be denoted by $x \in \mathscr{X}$, and we think of x as the realization of a random variate X which mathematically formalizes the data-generating mechanism. The target of statistical inference is the parameter ϑ which we regard as the unknown and unobservable state of nature.

Once the statistical model for the data-generating process at hand is defined, two general types of resulting multiplicity can be labeled as "one- or two- sample problems with multiple endpoints" and "k-sample problems with localized comparisons", where $k > 2$, respectively. In one- or two- sample problems with multiple endpoints, the sample space is often of the form $\mathscr{X} = \mathbb{R}^{m \times n}$. The same n observational units are measured with respect to m different endpoints, where we assumed for ease of presentation that every measurement results in a real number. The transfer to measurements of other type (for instance, allele pairs at genetic loci) is straightforward. For every of the m endpoints, an own scientific question can be of interest. On the contrary, in k-sample problems with localized comparisons, the sample space is typically of the form $\mathscr{X} = \mathbb{R}^{\sum_{i=1}^{k} n_i}$, meaning that $k > 2$ different groups of observational units (for instance, corresponding to k different doses of a drug) are considered, and that n_i observations are made in group i, where $1 \leq i \leq k$. In this, all $\sum_{i=1}^{k} n_i$ measurements concern one and the same endpoint (for instance, a disease status). The scientific questions in the latter case typically relate to differences between the k groups. Multiplicity arises, if not (only) general homogeneity or heterogeneity between the groups shall be assessed, but if differences, if any, are to be localized in the sense that we want to find out *which* groups are different. Two classical examples are the "all pairs" problem (all $m = k(k-1)/2$ pairwise group comparisons are of interest) and the "multiple comparisons with a control" problem (group k is a reference group and all other $m = k - 1$ groups are to be compared with group k).

We will primarily focus on these two kinds of problems. However, it has to be mentioned that they do not cover the whole spectrum of simultaneous statistical inference problems. For instance, flexible (group-sequential and adaptive) study designs induce a different type of multiplicity problem that we will not consider in the present work.

Throughout the remainder, we will try to stick to the notation developed in this section: m is the number of comparisons (the multiplicity of the problem), n or a subscripted n denotes a sample size and k refers to the total number of groups in a

k-sample problem or to the dimensionality of the parameter ϑ. Often, the two latter quantities are identical.

1.2 Multiple Hypotheses Testing

In what follows, we (sometimes implicitly) identify statistical hypotheses with non-empty subsets of the parameter space Θ. The tuple $(\mathscr{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta}, \mathscr{H})$ denotes a multiple test problem, where $\mathscr{H} = (H_i : i \in I)$ for an arbitrary index set I defines a family of null hypotheses. The resulting alternative hypotheses are denoted by $K_i = \Theta \setminus H_i, i \in I$. The intersection hypothesis $H_0 = \bigcap_{i\in I} H_i$ will be referred to as global hypothesis. Throughout the work, we assume that H_0 is non-empty.With very few exceptions, we will consider the case of finite families of hypotheses, meaning that $|I| = m \in \mathbb{N}$. In such cases, we will often write $\mathscr{H}_m$ instead of $\mathscr{H}$ and index the hypotheses such that $I = \{1, \ldots, m\}$. A (non-randomized) multiple test for $(\mathscr{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta}, \mathscr{H}_m)$ is a measurable mapping $\varphi = (\varphi_i)_{1\le i\le m} : \mathscr{X} \to \{0, 1\}^m$ the components of which have the usual interpretation of a statistical test for H_i versus K_i. Namely, H_i is rejected if and only if $\varphi_i(x) = 1$, where $x \in \mathscr{X}$ denotes the observed data.

1.2.1 Measuring and Controlling Errors

The general decision pattern of a multiple test for m hypotheses is summarized in Table 1.1. In contrast to usual, one-dimensional test problems, it becomes apparent that type I and type II errors can occur simultaneously. In Table 1.1, type I errors are counted by V_m and type II errors are counted by T_m. The total number of rejections is denoted by R_m. Notice that the quantities U_m, V_m, T_m, S_m and m_0, m_1 all depend on the unknown value of the parameter ϑ (although we suppressed this dependence on ϑ notationally in Table 1.1) and are therefore unobservable. Only m and R_m can be observed.

For a given $\vartheta \in \Theta$, we denote the index set of true null hypotheses in $\mathscr{H}_m$ by $I_0 \equiv I_0(\vartheta) = \{1 \le i \le m : \vartheta \in H_i\}$. Analogously, we define $I_1 \equiv I_1(\vartheta) = I \setminus I_0$. With this notation, we can formally define $V_m \equiv V_m(\vartheta) = |\{i \in I_0(\vartheta) : \varphi_i = 1\}|$, $S_m \equiv S_m(\vartheta) = |\{i \in I_1(\vartheta) : \varphi_i = 1\}|$, and $R_m \equiv R_m(\vartheta) = |\{i \in I : \varphi_i =$

Table 1.1 Decision pattern of a multiple test procedure

	Test decisions		
Hypotheses	0	1	
True	U_m	V_m	m_0
False	T_m	S_m	m_1
	W_m	R_m	m

$1\}| = V_m + S_m$. Based on these quantities, the following definition is concerned with measuring and controlling type I errors of a multiple test φ.

Definition 1.2 (Multiple type I error rates). Let $(\mathscr{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta}, \mathscr{H}_m)$ denote a multiple test problem and $\varphi = (\varphi_i : i \in I)$ a multiple test for $(\mathscr{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta}, \mathscr{H}_m)$.

(a) The number

$$\mathrm{FWER}_\vartheta(\varphi) = \mathbb{P}_\vartheta(V_m > 0) = \mathbb{P}_\vartheta\left(\bigcup_{i\in I_0(\vartheta)} \{\varphi_i = 1\}\right)$$

is called the family-wise error rate (FWER) of φ under ϑ.

(b) The random variable

$$\mathrm{FDP}_\vartheta(\varphi) = \frac{V_m}{\max(R_m, 1)}$$

is called the false discovery proportion (FDP) of φ under ϑ.

(c) The number

$$\mathrm{FDR}_\vartheta(\varphi) = \mathbb{E}_\vartheta[\mathrm{FDP}_\vartheta(\varphi)] = \mathbb{E}_\vartheta\left[\frac{V_m}{\max(R_m, 1)}\right]$$

is called the false discovery rate (FDR) of φ under ϑ.

(d) The number

$$\mathrm{pFDR}_\vartheta(\varphi) = \mathbb{E}_\vartheta\left[\frac{V_m}{R_m} \mid R_m > 0\right]$$

is called the positive false discovery rate (pFDR) of φ under ϑ.

(e) The multiple test φ is called a multiple test at local level $\alpha \in (0, 1)$, if each φ_i is a level α test for H_i versus K_i.

(f) The multiple test φ is said to control the FWER in the strong sense (strongly) at level $\alpha \in (0, 1)$, if

$$\sup_{\vartheta\in\Theta} \mathrm{FWER}_\vartheta(\varphi) \le \alpha. \tag{1.1}$$

(g) The multiple test φ is said to control the FWER in the weak sense (weakly) at level $\alpha \in (0, 1)$, if

$$\forall \vartheta \in H_0 : \mathrm{FWER}_\vartheta(\varphi) \le \alpha. \tag{1.2}$$

(h) The multiple test φ is said to control the FDR at level $\alpha \in (0, 1)$, if

$$\sup_{\vartheta\in\Theta} \mathrm{FDR}_\vartheta(\varphi) \le \alpha. \tag{1.3}$$

(i) We call a parameter value ϑ^* a least favourable parameter configuration (LFC) for the FWER or the FDR, respectively, of a given multiple test φ, if ϑ^* yields the supremum in (1.1) or (1.3), respectively.

The following lemma, though obvious, will be useful for the construction of closed test procedures, see Sect. 3.3.

Lemma 1.1. *Let $(\mathscr{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta}, \mathscr{H})$ denote a multiple test problem and $\varphi = (\varphi_i : i \in I)$ a multiple test for $(\mathscr{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta}, \mathscr{H})$.*

(a) Strong FWER control of φ implies weak FWER control of φ.
(b) Assume that φ controls the FWER weakly at level α. Then, a level α test for the (single) global hypothesis H_0 is given by the following rule: Reject H_0 if there exists an $i \in I$ such that $\varphi_i(x) = 1$.

For the relationships between the FWER, the FDR, and the pFDR, the following assertions hold true.

Lemma 1.2 (Relationships between FWER, FDR and pFDR). *Under the assumptions of Definition 1.2, we get:*

(a) $FDR_\vartheta(\varphi) = pFDR_\vartheta(\varphi)\mathbb{P}_\vartheta(R_m > 0)$.
(b) If $m_0(\vartheta) = m$, then $FDR_\vartheta(\varphi) = FWER_\vartheta(\varphi)$.
(c) For any $\vartheta \in \Theta$, it holds $FDR_\vartheta(\varphi) \le FWER_\vartheta(\varphi)$.

Proof. To prove part (a), we calculate straightforwardly

$$\begin{aligned}
\mathrm{FDR}_\vartheta(\varphi) &= \mathbb{E}_\vartheta\left[\frac{V_m}{R_m \vee 1}\right] \\
&= \mathbb{E}_\vartheta\left[\frac{V_m}{R_m \vee 1} | R_m > 0\right]\mathbb{P}_\vartheta(R_m > 0) \\
&\quad + \mathbb{E}_\vartheta\left[\frac{V_m}{R_m \vee 1} | R_m = 0\right]\mathbb{P}_\vartheta(R_m = 0) \\
&= \mathrm{pFDR}_\vartheta(\varphi)\mathbb{P}_\vartheta(R_m > 0) + 0.
\end{aligned}$$

For the proof of part (b), we notice that, if $m_0 = m$, $V_m = R_m$. Hence, $\mathrm{pFDR}_\vartheta(\varphi) \equiv 1$ in this case and, making use of part (a),

$$\mathrm{FDR}_\vartheta(\varphi) = \mathbb{P}_\vartheta(R_m > 0) = \mathbb{P}_\vartheta(V_m > 0) = \mathrm{FWER}_\vartheta(\varphi).$$

In the general case, we easily verify that $\mathrm{FDP}_\vartheta(\varphi) \le \mathbf{1}_{\{V_m>0\}}$. Thus, $\mathbb{E}_\vartheta[\mathrm{FDP}_\vartheta(\varphi)] \le \mathbb{E}_\vartheta\left[\mathbf{1}_{\{V_m>0\}}\right]$, which is equivalent to the assertion of part (c). □

Notice that the proof of part (b) of Lemma 1.2 implies that the pFDR cannot be controlled in the frequentist sense. The pFDR is only useful in Bayesian considerations (cf., e.g., Chap. 6). Throughout the remainder of this work, we will restrict our attention to the type I error rates defined in Definition 1.2. This is mainly due to