

# TESTING STATISTICAL HYPOTHESES OF EQUIVALENCE AND NONINFERIORITY

## SECOND EDITION

STEFAN WELLEK



CRC Press

Taylor & Francis Group

Boca Raton London New York

---

CRC Press is an imprint of the  
Taylor & Francis Group an **informa** business  
A CHAPMAN & HALL BOOK

MATLAB® is a trademark of The MathWorks, Inc. and is used with permission. The MathWorks does not warrant the accuracy of the text or exercises in this book. This book's use or discussion of MATLAB® software or related products does not constitute endorsement or sponsorship by The MathWorks of a particular pedagogical approach or particular use of the MATLAB® software.

Chapman & Hall/CRC  
Taylor & Francis Group  
6000 Broken Sound Parkway NW, Suite 300  
Boca Raton, FL 33487-2742

© 2010 by Taylor and Francis Group, LLC  
Chapman & Hall/CRC is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

Printed in the United States of America on acid-free paper  
10 9 8 7 6 5 4 3 2 1

International Standard Book Number: 978-1-4398-0818-4 (Hardback)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access [www.copyright.com](http://www.copyright.com) (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

**Trademark Notice:** Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

---

**Library of Congress Cataloging-in-Publication Data**

---

Wellek, Stefan.

Testing statistical hypothesis of equivalence and noninferiority / Stefan Wellek. -- 2nd ed.

p. cm.

Includes bibliographical references and index.

ISBN 978-1-4398-0818-4 (hardcover : alk. paper)

1. Statistical hypothesis testing. I. Title.

QA277.W46 2010

519.5'6--dc22

2010017666

---

Visit the Taylor & Francis Web site at  
<http://www.taylorandfrancis.com>

and the CRC Press Web site at  
<http://www.crcpress.com>

*To Brigitte*

---

# *Contents*

<b>Preface</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Statistical meaning of the concepts of equivalence and noninferiority . . . . .	1
1.2 Demonstration of equivalence as a basic problem of applied statistics . . . . .	2
1.3 Major fields of application of equivalence tests . . . . .	5
1.3.1 Comparative bioequivalence trials . . . . .	5
1.3.2 Clinical trials involving an active control . . . . .	7
1.3.3 Preliminary tests for checking assumptions underlying other methods of statistical inference . . . . .	8
1.4 Role of equivalence/noninferiority studies in current medical research . . . . .	8
1.5 Formulation of hypotheses . . . . .	10
1.6 Choosing the main distributional parameter . . . . .	13
1.7 Numerical specification of the limits of equivalence . . . . .	15
<b>2 General techniques for dealing with noninferiority problems</b>	<b>19</b>
2.1 Standard solution in the case of location parameter families . . . . .	19
2.1.1 Paired observations . . . . .	19
2.1.2 Two independent samples . . . . .	22
2.1.3 Power and sample size calculation based on tests for noninferiority under location-shift models . . . . .	23
2.2 Methods of constructing exact optimal tests for settings beyond the location-shift model . . . . .	24
2.3 Large-sample solutions for problems inaccessible for exact constructions . . . . .	26
2.4 Objective Bayesian methods . . . . .	27
2.5 Improved nonrandomized tests for discrete distributions . . . . .	28
2.6 Relationship between tests for noninferiority and two-sided equivalence tests . . . . .	30
2.7 Halving alpha? . . . . .	31

<b>3 General approaches to the construction of tests for equivalence in the strict sense</b>	<b>33</b>
3.1 The principle of confidence interval inclusion . . . . .	33
3.2 Bayesian tests for two-sided equivalence . . . . .	36
3.3 The classical approach to deriving optimal parametric tests for equivalence hypotheses . . . . .	40
3.4 Construction of asymptotic tests for equivalence . . . . .	45
<b>4 Equivalence tests for selected one-parameter problems</b>	<b>49</b>
4.1 The one-sample problem with normally distributed observations of known variance . . . . .	49
4.2 Test for equivalence of a hazard rate to some given reference value with exponentially distributed survival times . . . . .	55
4.3 Testing for equivalence of a single binomial proportion to a fixed reference success probability . . . . .	59
4.4 Confidence-interval inclusion rules as asymptotically UMP tests for equivalence . . . . .	64
4.5 Noninferiority analogues of the tests derived in this chapter .	68
<b>5 Equivalence tests for designs with paired observations</b>	<b>71</b>
5.1 Sign test for equivalence . . . . .	71
5.2 Equivalence tests for the McNemar setting . . . . .	76
5.2.1 Large-sample solution . . . . .	78
5.2.2 Corrected finite-sample version of the large-sample test . . . . .	81
5.2.3 Modifications for the noninferiority case . . . . .	85
5.3 Paired <i>t</i> -test for equivalence . . . . .	92
5.4 Signed rank test for equivalence . . . . .	99
5.5 A generalization of the signed rank test for equivalence for noncontinuous data . . . . .	108
<b>6 Equivalence tests for two unrelated samples</b>	<b>119</b>
6.1 Two-sample <i>t</i> -test for equivalence . . . . .	119
6.2 Mann-Whitney test for equivalence . . . . .	126
6.3 Two-sample equivalence tests based on linear rank statistics .	136
6.4 A distribution-free two-sample equivalence test allowing for arbitrary patterns of ties . . . . .	150
6.5 Testing for dispersion equivalence of two Gaussian distributions . . . . .	164
6.6 Equivalence tests for two binomial samples . . . . .	172
6.6.1 Exact Fisher type test for noninferiority with respect to the odds ratio . . . . .	172
6.6.2 Improved nonrandomized tests for noninferiority with respect to the odds ratio . . . . .	177

6.6.3	Tests for noninferiority using alternative parametrizations . . . . .	181
6.6.4	Exact test for two-sided equivalence with respect to the odds ratio . . . . .	186
6.6.5	An improved nonrandomized version of the UMPU test for two-sided equivalence . . . . .	193
6.6.6	Tests for two-sided equivalence with respect to the difference of success probabilities . . . . .	194
6.7	Log-rank test for equivalence of two survivor functions . . . . .	202
6.7.1	Rationale of the log-rank test for equivalence in the two-sided sense . . . . .	203
6.7.2	Power approximation and sample size calculation for the log-rank test for equivalence . . . . .	210
6.7.3	Log-rank test for noninferiority . . . . .	214
<b>7</b>	<b>Multisample tests for equivalence</b>	<b>219</b>
7.1	The intersection-union principle as a general solution to multisample equivalence problems . . . . .	219
7.2	<i>F</i> -test for equivalence of $k$ normal distributions . . . . .	221
7.3	Modified studentized range test for equivalence . . . . .	225
7.4	Testing for dispersion equivalence of more than two Gaussian distributions . . . . .	227
7.5	A nonparametric $k$ -sample test for equivalence . . . . .	231
<b>8</b>	<b>Equivalence tests for multivariate data</b>	<b>235</b>
8.1	Equivalence tests for several dependent samples from normal distributions . . . . .	235
8.1.1	Generalizing the paired <i>t</i> -test for equivalence by means of the $T^2$ -statistic . . . . .	235
8.1.2	A many-one equivalence test for dependent samples based on Euclidean distances . . . . .	241
8.1.3	Testing for equivalence with dependent samples and an indifference zone of rectangular shape . . . . .	248
8.1.4	Discussion . . . . .	252
8.2	Multivariate two-sample tests for equivalence . . . . .	253
8.2.1	A two-sample test for equivalence based on Hotelling's $T^2$ . . . . .	253
8.2.2	Behavior of the two-sample $T^2$ -test for equivalence under heteroskedasticity . . . . .	259
8.2.3	Multivariate two-sample tests for equivalence regions of rectangular shape . . . . .	262
<b>9</b>	<b>Tests for establishing goodness of fit</b>	<b>265</b>
9.1	Testing for equivalence of a single multinomial distribution with a fully specified reference distribution . . . . .	265

9.2	Testing for approximate collapsibility of multiway contingency tables . . . . .	270
9.3	Establishing goodness of fit of linear models for normally distributed data . . . . .	278
9.3.1	An exact optimal test for negligibility of interactions in a two-way ANOVA layout . . . . .	278
9.3.2	Establishing negligibility of carryover effects in the analysis of two-period crossover trials . . . . .	284
9.4	Testing for approximate compatibility of a genotype distribution with the Hardy-Weinberg condition . . . . .	290
9.4.1	Genetical background, measures of HW disequilibrium	290
9.4.2	Exact conditional tests for absence of substantial disequilibrium . . . . .	293
9.4.3	Confidence-interval-based procedures for assessing HWE . . . . .	303
<b>10</b>	<b>The assessment of bioequivalence</b>	<b>311</b>
10.1	Introduction . . . . .	311
10.2	Methods of testing for average bioequivalence . . . . .	315
10.2.1	Equivalence with respect to nonstandardized mean bioavailabilities . . . . .	315
10.2.2	Testing for scaled average bioequivalence . . . . .	322
10.3	Individual bioequivalence: Criteria and testing procedures . . . . .	325
10.3.1	Introduction . . . . .	325
10.3.2	Distribution-free approach to testing for probability-based individual bioequivalence . . . . .	327
10.3.3	An improved parametric test for probability-based individual bioequivalence . . . . .	329
10.4	Approaches to defining and establishing population bioequivalence . . . . .	340
10.4.1	Introduction . . . . .	340
10.4.2	A testing procedure for establishing one-sided bioequivalence with respect to total variability . . . . .	342
10.4.3	Complete disaggregate testing procedure and illustrating example . . . . .	344
10.4.4	Some results on power and sample sizes . . . . .	345
10.4.5	Discussion . . . . .	347
10.5	Bioequivalence assessment as a problem of comparing bivariate distributions . . . . .	348
<b>11</b>	<b>Tests for relevant differences between treatments</b>	<b>355</b>
11.1	Introduction . . . . .	355
11.2	Exploiting the duality between testing for two-sided equivalence and existence of relevant differences . . . . .	356

11.3	Solutions to some special problems of testing for relevant differences . . . . .	359
11.3.1	One-sample problem with normally distributed data of known variance . . . . .	359
11.3.2	Two-sample $t$ -test for relevant differences . . . . .	361
11.3.3	Exact Fisher type test for relevant differences between two binomial distributions . . . . .	364
<b>Appendix A Basic theoretical results</b>		<b>369</b>
A.1	UMP tests for equivalence problems in STP <sub>3</sub> families . . . . .	369
A.2	UMPU equivalence tests in multiparameter exponential families . . . . .	375
A.3	A sufficient condition for the asymptotic validity of tests for equivalence . . . . .	376
<b>Appendix B List of special computer programs</b>		<b>379</b>
<b>Appendix C Frequently used special symbols and abbreviations</b>		<b>383</b>
<b>References</b>		<b>387</b>
<b>Author index</b>		<b>403</b>
<b>Subject index</b>		<b>407</b>

---

## *Preface*

During the time since finalization of the manuscript of the first edition of this book, research in the field of equivalence testing methods expanded at an unexpectedly fast rate so that there seems to be a considerable need for updating its coverage. Furthermore, in clinical research, there developed an increasing preference for replacing trials following the classical placebo-controlled design with active-control trials requiring methods of testing for noninferiority rather than equivalence in the strict, i.e., two-sided sense. On the one hand, noninferiority problems are nothing else but generalized one-sided testing problems in the usual sense arising from a shift of the upper bound set under the null hypothesis to the parameter of interest away from zero or unity. Furthermore, from a technical point of view, the modifications required for transforming a test for two-sided equivalence into a test for noninferiority for the same setting are largely straightforward. On the other hand, it cannot be ignored that a book on the topic is likely to better serve the needs of readers mainly interested in applications when for each specific scenario the noninferiority version of the testing procedure is also described in full detail. Another extension of considerable interest for research workers in a multitude of empirical areas refers to testing for “relevant differences” between treatments or experimental conditions. Testing problems of this latter kind are dual to two-sided equivalence problems in that the assumption of nonexistence of differences of clinical or otherwise practical relevance plays the role of the null hypothesis to be as-

sessed. The new edition discusses solutions to such problems in an additional chapter.

Roughly speaking, tests for equivalence in the strict sense provide the adequate answer to the most natural question of how to proceed in a traditional two-sided testing problem if it turns out that primary interest is in verifying rather than rejecting the null hypothesis. Put in more technical terms, equivalence assessment deals with a particular category of testing problems characterized by the fact that the alternative hypothesis specifies a sufficiently small neighborhood of the point in the space of the target parameter which indicates perfect coincidence of the distributions to be compared.

The relevance of inferential procedures which, in the sense of this notion, enable one to “prove the null hypothesis” for many areas of applied statistical data analysis, is obvious enough. A particularly striking phenomenon which demonstrates the real need for such methods, is the adherence of generations of authors to using the term “goodness-of-fit tests” for methods which are actually tailored for solving the reverse problem of establishing absence or lack of fit. From a “historical” perspective (the first journal article on an equivalence test appeared as late as in the sixties of the twentieth century), the interest of statistical researchers in equivalence assessment was almost exclusively triggered by the introduction of special approval regulations for so-called generic drugs by the Food and Drug Administration (FDA) of the U.S. as well as the drug regulation authorities of many other industrial countries. Essentially, these regulations provide that the positive result of a test, which enables one to demonstrate with the data obtained from a so-called comparative bioavailability trial the equivalence of the new generic version of a drug to the primary manufacturer’s formulation, shall be accepted as a sufficient condition for approval of the generic formulation to the market. The overwhelming practical importance of the entailed problems of bioequivalence assessment (drugs whose equivalence with respect to the measured bioavailabilities can be taken for granted, are termed “bioequivalent” in clinical pharmacology literature), arises mainly out of quantity: Nowadays, at least half of the prescription drug units sold in the leading industrial countries are generic drugs that have been approved to be marketed on the basis of some bioequivalence trial.

Considerations of one-sided equivalence (noninferiority) play an increasingly important role in the design and analysis of genuine clinical trials of therapeutic methods. The subjects recruited for such trials are patients suffering from some disease rather than healthy volunteers. Whenever well-established therapeutic strategies of proven efficacy and tolerability are already available for the disease under consideration, it would be unethical to launch a new trial involving a negative control (in particular, placebo). From the statistical perspective, using a positive or active control instead frequently implies that a classical procedure tailored for establishing superiority of the experimental treatment over the control condition has to be replaced with the corresponding test for noninferiority.

Although noninferiority testing is given much more attention in the new

book as compared to the first edition, the core of this monograph still deals with methods of testing for equivalence in the strict, i.e., two-sided sense. The spectrum of specific equivalence testing problems of both types it covers range from the one-sample problem with normally distributed observations of fixed known variance (which will serve as the basis for the derivation of asymptotic equivalence tests for rather complex multiparameter and even semi- and nonparametric models), to problems involving several dependent or independent samples and multivariate data. A substantial part of the testing procedures presented here satisfy rather strong optimality criteria, which is to say that they maximize the power of detecting equivalence uniformly over a large class of valid tests for the same (or an asymptotically equivalent) problem. In equivalence testing, the availability of such optimal procedures seems still more important than in testing conventional one- or two-sided hypotheses. The reason is that even those equivalence tests which can be shown to be uniformly most powerful among all valid tests of the same hypotheses, turn out to require much higher sample sizes in order to maintain some given bounds on both types of error risks than do ordinary one- or two-sided tests for the same statistical models, unless one starts from an extremely liberal specification of the equivalence limits.

The theoretical basis of the construction of optimal tests for interval hypotheses was laid within the mathematical statistics literature of the nineteen fifties. However, up to now the pertinent results have only rarely been exploited in the applied, in particular the biostatistical, literature on equivalence testing. In a mathematical appendix to this book, they will be presented in a coherent way and supplemented with a corollary which allows great simplification of the computation of the critical constants of optimal equivalence tests under suitable symmetry restrictions. An additional appendix contains a listing of all computer programs supplied at the URL <http://www.crcpress.com/product/isbn/9781439808184> for facilitating as much as possible the routine application of all testing procedures discussed in the book. The collection of all program files contained in that directory is referenced as the **WKTSEQ2 Source Code Package** throughout the text. In contrast to the Web material which accompanied the first edition, the majority of the programs have now been made available also as R scripts or shared objects which can be called within the R system. Most of the concrete numerical examples given in the text for purposes of illustrating the individual methods, are taken from the author's own field of application, i.e., from medical research.

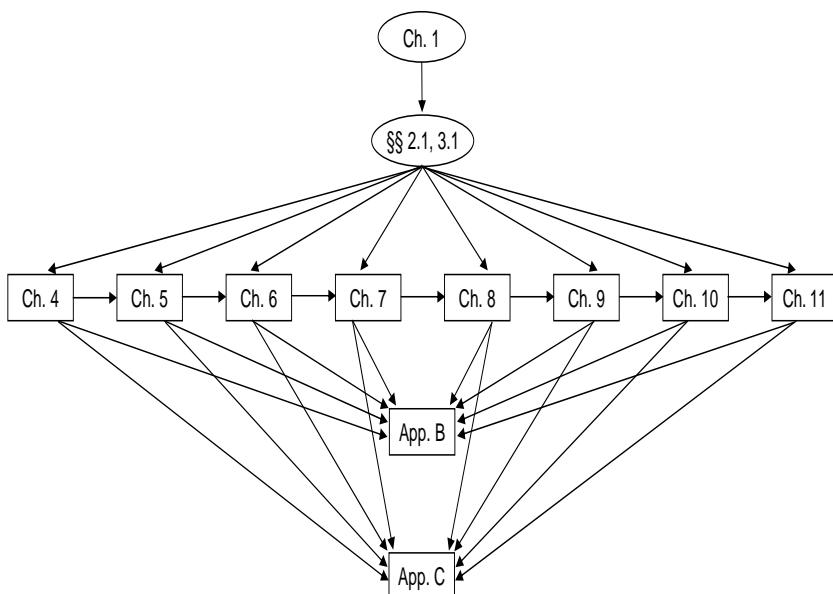
The book can be used in several ways depending on the reader's interests and level of statistical background. Chapter 1 gives a general introduction to the topic and should at least be skimmed by readers of any category. Chapters 2 and 3 deal with general approaches to problems of testing for noninferiority and two-sided equivalence and are mainly intended for readers with interests in a systematic account of equivalence testing procedures and their mathematical basis. Readers seeking information about specific procedures and their

practical implementation are advised to skip these chapters except for the sections on noninferiority testing in location-shift models (§ 2.1) and confidence interval inclusion rules (§ 3.1).

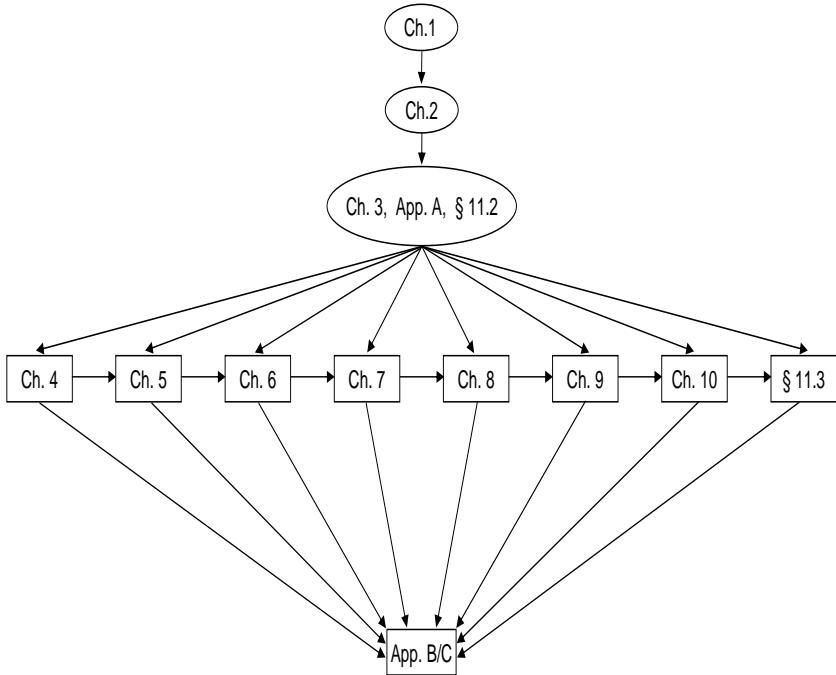
Apart from occasional cross-references, all remaining chapters (and even individual sections of them) can be read independently of each other. The material they contain is to provide the working statistician of any background and level of sophistication with a sufficiently rich repertoire of efficient solutions to specific equivalence and noninferiority testing problems frequently encountered in the analysis of real data sets. Except for Section 4.4 which introduces no additional testing procedure, Chapters 4–10 should even be suited for serving as a procedure reference book in the field of equivalence and noninferiority testing. The last chapter summarizes both some basic theoretical results about tests for relevant differences (arising from switching the roles of both hypotheses in a two-sided equivalence problem) and describes solutions for some specific settings frequently arising in practice. In order to keep the page count within reasonable limits, the coverage in this book is confined to methods for samples of fixed size. Fully and group sequential methods for equivalence testing problems are left out of account.

All in all, one of the following alternative guides to the book should be followed:

A) [for readers primarily interested in practical applications]



- B) [for the reader particularly interested in theory and mathematical background]



I have many people to thank for helping me in the endeavor of preparing this book, without being able to mention here more than a few of them. Niels Keiding played an initiating role, not only by encouraging me to make the material contained in a book on the same topic I had published in 1994 in German accessible to an international readership, but also by bringing me in contact with Chapman & Hall/CRC. Cooperation with the staff of this publisher proved remarkably smooth and constructive, and I would like to acknowledge in particular the role of Rob Calver as the present statistics editor at Taylor & Francis during the whole phase until finalizing the manuscript of the new edition. The former vice-president of Gustav Fischer Verlag Stuttgart, Dr. Wolf D. von Lucius, is gratefully acknowledged for having given his permission to utilize in part the content of my book of 1994. As was already the case with the first edition, there are two people from the staff in my de-

partment at the Central Institute of Mental Health at Mannheim to whom I owe a debt of gratitude: Mireille Lukas spent her expert skills in handling the L<sup>A</sup>T<sub>E</sub>X system on typesetting the new parts of the book and reorganizing the whole document. The editorial component of my job as the author was greatly facilitated by the fact that I could delegate to her a considerable part of the work entailed in compiling the bibliography and both indices by means of special T<sub>E</sub>X-based tools. Peter Ziegler took over the task of making available within R more than 30 computer programs originally written in Fortran or SAS. Moreover, he generated the better part of the figures contained in the book writing suitable source scripts in SAS/GRAFH. Last but not least, special thanks are due to the following two colleagues from external departments in statistics and related fields: Arnold Janssen (University of Düsseldorf) for agreeing to present results from an unpublished joint paper by him and myself, and Andreas Ziegler (University at Lübeck) for the fruitful cooperation on the topic of Chapter 9.4

Mannheim  
January 2010

STEFAN WELLEK

## Disclaimers

1. All computer programs included with this book are provided in good faith and after a reasonable amount of validation. However, the author, publishers and distributors do not guarantee their accuracy and take no responsibility for the consequences of their use.
2. MATLAB® is a registered trademark of The MathWorks, Inc. For product information, please contact:

The MathWorks, Inc.  
3 Apple Hill Drive  
Natick, MA 01760-2098 USA  
Tel: 508 647 7000  
Fax: 508-647-7001  
E-mail: [info@mathworks.com](mailto:info@mathworks.com)  
Web: [www.mathworks.com](http://www.mathworks.com)

# 1

---

## *Introduction*

---

### **1.1 Statistical meaning of the concepts of equivalence and noninferiority**

Although the notions of equivalence and noninferiority have nowadays become part of standard terminology of applied statistics, the precise meaning of these terms is not self-explanatory. The first of them is used in statistics to denote a weak or, more adequately speaking, fuzzy form of an identity relation referring to the distribution(s) which underly the data under analysis. The fuzziness of equivalence hypotheses as considered in this book is induced by enlarging the null hypothesis of the traditional two-sided testing problem referring to the same statistical setting, through adding an “indifference zone” around the corresponding region (or point) in the parameter space. In other words, *equivalence* means here *equality except for practically irrelevant deviations*. Such an indifference zone is a basic and necessary ingredient of any kind of testing problem to be addressed in the planning and confirmatory analysis of a study, trial or experiment run with the objective of demonstrating equivalence. Admittedly, finding a consensus on how to specify that indifference zone concretely is far from easy in the majority of applications. However, it is an indispensable step without which the testing problem the experimenter proposes would make no statistical sense at all. The reason behind this fact whose proper understanding is an elementary prerequisite for a sensible use of the methods discussed in this book, will be made precise in § 1.5.

Recalling the way the word noninferiority is used in everyday language provides little if any insight into the real meaning of the second of the concepts to be introduced here. The term has been originally coined in the clinical trials literature in order to denote a study which aims at demonstrating that some new, experimental therapy falls short in efficacy by a clinically acceptable amount at most as compared to a well-established reference treatment for the same disease. Thus, noninferiority means absence of a relevant difference in favor of the comparator against which the experimental treatment has to be assessed. Formalizing a noninferiority problem by translating it into a pair of statistical hypotheses leads to a generalized one-sided testing problem. The only difference to a standard testing problem of the one-sided type is that the common boundary of the two hypotheses is now shifted to the left, away from

the null which often, yet not always, coincides with the origin of the real line.

At the same time, noninferiority problems exhibit a clear-cut relationship to equivalence problems: Obviously, every equivalence hypothesis can be modified into a hypothesis of noninferiority simply by letting the right-hand limit (which in some cases will be a curve rather than a single point) increase to infinity. This justifies considering noninferiority as a one-sided form of equivalence. In the earlier literature in the field, some authors (for notable examples, see Dunnett and Gent, 1977; Mehta et al., 1984) did little care to distinguish between equivalence and noninferiority. In order to avoid potential confusion resulting from such usage, we will adhere to the following terminological rule: When referring to specific problems and procedures, equivalence per se will always used in the strict, two-sided sense of the term. Noninferiority problems will be either called that way or, alternatively, addressed as one-sided equivalence problems.

---

## 1.2 Demonstration of equivalence as a basic problem of applied statistics

It is a basic fact well known to every statistician that in any hypotheses testing problem there is an inherent logical asymmetry concerning the roles played by the two statements (traditionally termed null and alternative hypotheses) between which a decision shall be taken on the basis of the data collected in a suitable trial or experiment: Any valid testing procedure guarantees that the risk of deciding erroneously in favor of the alternative does not exceed some prespecified bound whereas the risk of taking a wrong decision in favor of the null hypothesis can typically be as high as 1 minus the significance level (i.e., 95% in the majority of practical applications). On the other hand, from an experimental or clinical researcher's perspective, there seems little reason why he should not be allowed to switch his views towards the problem under consideration and define what he had treated as the null hypothesis in a previous study, as the hypothesis of primary interest in a subsequent trial.

If, as is so often the case in practice, the “traditional” formulation of the testing problem has been a two-sided one specifying equality of the effects of, say, two treatments under the null hypothesis, then such a switch of research interest leads to designing a study which aims at proving absence of a (relevant) difference between both treatment effects, i.e., equivalence. The term treatment is used here in the generic sense covering also experimental conditions etc. being compared in a fully non-medical context. Typically, the rigorous construction of a testing procedure for the confirmatory analysis of an equivalence trial requires rather heavy mathematical machinery. Nevertheless, the basic idea leading to a logically sound formulation of an equivalence

testing problem and the major steps making up an appropriate statistical decision procedure can be illustrated by an example as simple as the following.

### Example 1.1

Of an antihypertensive drug which has been in successful use for many years, a new generic version has recently been approved to the market and started to be sold in the pharmacies at a price undercutting that of the reference formulation ( $R$ ) by about 40%. A group of experts in hypertension doubt the clinical relevance of existing data showing the *bioequivalence* of the generic to the reference formulation, notwithstanding the fact that these data had been accepted by the drug regulation authorities as sufficient for approving the new formulation. Accordingly, the hypertensionologists agree to launch into a clinical trial aiming at establishing the *therapeutic equivalence* of the new formulation of the drug. Instead of recruiting a control group of patients to be treated with  $R$ , one decides to base the assessment of the therapeutic equivalence of the new formulation on comparison to a fixed responder rate of 60% obtained from long-term experience with formulation  $R$ . Out of  $n = 125$  patients eventually recruited for the current trial, 56% showed a positive response in the sense of reaching a target diastolic blood pressure below 90 mmHg. Statistical assessment of this result was done by means of a conventional binomial test of the null hypothesis that the probability  $p$ , say, of obtaining a positive response in a patient given the generic formulation, equals the reference value  $p_o = .60$ , versus the two-sided alternative  $p \neq p_o$ . Since the significance probability (p-value) computed in this way turned out to be as high as .41, the researchers came to the conclusion that the therapeutic equivalence of the generic to the reference formulation could be taken for granted, implying that the basic requirement for switching to the new formulation whenever confining the costs of treatment is an issue, was satisfied.

Unfortunately, it follows from the logical asymmetry between null and alternative hypothesis mentioned at the beginning that such kind of reasoning misses the following point of fundamental importance: “Converting” a traditional two-sided test of significance by inferring equivalence of the treatments under comparison from a nonsignificant result of the former, generally fails to yield a valid testing procedure. In a word: *A nonsignificant difference must not be confused with significant homogeneity*, or, as Altman and Bland (1995) did put it, “*absence of evidence is not evidence of absence*.” Even in the extremely simple setting of the present example, i.e., of a one-arm trial conducted for the purpose of establishing (therapeutic) equivalence of a single treatment with regard to a binary success criterion, correct inference requires the application of a testing procedure exhibiting genuinely new features.

- (i) First of all, it is essential to notice that the problem of establishing the alternative hypothesis of *exact* equality of the responder rate  $p$  associated

with the generic formulation of the drug, to its reference value  $p_o = .60$  by means of a statistical test admits no sensible solution (the reader interested in the logical basis of this statement is referred to § 1.5). The natural way around this difficulty consists of *introducing a region of values of  $p$  close enough to the target value  $p_o$  for considering the deviations practically irrelevant*. For the moment, let us specify this region as the interval  $(p_o - .10, p_o + .10) = (.50, .70)$ . Hence, by equivalence of  $p$  to  $p_o$  we eventually mean a weakened form of identity specifying equality except for ignorable differences.

- (ii) The closer the observed responder rate  $X/n$  comes up to the target rate  $p_o = .60$ , the stronger the evidence in favor of equivalence provided by the available data. Thus, a reasonable test for equivalence of  $p$  to  $p_o$  will use a decision rule of the following form: The *null hypothesis of inequivalence is rejected if and only if* the difference  $X/n - p_o$  between the observed and the target responder rate falls between suitable critical bounds, say  $c_1$  and  $c_2$ , such that  $c_1$  is some negative and  $c_2$  some positive real number, respectively.
- (iii) Optimally, the rejection region of the desired test, i.e., the set of possible outcomes of the trial allowing a decision in favor of equivalence, should be defined in such a way that the associated requirement on the degree of closeness of the observed responder rate  $X/n$  to the reference rate  $p_o$  is as weak as possible without increasing the risk of an erroneous equivalence decision over  $\alpha$ , the prespecified level of significance (chosen to be .05 in the majority of practical applications).
- (iv) As follows from applying the results to be presented in § 4.3 with  $p_o \mp .10$  as the *theoretical range of equivalence* and at level  $\alpha = 5\%$ , the optimal critical bounds to  $X/n - p_o$  to be used in a test for equivalence of  $p$  to  $p_o = .60$  based on a random sample of size  $n = 125$  are given by  $c_1 = -2.4\%$  and  $c_2 = 3.2\%$ , respectively. Despite the considerable size of the sample recruited to the trial, the rejection interval for  $X/n - p_o$  corresponding to these values of  $c_1$  and  $c_2$  is pretty narrow, and the observed rate 56% of responders falls relatively far outside giving  $X/n - p_o = -4.0\%$ . Consequently, at significance level 5%, the data collected during the trial do not contain sufficient evidence in favor of equivalence of the generic to the reference formulation in the sense of  $|p - p_o| < .10$ .

The confirmatory analysis of experimental studies, clinical trials etc. which are performed in order to establish equivalence of treatments is only one out of many inferential tasks of principal importance which can adequately be dealt with only by means of methods allowing to establish the (suitably enlarged) null hypothesis of a conventional two-sided testing problem. Another category of problems for which exactly the same holds true, refers to the verification of statistical model assumptions of any kind. Notwithstanding the traditional

usage of the term “goodness-of-fit test” obscuring the fact that the testing procedures subsumed in the associated category are tailored for solving the reverse problem of establishing lack of fit, in the majority of cases it is much more important to positively demonstrate the compatibility of the model with observed data. Thus, if a goodness-of-fit test is actually to achieve what is implied by its name, then it has to be constructed as an equivalence test in the sense that a positive result supports the conclusion that the true distribution from which the data have been taken, except for minor discrepancies, coincides with the distribution specified by the model.

The primary objective of this book is a systematic and fairly comprehensive account of testing procedures for problems such that the *alternative hypothesis* specifies a sufficiently *small neighborhood* of the point in the space of the target parameter (or functional) which indicates perfect coincidence of the probability distributions under comparison. As will become evident from the numerical material presented in the subsequent chapters, the sample sizes required in an equivalence test in order to achieve a reasonable power typically tend to be considerably larger than in an ordinary one- or two-sided testing procedure for the same setting unless the range of tolerable deviations of the distributions from each other is chosen so wide that even distributions exhibiting pronounced dissimilarities would be declared “equivalent”. This is the reason why in equivalence testing optimization of the procedures with respect to power is by no means an issue of purely academic interest but a necessary condition for keeping sample size requirements within the limits of practicality. The theory of hypotheses testing as developed in the fundamental work of E.L. Lehmann having appeared a few years ago in a third edition (Lehmann and Romano, 2005) provides in full mathematical generality methods for the construction of optimal procedures for four basic types of testing problems covering equivalence problems in the sense of the present monograph as well. Converting these general results into explicit decision rules suitable for routine applications will be a major objective in the chapters to follow.

---

## 1.3 Major fields of application of equivalence tests

### 1.3.1 Comparative bioequivalence trials

It was not until the late nineteen sixties that statistical researchers started to direct some attention to methods of testing for equivalence of distributions in the sense made precise in the previous section. In this initial phase, work on equivalence assessment was almost exclusively triggered by the introduction of special approval regulations for so-called generic drugs by the Food and Drug Administration (FDA) of the U.S. as well as the drug regulation authorities of many other industrialized countries. Loosely speaking, a generic drug is an

imitation of a specific drug product of some primary manufacturer that has already been approved to the market and prescribed for therapeutic purposes for many years but is no longer protected by patent. As a matter of course, with regard to the biologically active ingredients, every such generic drug is chemically identical to the original product. However, the actual biological effect of a drug depends on a multitude of additional factors referring to the whole process of the pharmaceutical preparation of the drug. Examples of these are

- chemical properties and concentrations of excipients
- kind of milling procedure
- choice of tablet coatings
- time and strength of compression applied during manufacture.

For the approval of a generic drug, the regulatory authorities do not require evidence of therapeutic efficacy and tolerability based on comparative clinical trials. Instead, it is considered sufficient that in a trial on healthy volunteers comparing the generic to the original formulation of the drug, the hypothesis of absence of relevant differences in basic pharmacokinetic characteristics (called “measures of bioavailability”) can be established. If this is the case, then the generic drug is declared equivalent with respect to bioavailability or, for short, bioequivalent to the original formulation.

Assessment of bioequivalence between a new and a reference formulation of some drug is still by far the largest field of application for statistical tests of the type this book focusses upon, and can be expected to keep holding this position for many years to come. The overwhelming importance of the problem of bioequivalence assessment has to do much more with economic facts and public health policies than with truly scientific interest: During the last two decades, the market share of generic drugs has been rising in the major industrial countries to levels ranging between 43% (U.S., 1996) and 67.5% (Germany, 1993)! From the statistical perspective, the field of bioequivalence assessment is comparatively narrow. In fact, under standard model assumptions [to be made explicit in Ch. 10], the confirmatory analysis of a prototypical bioequivalence study reduces to a comparison of two Gaussian distributions. In view of this, it is quite misleading that equivalence testing is still more or less identified with *bioequivalence* assessment by many (maybe even the majority) of statisticians. As will hopefully become clear enough from further reading of the present monograph, problems of equivalence assessment are encountered in virtually every context where the application of the methodology of testing statistical hypotheses makes any sense at all. Accordingly, it is almost harder to identify a field of application of statistics where equivalence problems play no or at most a minor role, than to give reasons why they merit particular attention in some specific field.

### 1.3.2 Clinical trials involving an active control

In medical research, clinical trials which involve an active (also called positive) control make up the second largest category of studies commonly analyzed by means of equivalence testing methods. An active rather than negative control (typically placebo) is used in an increasing number of clinical trials referring to the treatment of diseases for which well-established therapeutic strategies of proven efficacy and tolerability already exist. Under such circumstances it would be clearly unethical to leave dozens or hundreds of patients suffering from the respective disease without any real treatment until the end of the study. What has to be and is frequently done instead, is replacing the traditional negative control by a group to which the best therapy having been in use up to now, is administered. Usually, it is not realistic to expect then that the group which is given the new treatment will do still better than the control group with respect to efficacy endpoints. In return, the experimental therapy is typically known in advance to have much better tolerability so that its use can and should be recommended as soon as there is convincing evidence of equivalent efficacy. A particularly important example are trials of modifying adjuvant chemotherapy regimes well established in oncology, by reducing dosages and/or omitting the most toxic of the substances used. For such a reduced regime, superiority with respect to tolerability can be taken for granted without conducting any additional trial at all, and it is likewise obvious that demonstrating merely noninferiority with regard to efficacy would entail a valuable success.

In the clinical trials methodology literature, it has sometimes been argued (cf. Windeler and Trampisch, 1996) that tests for equivalence in the strict, i.e., two-sided sense are *generally* inappropriate for an active-control study and should always be replaced by one-sided equivalence tests or tests for noninferiority. In contrast, we believe that there are several convincing points (not to be discussed in detail here) for the view that the question whether a one- or a two-sided formulation of the equivalence hypothesis eventually to be tested is the appropriate one, should be carefully discussed with the clinicians planning a specific active-control trial rather than decided by biostatistical decree once and for all.

An undisputed major difference between clinical trials involving an active control, and comparative bioavailability studies (the consensus about the adequacy of two-sided equivalence tests for the confirmatory analysis of the latter has never been seriously challenged) refers to the structure of the distributions which the variables of primary interest typically follow: Quite often, the analysis of an active-control trial has to deal with binomial proportions [→ § 6.6] or even empirical survivor functions computed from partially censored observations [→ § 6.7] rather than with means and variances determined from samples of normally distributed observations.

### 1.3.3 Preliminary tests for checking assumptions underlying other methods of statistical inference

Looking through statistical textbooks of virtually all kinds and levels of sophistication, it is hard to find any which does not give at least some brief account of methods for checking the assumptions that the most frequently used inferential procedures have to rely upon. All of them approach this basic problem from the same side: The testing procedures provided are tests of the null hypothesis that the assumptions to be checked hold true, versus the alternative hypothesis that they are violated in one way or the other. Since the aim a user of such a preliminary test commonly has in mind is to give evidence of the correctness of the required assumptions, one cannot but state that the usual approach is based on an inadequate formulation of the hypotheses. It is clear that equivalence tests in the sense of § 1.2 are exactly the methods needed for finding a way around this logical difficulty so that another potentially huge field of applications of equivalence testing methods comes within view.

One group of methods needed in this context are of course, tests for goodness rather than lack of fit since they allow in particular the verification of parametric distributional assumptions of any kind. Other important special cases covered by the methods presented in subsequent chapters refer to restrictions on nuisance parameters in standard linear models such as

- homoskedasticity [ $\rightarrow$  §§ 6.5, 7.4]
- additivity of main effects [ $\rightarrow$  § 9.3.1]
- identity of carryover effects in crossover trials [ $\rightarrow$  § 9.3.2].

Establishing goodness of fit by means of equivalence testing procedures is even an important issue in genetic epidemiology. This will be explained in detail in § 9.4 which is devoted to methods for assessing the validity of the Hardy-Weinberg assumption upon which some of the most basic and widely used techniques for the analysis of genetic association studies have to rely.

---

## 1.4 Role of equivalence/noninferiority studies in current medical research

The increasing relevance of both types of an equivalence study for current medical research is reflected in a number of facts comparatively easy to grasp from widely accessible sources and databases.

A well-accepted way of getting an objective basis for statements about the development of some specific area of scientific research is through a systematic search over the pertinent part of published literature. A rough summary

of results to be obtained following that line is presented in Figure 1.1 where the count of entries in the PubMed database containing a keyword indicating the use of equivalence/noninferiority testing methods in the respective paper, is plotted by calendar year ranging from 1990 through 2008. The keywords which were selected for that purpose were (i) *bioequivalence*, (ii) *non(-)inferiority study*, and (iii) *equivalence study*, with the parentheses around the hyphenation sign indicating that both spellings in use were covered.

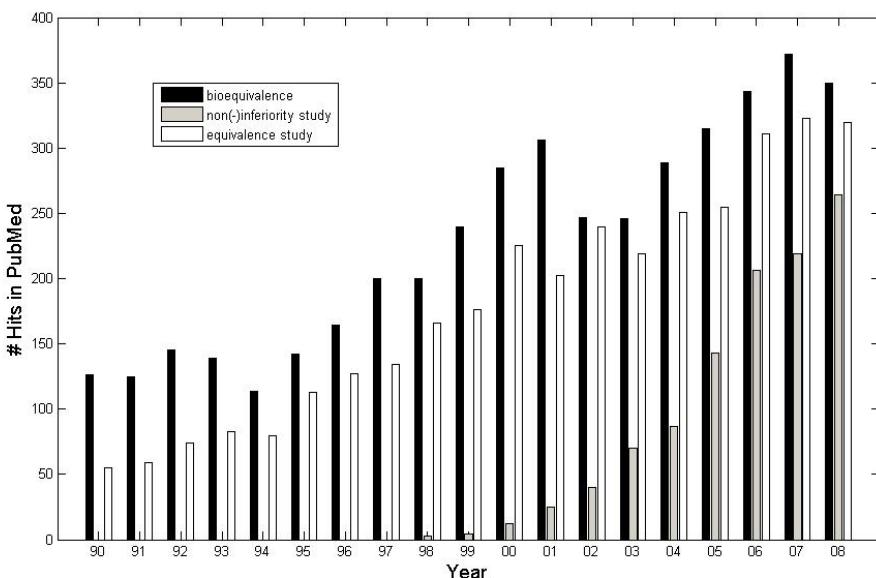


Figure 1.1 *Count of hits obtained by entering the keywords (i) bioequivalence, (ii) non(-)inferiority study, and (iii) equivalence study in PubMed, by year of publication.*

Admittedly, these figures must be interpreted with considerable caution, due to the unknown extent of the overlap between the scopes of the keywords of interest. Nevertheless, the following conclusions seem to be justified:

- The majority of published studies aiming to establish equivalence are comparative bioavailability trials as described in § 1.3.1. Since 1990, the number of bioequivalence-related publications has increased by about 200%.
- The time-lag in the quantitative development of published literature on studies devoted to establishing noninferiority as compared to equivalence, is likely to reflect mainly a change in terminological usage: Adoption of the term noninferiority by the medical community is a process which started only quite recently; before the beginning of the new millennium, systematic distinction between one- and two-sided equivalence

was lacking, and even nowadays, outside the area of bioequivalence, noninferiority and equivalence are often used more or less as synonyms.

- Since 2002, the number of publications dealing with methodological issues and results of equivalence/noninferiority studies involving patients rather than healthy volunteers has reached the same order of magnitude as the number of bioequivalence-related publications per year.

Although the proportion of clinical trials leading to publications in the medical sciences literature is hard to estimate, it is likely that the vast majority remain unpublished. Thus, even a better basis for assessing the relevance of the equivalence paradigm for the medical progress than searching for pertinent literature might be provided by looking at the proportion of prescription drugs which have been approved to the market due to positive results of equivalence trials. Unfortunately, the question of how large this proportion is, admits of a well-grounded answer only for generic drugs, namely 100%, simply by definition. However, even if the corresponding proportion of innovator drugs were as low as 10%, calculation of the overall rate for the U.S. would yield the following result: In the Orange Book, Version 12/2008 of the FDA (2008) 12,751 prescription drug products were listed of which only 3,154 are innovator products; assuming that for 10% of the latter, the positive approval decision was based on equivalence trials, the rate among all authorized prescription drugs is obtained to be  $100 \times (9,597 + 315)/12,751 \approx 78\%$ . This figure, which is likely to apply also to a number of other industrialized countries, gives sufficient evidence of the extent to which in our era, the medical sciences have to rely on a well-developed statistical methodology for the planning and analysis of studies conducted with the objective of establishing equivalence or noninferiority.

---

## 1.5 Formulation of hypotheses

As explained in nontechnical terms in Section 1.1, equivalence problems are distinguished from conventional testing problems by the form of the hypotheses to be established by means of the data obtained from the experiment or study under analysis. Typically [for an exception see § 10.3], the hypothesis formulation refers to some real-valued parameter  $\theta$  which provides a sensible measure of the degree of dissimilarity of the probability distributions involved. For example, in the specific case of a standard parallel group design used for the purpose of testing for equivalence of two treatments  $A$  and  $B$ , an obvious choice is  $\theta = \mu_1 - \mu_2$  with  $\mu_1$  and  $\mu_2$  denoting a measure of location for the distribution of the endpoint variable under  $A$  and  $B$ , respectively. The equivalence hypothesis whose compatibility with the data one wants to

assess, specifies that  $\theta$  is contained in a suitable neighborhood around some reference value  $\theta_o$  taken on by  $\theta$  if and only if the distributions under comparison are exactly equal. This neighborhood comprises those values of  $\theta$  whose distance from  $\theta_o$  is considered compatible with the notion of equivalence for the respective setting. It will be specified as an open interval throughout with endpoints denoted by  $\theta_o - \varepsilon_1$  and  $\theta_o + \varepsilon_2$ , respectively. Of course, both  $\varepsilon_1$  and  $\varepsilon_2$  are positive constants whose numerical values must be assigned *a priori*, i.e., without knowledge of the data under analysis. Specifically, in the case of the simple parallel group design with  $\theta = \mu_1 - \mu_2$ , the usual choice of  $\theta_o$  is  $\theta_o = 0$ , and the equivalence interval is frequently chosen symmetrical about  $\theta_o$ , i.e., in the form  $(-\varepsilon, \varepsilon)$ .

Accordingly, in this book, our main objects of study are statistical decision procedures which define a valid statistical test at some prespecified level  $\alpha \in (0, 1)$  of the *null hypothesis*

$$H : \theta \leq \theta_o - \varepsilon_1 \quad \text{or} \quad \theta \geq \theta_o + \varepsilon_2 \quad (1.1a)$$

of *nonequivalence*, versus the *equivalence assumption*

$$K : \theta_o - \varepsilon_1 < \theta < \theta_o + \varepsilon_2 \quad (1.1b)$$

as the *alternative hypothesis*. Such a decision rule has not necessarily to exhibit the form of a significance test in the usual sense. For example, it can and will [see § 3.2] also be given by a Bayes rule for which there is additional evidence that the “objective probability” of a false decision in favor of equivalence will never exceed the desired significance level  $\alpha$ . Bayesian methods for which we cannot be sure enough about this property taken for crucial from the frequentist point of view, are of limited use in the present context as long as the regulatory authorities to which drug approval applications based on equivalence studies have to be submitted, keep insisting on the maintenance of a prespecified significance level in the classical sense.

It is worth noticing that an equivalence hypothesis of the general form (1.1b) will never be the same as the null hypothesis  $H_0 : \theta = \theta_o$  of the corresponding two-sided testing problem, irrespective of what particular positive values are assigned to the constants  $\varepsilon_1$  and  $\varepsilon_2$ . In other words, switching attention from an ordinary two-sided to an equivalence testing problem entails not simply an exchange of both hypotheses involved but in addition a more or less far-reaching modification of them. Replacing the nondegenerate interval  $K$  of (1.1b) by the singleton  $\{\theta_o\}$  would give rise to a testing problem admitting of no worthwhile solution at all. In fact, in all families of distributions being of interest for concrete applications, the rejection probability of any statistical test is a *continuous* function of the target parameter  $\theta$ . But continuity of the power function  $\theta \mapsto \beta(\theta)$ , say, clearly implies, that the test can maintain level  $\alpha$  on  $\{\theta \neq \theta_o\}$  only if its power against  $\theta = \theta_o$  exceeds  $\alpha$  neither. Consequently, if we tried to test the null hypothesis  $\theta \neq \theta_o$  against the alternative  $\theta = \theta_o$ , we would not be able to replace the trivial “test” rejecting the null hypothesis independently of the data with probability  $\alpha$ , by a useful decision rule.

The inferential problems to be treated in the subsequent chapters under the heading of noninferiority assessment share two basic properties with equivalence testing problems in the strict sense. In the first place, they likewise arise from modifying the hypotheses making up some customary type of testing problem arising very frequently in routine data analysis. In the second place, modification of hypotheses again entails the introduction of a region in the space of the target distributional parameter  $\theta$  within which the difference between the actual value of  $\theta$  and its reference value  $\theta_0$  is considered practically irrelevant. However, there remains one crucial difference of considerable importance for the correct interpretation of the results eventually established by means of the corresponding testing procedures, as well as the mathematical treatment of the testing problems: The region of tolerable discrepancies between  $\theta$  and  $\theta_0$  is now bounded to below only whereas excesses in value of  $\theta$  over  $\theta_0$  of arbitrary magnitude are considered acceptable or even desirable. In other words, the testing procedures required in this other context have to enable the experimenter to make it sufficiently sure that the experimental treatment  $A$ , say, is not substantially inferior to some standard treatment  $B$ , without ruling out the possibility that  $A$  may even do considerably better than  $B$ . In contrast, for an equivalence trial in the strict sense made precise before, the idea is constitutive that one may encounter hypo- as well as hyperefficacy of the new as compared to the standard treatment and that protecting oneself against both forms of a substantial dissimilarity between  $A$  and  $B$  is a definite requirement.

Formally speaking, the crucial difference between equivalence testing and testing for absence of substantial inferiority is that in the latter type of problem the right-hand boundary  $\theta_0 + \varepsilon_2$  of the equivalence interval is replaced with  $+\infty$  or, in cases where the parameter space  $\Theta$  of  $\theta$  is bounded to the right itself, by  $\theta^* = \sup \Theta$ . The corresponding hypothesis testing problem reads

$$H_1 : \theta \leq \theta_0 - \varepsilon \quad \text{versus} \quad K_1 : \theta > \theta_0 - \varepsilon \quad (1.2)$$

with sufficiently small  $\varepsilon > 0$ .

From a mathematical point of view, the direction of the shift of the common boundary of an ordinary one-sided testing problem does not matter. In fact, approaches well suited for the construction of tests for one-sided equivalence in the sense of (1.2) can also be used for the derivation of tests for one-sided problems with a boundary of hypotheses shifted to the right, and vice versa. If  $\theta$  keeps denoting a meaningful measure for the extent of superiority of a new treatment  $A$  over some standard treatment  $B$  and  $\theta = \theta_0$  indicates identity in effectiveness of both treatments, testing  $\theta \leq \theta_0 + \varepsilon$  versus  $\theta > \theta_0 + \varepsilon$  rather than  $\theta \leq \theta_0 - \varepsilon$  versus  $\theta > \theta_0 - \varepsilon$  makes sense whenever one wants to ensure that a significant result of the corresponding test indicates that replacing  $A$  by  $B$  entails a relevant improvement. As pointed out by Victor (1987) this holds true for the majority of clinical trials aiming at giving evidence of treatment differences rather than equivalence.

---

## 1.6 Choosing the main distributional parameter

Except for single-parameter problems, the scientific relevance of the result of an equivalence testing procedure highly depends on a careful and sensible choice of the target parameter  $\theta$  [recall (1.1a), (1.1b) and (1.2)] in terms of which the hypotheses have been formulated. The reason is that, in contrast to the corresponding conventional testing problems with the common boundary of null and alternative hypothesis being given by zero, equivalence problems remain generally not invariant under redefinitions of the main distributional parameter. A simple, yet practically quite important example which illustrates this fact, is the two-sample setting with binomial data. If we denote the two unknown parameters in the usual way, i.e., by  $p_1$  and  $p_2$ , and define  $\delta$  and  $\rho$  as the difference  $p_1 - p_2$  and the odds ratio  $p_1(1 - p_2)/((1 - p_1)p_2)$ , respectively, then the null hypotheses  $\delta = 0$  and  $\rho = 1$  correspond of course to exactly the same subset in the space  $[0, 1] \times [0, 1]$  of the primary parameter  $(p_1, p_2)$ . On the other hand, the set  $\{(p_1, p_2) | -\delta_1 < \delta < \delta_2\}$  will be different from  $\{(p_1, p_2) | 1 - \varepsilon_1 < \rho < 1 + \varepsilon_2\}$  for *any* choice of the constants  $0 < \delta_1, \delta_2 < 1$  and  $0 < \varepsilon_1 < 1, \varepsilon_2 > 0$  determining the equivalence limits under both specifications of the target parameter.

On the one hand, there are no mathematical or otherwise formal criteria leading to a unique answer to the question about the appropriate choice of the parameter of main interest for purposes of formulating equivalence hypotheses for a given model or setting. On the other, this is by no means a matter of purely subjective taste but in numerous cases there are convincing arguments for preferring a specific parametrization to an alternative one, with the discrimination between the difference of the responder rates and the odds ratio in the binomial two-sample setting giving an interesting case in point. Although simplicity and ease of interpretability even for the mathematically less educated user clearly speak in favor of the difference  $\delta$ , plotting the regions corresponding to the two equivalence hypotheses  $-\delta_1 < \delta < \delta_2$  and  $1 - \varepsilon_1 < \rho < 1 + \varepsilon_2$  as done in Figure 1.2, shows to the contrary that defining equivalent binomial distributions in terms of  $p_1 - p_2$  entails a serious logical flaw: Whereas the hypothesis of equivalence with respect to the odds ratio corresponds to a proper subset of the parameter space for  $(p_2, \delta)$ , the range of  $\delta$ -coordinates of points equivalent to 0 in the sense of the first hypothesis formulation, is distinctly beyond the limits imposed by the side conditions  $-p_2 \leq \delta \leq 1 - p_2$ , for all sufficiently small and large values of the baseline responder rate  $p_2$ . This fact suggests that the choice  $\theta = \delta$ , notwithstanding its popularity in the existing literature on equivalence testing with binomially distributed data [for a selection of pertinent references see § 6.6.3] leads to an ill-considered testing problem.

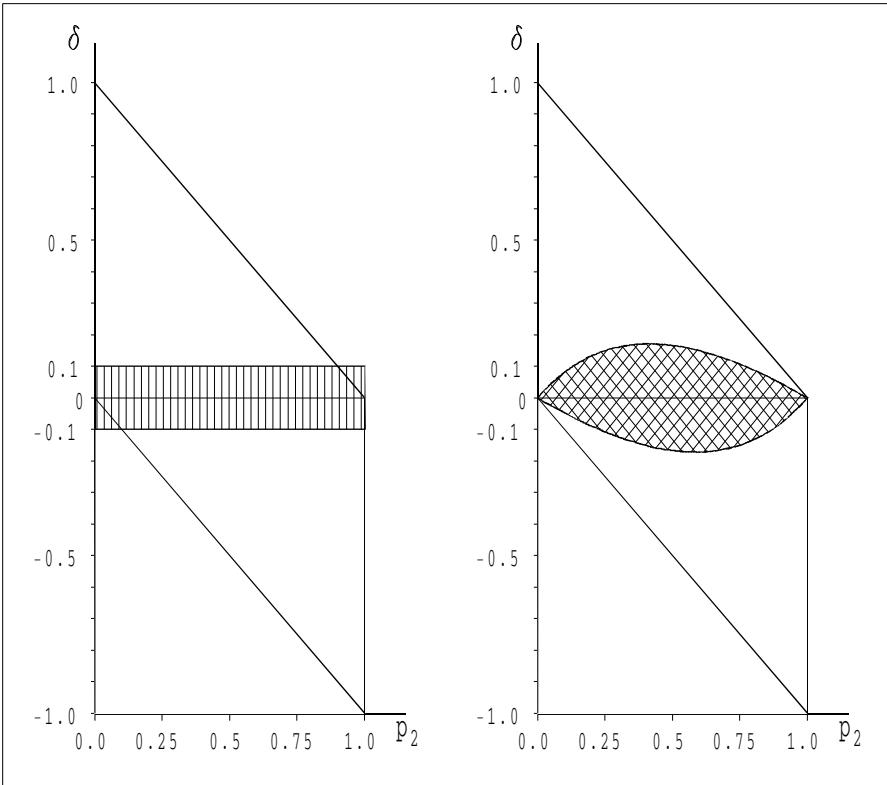


Figure 1.2 *Equivalence hypotheses in terms of the difference between both responder rates [left] and the odds ratio [right] as regions in the  $p_2 \times \delta$ -plane with  $\delta = p_1 - p_2$ . [Rhomboide  $\hat{=}$  set of possible values of  $(p_2, \delta)$ .]*

Another special case whose importance for practical work can hardly be overestimated since, under standard parametric modeling, bioequivalence assessment with data from two-period crossover studies reduces to it, concerns the comparison of two Gaussian distributions with common but unknown variance on the basis of independent samples. Denoting, as usual, the two expected values and the common standard deviation by  $\mu_1$ ,  $\mu_2$  and  $\sigma$ , respectively, the predominant approach starts from the choice  $\theta = \mu_1 - \mu_2$  although there are clear reasons why setting  $\theta = (\mu_1 - \mu_2)/\sigma$  yields a much more sensible measure of distance between the distributions to compare in the setting of the two-sample  $t$ -test: It is an elementary fact (whose implications are given due attention in Lehmann and Romano, 2005, §5.3) that, given any whatever large value of  $|\mu_1 - \mu_2|$ , both distributions become practically indistinguishable if  $\sigma$  is large enough, whereas the areas under the corresponding densities are next to disjoint if  $\sigma$  approaches zero. At the same time, focusing on the standardized rather than the raw difference between the means, facilitates

the step to be discussed in some more detail in the subsequent section: Even in discussions with clinical researchers caring little for statistical subtleties, presentation of a handful of graphs usually suffices to reach a consensus that  $|\mu_1 - \mu_2|/\sigma \geq 1$  is incompatible with the notion of equivalence of two Gaussian distributions, and so on.

Interestingly enough, under some circumstances, a thoughtful discussion of the question how the target parameter for defining the equivalence region should most appropriately be chosen, will even lead to the conclusion that the equivalence testing problem originally in mind should better be replaced by an ordinary one-sided testing problem. An example of this kind arises in bioequivalence studies of which one knows (or feels justified to assume) that no period effects have to be taken into account (Anderson and Hauck, 1990; Wellek, 1990, 1993a) [see also Ch. 10.3 of the present book].

---

## 1.7 Numerical specification of the limits of equivalence

The first question which arises when we want to reach a decision on what numerical values shall be assigned to the equivalence limits  $\theta_o - \varepsilon_1$ ,  $\theta_o + \varepsilon_2$  defining the hypotheses in a testing problem of the form (1.1), is whether or not the equivalence interval has to be symmetric about the reference value  $\theta_o$ . More often than not it seems reasonable to answer this in the affirmative, although virtually all procedures presented in the chapters following the next allow full flexibility in that respect. Perhaps the still best known example of a whole area of application for methods of establishing equivalence in a nonsymmetric sense is bioequivalence assessment along the former FDA guidelines. Before the 1992 revision of its guidance for bioequivalence studies, the FDA strongly recommended to use the specifications  $\theta_o - \varepsilon_1 = 2 \log(.80) \approx -.446$ ,  $\theta_o + \varepsilon_2 = 2 \log(1.20) \approx .365$  for the maximally tolerable shift between the Gaussian distributions eventually to compare. Essentially, the corresponding interval is the log-transform of what is customarily called the 80 to 120% range for the ratio of the true drug formulation effects. In the revised version of the guidelines, the latter has been replaced with the range 80–125%.

With regard to the question whether it is advisable for the statistician to give general recommendations concerning the form of the equivalence interval, we take the same position as on the one- versus two-sidedness controversy in the context of active-control trials [recall § 1.3.2]: This is a point for careful discussion with the researcher planning an individual study and should not made subject to fixed general rules. Instead, full generality should be aimed at in developing the pertinent statistical methods so that we can provide the experimental or clinical researcher with a range of options sufficiently large for allowing him to cover the question he really wants to answer by means of

his data.

Even if the problem is one of testing for noninferiority or has been symmetrized by introducing the restriction  $\varepsilon_1 = \varepsilon_2 = \varepsilon$  in (1.1a) and (1.1b), coming to an agreement with the experimenter about a specific numerical value to be assigned to the only remaining constant determining the equivalence interval, is not always easy. The following table is intended to give some guidance for some of the most frequently encountered settings:

Table 1.1 *Proposals for choosing the limits of a symmetrical equivalence interval or the noninferiority margin in some standard settings.*

(Serial No.)	Setting	Target Parameter or Functional	Reference Value	Tolerance $\varepsilon$ : Strict Choice	Liberl Choice
(i)	Sign test	$p_+ = P[D > 0]^{\dagger)}$	1/2	.10	.20
(ii)	Mann-Whitney	$\pi_+ = P[X > Y]^{\ddagger})$	1/2	.10	.20
(iii)	Two binomial samples	$\log \rho = \log \left[ \frac{p_1(1-p_2)}{(1-p_1)p_2} \right]$	0	.41	.85
(iv)	Paired <i>t</i> -Test	$\delta/\sigma$	0	.25	.50
(v)	Two-Sample <i>t</i> -Test	$(\mu_1 - \mu_2)/\sigma$	0	.36	.74
(vi)	Two Gaussian distr., comparison of var.	$\log(\sigma_1/\sigma_2)$	0	.41	.69
(vii)	Two exponential distr.	$\log(\sigma_1/\sigma_2)$	0	.405	.847

<sup>†)</sup>  $D \equiv$  intraindividual difference for a randomly chosen observational unit

<sup>‡)</sup>  $X, Y \equiv$  independent observations from different distributions

Here are some reasons motivating the above suggestions:

- (i),(ii): Everyday experience shows that most people will rate probabilities of medium size differing by no more than 10%, as rather similar; 20% or more is usually considered indicating a different order of magnitude in the same context.
- (iii): Assuming that the reference responder rate is given by  $p_2 = 1/2$ , straightforward steps of converting inequalities show the condition  $-\varepsilon < p_1 - p_2 < \varepsilon$  to be equivalent to  $|\log \rho| < \log(\frac{1+2\varepsilon}{1-2\varepsilon}) \equiv \varepsilon_{\tilde{\rho}}$ . According to this relationship, the choices  $\varepsilon = .10$  and  $\varepsilon = .20$  [recall (i)] correspond to  $\varepsilon_{\tilde{\rho}} = \log(12/8) = .4055 \approx .41$  and  $\varepsilon_{\tilde{\rho}} = \log(14/6) = .8473 \approx .85$ , respectively.

→ (iv): Under the Gaussian model  $D \sim \mathcal{N}(\delta, \sigma^2)$ , we can write:

$$\begin{aligned} & 1/2 - \varepsilon < p_+ \equiv P[D > 0] < 1/2 + \varepsilon \\ \Leftrightarrow & \Phi^{-1}(1/2 - \varepsilon) < \delta/\sigma < \Phi^{-1}(1/2 + \varepsilon) \\ \Leftrightarrow & -\Phi^{-1}(1 - (1/2 - \varepsilon)) < \delta/\sigma < \Phi^{-1}(1/2 + \varepsilon) \\ \Leftrightarrow & -\Phi^{-1}(1/2 + \varepsilon) < \delta/\sigma < \Phi^{-1}(1/2 + \varepsilon) \end{aligned}$$

where  $\Phi^{-1}$  denotes the quantile function for the standard normal distribution. Hence, the choice  $\varepsilon = .10$  and  $\varepsilon = .20$  in case (i) corresponds here to  $\varepsilon = \Phi^{-1}(.60) = .2529$  and  $\varepsilon = \Phi^{-1}(.70) = .5240$ , respectively.

→ (v): Analogously, relating (ii) to the special case  $X \sim \mathcal{N}(\mu_1, \sigma^2)$ ,  $Y \sim \mathcal{N}(\mu_2, \sigma^2)$ , yields the first in the following chain of inequalities:

$$\begin{aligned} & 1/2 - \varepsilon < \Phi((\mu_1 - \mu_2)/\sqrt{2}\sigma) < 1/2 + \varepsilon \\ \Leftrightarrow & \Phi^{-1}(1/2 - \varepsilon) < (\mu_1 - \mu_2)/\sqrt{2}\sigma < \Phi^{-1}(1/2 + \varepsilon) \\ \Leftrightarrow & -\sqrt{2}\Phi^{-1}(1/2 + \varepsilon) < (\mu_1 - \mu_2)/\sigma < \sqrt{2}\Phi^{-1}(1/2 + \varepsilon). \end{aligned}$$

Thus, the choices suggested for (ii) are this time equivalent to setting  $\varepsilon = \sqrt{2}\Phi^{-1}(.60) = .3577$  and  $\varepsilon = \sqrt{2}\Phi^{-1}(.70) = .7411$ , respectively.

→ (vi): Unlike (ii)–(v), this setting cannot be related in a natural way to case (i). The suggested values of  $\varepsilon$ , except for rounding to two significant decimals, are equivalent to the requirements  $2/3 < \sigma_1/\sigma_2 < 3/2$  and  $1/2 < \sigma_1/\sigma_2 < 2$ , respectively. The latter seem again plausible for common statistical sense.

→ (vii): The exponential scale model is a particularly important special case of a proportional hazards model so that the general considerations of § 6.3 about the latter apply.

---

*General techniques for dealing with  
noninferiority problems*

---

## 2.1 Standard solution in the case of location parameter families

Whenever the target parameter  $\theta$  is a measure of the shift in location of the distributions of interest, shifting the common boundary of a pair of one-sided hypotheses produces a testing problem which is new only when we look at the concrete meaning of a positive decision in favor of the alternative. From a purely statistical point of view, no more than a trivial modification of the usual test for the corresponding conventional problem  $\theta \leq \theta_0$  versus  $\theta > \theta_0$  is required. Subsequently we describe the rationale behind this modification in some detail for the case of comparing two treatments  $A$  and  $B$  on the basis of paired and of independent samples of univariate observations, respectively.

### 2.1.1 Paired observations

The data to be analyzed in any trial following the basic scheme of paired comparisons consists of random pairs  $(X, Y)$ , say, such that  $X$  and  $Y$  gives the result of applying treatment  $A$  and  $B$  to the same arbitrarily selected observational unit. Except for the treatment, the conditions under which  $X$  and  $Y$  are taken are supposed to be strictly balanced allowing the experimenter to interpret the intra-subject difference  $D = X - Y$  as quantifying in that individual case the superiority in effectiveness of treatment  $A$  as compared to  $B$ . In this setting, speaking of a location problem means to make the additional assumption that in the underlying population of subjects, any potential difference between both treatments is reflected by a shift  $\theta$  in the location of the distribution of  $D$  away from  $\theta_0 = 0$  leaving the distributional shape per se totally unchanged. In absence of any treatment difference at all, let this distribution be given by some continuous cumulative distribution function (cdf)  $F_0 : \mathbb{R} \rightarrow [0, 1]$  symmetric about zero. For the time being, we do not specify whether the baseline cdf has some known form (e.g.,  $F_0 = \Phi(\cdot/\sigma)$  with  $\Phi$  denoting the standard normal cdf), or is allowed to vary over the whole class of all continuous cdf's on the real line being symmetric about zero [→

nonparametric one-sample location problem, cf. Randles and Wolfe (1979), Ch. 10)].

Regarding the full data set, i.e., the sample  $(X_1, Y_1), \dots, (X_n, Y_n)$  of all  $n$  pairs of measurements obtained under both treatments, statistical inference is based on the following assumptions about the corresponding intraindividual differences  $D_i = X_i - Y_i$ :

- (a) The vector  $(D_1, \dots, D_n)$  is independent and identically distributed (iid);
- (b)  $P[D_i \leq d] = F_\circ(d - \theta)$  for arbitrary  $d \in \mathbb{R}$  and all  $i = 1, \dots, n$ , with  $F_\circ(-d) = 1 - F_\circ(d)$   $\forall d$  and  $\theta$  denoting the parameter of interest. (By convention, positive values of  $\theta$  are assumed to indicate a tendency towards “better” results under treatment  $A$  as compared to  $B$ ).

Now, the general form of the rejection region of a test at level  $\alpha$  for the traditional one-sided testing problem  $H_1^\circ : \theta \leq 0$  versus  $K_1^\circ : \theta > 0$  [→ (1.2), specialized to the case that  $\theta_\circ = 0, \varepsilon = 0$ ] is well known to be given by

$$\{T(D_1, \dots, D_n) > c_\alpha\}$$

where  $T(\cdot)$  denotes a suitable real-valued function of  $n$  arguments (usually called the test statistic), and  $c_\alpha$  the upper  $100\alpha$  percentage point of the distribution of the random variable  $(D_1, \dots, D_n)$  under  $\theta = 0$ . If we change the directly observed intra-subject differences  $D_i$  to  $\tilde{D}_i = D_i + \varepsilon$ , then the modified sample  $(\tilde{D}_1, \dots, \tilde{D}_n)$  obviously satisfies again (a) and (b), provided the parameter  $\theta$  is likewise shifted introducing the transform  $\tilde{\theta} = \theta + \varepsilon$ . Furthermore, the testing problem  $H_1 : \theta \leq -\varepsilon$  versus  $K_1 : \theta > -\varepsilon$  we are primarily interested in, is clearly the same as the ordinary one-sided problem  $\tilde{\theta} \leq 0$  versus  $\tilde{\theta} > 0$  relating to the transformed intra-subject differences  $\tilde{D}_i$ . Hence, in the present setting we obtain the desired test for the one-sided equivalence problem  $H_1$  vs.  $K_1$  simply by using the rejection region of the test for the associated nonshifted null hypothesis in terms of the observations shifted the same distance as the common boundary of the hypotheses but in the opposite direction. The test obtained in this way rejects the null hypothesis  $H_1 : \theta \leq -\varepsilon$  [→ relevant inferiority] if and only if we find that  $T(D_1 + \varepsilon, \dots, D_n + \varepsilon) > c_\alpha$  where  $T(\cdot)$  and  $c_\alpha$  are computed in exactly the same way as before.

### *Example 2.1*

We illustrate the approach described above by reanalyzing the data from a study (Miller et al., 1990) of possible effects of the fat substitute olestra on the absorption of highly lipophilic oral contraceptives. The sample recruited for this trial consisted of 28 healthy premenopausal women. During the verum phase, each subject consumed 18 gm/day olestra for 28 days while taking a combination of norgestrel (300 $\mu$ g) and ethinyl estradiol (30 $\mu$ g) as an oral con-

Table 2.1 *Maximal concentrations of norgestrel (ng/ml) in the sera of 28 women while consuming olestra ( $X_i$ ) and meals containing ordinary triglycerides ( $Y_i$ ), respectively [ $\tilde{D}_i = D_i + \varepsilon$ ;  $\tilde{R}_i^+ = \text{rank of the } i\text{th subject with respect to } |\tilde{D}_i|$ ;  $\varepsilon = 1.5$ ].*

$i$	$X_i$	$Y_i$	$\tilde{D}_i$	$\tilde{R}_i^+$	$i$	$X_i$	$Y_i$	$\tilde{D}_i$	$\tilde{R}_i^+$
1	6.03	6.62	0.91	12	15	11.81	11.19	2.12	21
2	5.62	6.78	0.34	5	16	8.72	9.55	0.67	9
3	6.93	6.85	1.58	18	17	7.01	5.53	2.98	26
4	5.86	8.09	-0.73	11	18	7.13	6.71	1.92	20
5	8.91	9.18	1.23	15	19	6.56	6.53	1.53	17
6	5.86	7.47	-0.11	1	20	4.22	5.39	0.33	4
7	9.43	9.90	1.03	13	21	4.13	4.92	0.71	10
8	5.30	4.29	2.51	22	22	6.57	9.92	-1.85	19
9	4.99	3.80	2.69	24	23	8.83	10.51	-0.18	2
10	6.12	7.01	0.61	8	24	9.05	10.15	0.40	6
11	12.45	9.53	4.42	28	25	9.31	9.55	1.26	16
12	5.48	6.39	0.59	7	26	7.67	8.95	0.22	3
13	6.04	4.63	2.91	25	27	7.66	6.63	2.53	23
14	8.32	5.54	4.28	27	28	5.45	8.01	-1.06	14

traceptive. Blood samples were taken on days 12 to 14 of the cycle and analyzed for ethinyl and estradiol concentrations. For the placebo phase, the experimental and measurement procedure was exactly the same as for verum except for replacing olestra with conventional triglycerides at each meal. Table 2.1 gives the individual results for norgestrel and the maximum concentration  $C_{max}$  as the pharmacokinetic parameter of interest.

According to the general objective of the trial, let us aim at establishing that the consumption of olestra does not reduce the bioavailability of norgestrel (as measured by  $C_{max}$ ) to a relevant extent. Further, let us define  $\theta$  as denoting the population median of the distribution of the intra-subject differences  $D_i = X_i - Y_i$  with  $\varepsilon = 1.5$  as the limit of relevance, and base the confirmatory analysis of the data on the Wilcoxon signed rank statistic. Then, the computational steps which have to be carried out in order to test for one-sided equivalence are as follows.

- (i) For each  $i = 1, \dots, 28$ , the shifted intra-subject difference  $\tilde{D}_i$  [→ Table 2.1, 4th column] and the rank  $\tilde{R}_i^+$  with respect to  $|\tilde{D}_i|$  [→ Table 2.1, 5th column] have to be determined.
- (ii) Denoting the sum of ranks of subjects with a positive sign of  $\tilde{D}_i$  by  $\tilde{V}_s^+$ , the value of this modified signed rank statistic is computed to be  $\tilde{V}_s^+ = 359$ .

- (iii) Under  $\theta = -\varepsilon = -1.5$ ,  $\tilde{V}_s^+$  has an asymptotic normal distribution with expected value  $E_0(\tilde{V}_s^+) = n(n+1)/4 = 28 \cdot 29/4 = 203$  and variance  $Var_0(\tilde{V}_s^+) = n(n+1)(2n+1)/24 = 1928.5$ . Hence, the usual approximation with continuity correction gives the p-value (observed significance probability)  $p_{obs} = \Phi[(203 - 359 + .5)/\sqrt{1928.5}] = \Phi[-3.5410] = .0002$ .

In view of the order of magnitude of the significance probability obtained in this way, the decision of the modified signed rank test for one-sided equivalence is positive even at the 1% level in the present case. In other words, the results of the study performed by Miller et al. (1990) contain sufficient evidence in favor of the hypothesis that the consumption of olestra does not lead to a relevant decrease of the bioavailability of norgestrel.

### 2.1.2 Two independent samples

If a comparative study of two treatments  $A$  and  $B$  follows the parallel group design, the data set to be analyzed consists of values of  $m + n$  mutually independent random variables  $X_1, \dots, X_m, Y_1, \dots, Y_n$ . By convention, it is assumed that the  $X_i$  are observed in subjects who are given treatment  $A$  whereas the  $Y_j$  relate to the other treatment, i.e., to  $B$ . In this setting, the shift model implies that any possible treatment difference can be represented by means of the relationship  $X_i \stackrel{d}{=} Y_j + \theta$  where, as usual (cf. Randles and Wolfe, 1979, §1.3), the symbol “ $\stackrel{d}{=}$ ” indicates identity of the distributions of the two random variables appearing on its left- and right-hand side, respectively. In other words, in the case of the parallel group design the shift model assumes that the distributions associated with the two treatments have exactly the same shape, implying that distribution  $A$  can be generated by shifting all individual values making up distribution  $B$  the same distance  $|\theta|$  to the right (for  $\theta > 0$ ) or left (for  $\theta < 0$ ). Making this idea mathematically precise leads to specifying the following assumptions about the two distributions under comparison:

- (a\*) The complete data vector  $X_1, \dots, X_m, Y_1, \dots, Y_n$  is independent, and all  $X_i$  and  $Y_j$  have the same continuous distribution function  $F$  and  $G$ , respectively.
- (b\*) There exists a real constant  $\theta$  such that  $F(x) = G(x - \theta)$  for all  $x \in \mathbb{R}$ .

Reduction of the one-sided equivalence problem  $H_1 : \theta \leq -\varepsilon$  vs.  $K_1 : \theta > -\varepsilon$  to the corresponding ordinary one-sided testing problem  $\tilde{\theta} \leq 0$  vs.  $\tilde{\theta} > 0$  proceeds here along analogous lines as in the case of paired observations discussed in the previous subsection. To start with, one has to select a suitable test for the nonshifted hypothesis which rejects if and only if one has  $T(X_1, \dots, X_m, Y_1, \dots, Y_n) > c_\alpha$ , where the test statistic  $T(\cdot)$  is a real-valued function of  $m + n$  arguments and  $c_\alpha$  stands for the upper  $100\alpha$  percentage point of the distribution of  $T(X_1, \dots, X_m, Y_1, \dots, Y_n)$  under  $\theta = 0$ . As before,

this test has to be carried out with suitable transforms  $\tilde{X}_i$  and  $\tilde{Y}_j$ , say, of the primary observations  $X_i$  and  $Y_j$ , given by  $\tilde{X}_i = X_i + \varepsilon$  (for  $i = 1, \dots, m$ ) and  $\tilde{Y}_j = Y_j$  (for  $j = 1, \dots, n$ ), respectively. In view of the close analogy to the paired observations case, illustration of the approach by another numerical example is dispensable. Its implementation gets particularly simple in the parametric case with  $F(x) = \Phi((x - \theta)/\sigma)$ ,  $G(y) = \Phi(y/\sigma)$  and  $T$  chosen to be the ordinary two-sample  $t$ -statistic: One has just to replace  $\bar{x} - \bar{y}$  by  $\bar{x} - \bar{y} + \varepsilon$  in the numerator of  $T$  then and proceed exactly as usual in all remaining steps.

### 2.1.3 Power and sample size calculation based on tests for noninferiority under location-shift models

The changes required for adapting the usual formula for power and sample size calculation for tests for one-sided location-shift hypotheses to the noninferiority case are likewise straightforward. Suppose the specific alternative to be detected in a test for noninferiority of the form discussed in this section is given by some fixed value  $\theta_a$  of the shift parameter to which the proposed hypothesis is referring. Then, the power of the test for noninferiority with margin  $\varepsilon$  is the same as that of the ordinary one-sided test for the respective setting against the alternative  $\tilde{\theta}_a = \varepsilon + \theta_a$ , and the sample size required for ensuring that the test for noninferiority rejects with given probability  $\beta$  under this alternative, is likewise obtained by replacing  $\theta_a$  with  $\tilde{\theta}_a$  in the formula or algorithm for the one-sided case.

In the majority of practical applications, the alternative of primary interest is given by  $\theta_a = 0$  specifying that the effects of both treatments are identical. The power of the noninferiority test against this “null alternative” is obtained by calculating the rejection probability of the corresponding test for  $H_1^\circ : \theta \leq 0$  versus  $K_1^\circ : \theta > 0$  under  $\theta = \varepsilon$ . Specifically, *for the shifted t-tests for noninferiority*, the power against  $\theta = 0$  is given by

$$\text{POW}_0 = 1 - G_{\sqrt{n\varepsilon}/\sigma_D}(t_{n-1;1-\alpha}) \quad (2.1a)$$

and

$$\text{POW}_0 = 1 - G_{\sqrt{mn/N\varepsilon}/\sigma}^*(t_{N-2;1-\alpha}) \quad (2.1b)$$

in the paired-sample and independent-sample case, respectively. In the first of these formula,  $G_{\lambda_{nc}}(\cdot)$  stands for the cdf of the noncentral  $t$ -distribution with noncentrality parameter  $\lambda_{nc} \in \mathbb{R}$  and  $n - 1$  degrees of freedom. Furthermore,  $\sigma_D$  denotes the population standard deviation of the intraindividual differences,  $t_{n-1;1-\alpha}$  the  $(1 - \alpha)$ -quantile of the central  $t$ -distribution with  $df = n - 1$ .  $G_{\lambda_{nc}}^*(\cdot)$  differs from  $G_{\lambda_{nc}}(\cdot)$  by changing the number of degrees of freedom from  $n - 1$  to  $N - 2 \equiv m + n - 2$ , and the analogous change has to be made concerning the central  $t$ -quantile when proceeding from the one- to the two-sample case. Evaluation of the above expression for the power is

particularly easy in a programming environment like R and SAS providing a predefined function for computing the noncentral  $t$ -distribution function.

---

## 2.2 Methods of constructing exact optimal tests for settings beyond the location-shift model

Clearly, the simple trick behind the approach of § 2.1 works only with problems relating to location-shift models, and the repertoire of noninferiority tests which can be constructed in that way is much too narrow for covering even the standard settings occurring in real applications. However, from a theoretical point of view, noninferiority problems are nothing but one-sided testing problems with a nonzero cutoff specified as the boundary point of the hypotheses between which one wants to decide. Fortunately, the mathematical principles leading to exact optimum solutions of one-sided hypothesis testing problems apply for arbitrary specifications of that boundary point, and the only modification required when proceeding from the classical to the noninferiority formulation concerns the way of determining the critical bounds and constants. In the noninferiority case, a suitable *noncentral* version of the sampling distribution of the test statistic involved has to be used.

In order to make these general statements more concrete, let us denote by  $\mathbf{X}$  the collection of all observations obtained in terms of the experiment or study under current analysis, i.e., a random vector of dimension at least as large as the sum of all sample sizes involved [e.g., in an ordinary parallel group design for a trial of two treatments one has  $\mathbf{X} = (X_1, \dots, X_m, Y_1, \dots, Y_n)$ ]. Let us further assume that this primary data vector can be reduced to some real valued statistic  $T(\mathbf{X})$  such that the possible distributions of  $T(\mathbf{X})$  constitute a family with monotone likelihood ratios in the parameter  $\theta$  of interest (also called a strictly totally positive family of order 2 or STP<sub>2</sub> family — see Definition A.1.1 in the Appendix). Then, it is a well-known fact of the mathematical theory of hypotheses testing (see, e.g., Lehmann and Romano, 2005, § 3.4) that for any choice of the noninferiority margin  $\varepsilon$ , there is a test for  $H_1 : \theta \leq \theta_o - \varepsilon$  versus  $K_1 : \theta > \theta_o - \varepsilon$  which is uniformly most powerful among all tests at the same significance level  $\alpha$  depending on the data only through  $T$ . The rejection region of this test is given by

$$\{ \mathbf{x} \mid T(\mathbf{x}) > k \} . \quad (2.2)$$

The way in which the critical constant  $k$  has to be determined depends on the type of the distribution which the test statistic follows under  $\theta = \theta_o - \varepsilon$ . In the continuous case, the optimal  $k$  is obtained by solving the equation

$$P_{\theta_o - \varepsilon}[T(\mathbf{X}) > k] = \alpha, \quad \infty < k < \infty . \quad (2.3)$$

For noncontinuous  $T(\mathbf{X})$ , a solution to (2.3) will typically not exist. However, it will always be possible to uniquely determine a pair  $(k, \gamma)$  of real numbers the second of which belongs to the half-open unit interval  $[0, 1)$  such that there holds

$$P_{\theta_0 - \varepsilon}[T(\mathbf{X}) > k] + \gamma P_{\theta_0 - \varepsilon}[T(\mathbf{X}) = k] = \alpha, \quad \infty < k < \infty. \quad (2.4)$$

For  $\gamma > 0$ , the exact level- $\alpha$  test has to be carried out entailing a randomized decision between  $H_1$  and  $K_1$  when  $T(\mathbf{X})$  falls on the critical point  $k$ . In this case which can almost surely be ruled out for continuously distributed  $T(\mathbf{X})$ , the null hypothesis  $H_1$  of (relevant) inferiority has to be rejected [accepted] with probability  $\gamma$  [ $1 - \gamma$ ]. Since randomized decision rules are rarely applicable in the confirmatory statistical analysis of real research data, the point  $k$  is usually incorporated in the acceptance region even if the event  $\{T(\mathbf{X}) = k\}$  has positive probability, giving a test which is more or less conservative. Promising techniques for reducing this conservatism will be discussed in § 2.5.

Even if the STP<sub>2</sub>-property of the family of distributions of  $T(\mathbf{X})$  can be taken for granted, the precise meaning of the adjective “optimal” we used above with regard of a test of the form (2.2) (or, in the noncontinuous case, its randomized counterpart) depends on the relationship between  $T(\mathbf{X})$  and the distributions from which the primary data  $\mathbf{X}$  have been taken. The most important cases to be distinguished from this point of view are the following:

- (i)  $T(\mathbf{X})$  is sufficient for the family of the possible distributions of  $\mathbf{X}$ ; then, optimal means uniformly most powerful (UMP).
- (ii) The model underlying the data corresponds to a multiparameter exponential family of distributions (see Definition A.2.1 in the Appendix); then, in (2.3) and (2.4), the symbol  $P_{\theta_0 - \varepsilon}[\cdot]$  must be interpreted as denoting the conditional distribution of  $T(\mathbf{X})$  given some fixed value of another (maybe multidimensional statistic) being sufficient for the remaining parameters of the model, and the test rejecting for sufficiently large values of  $T(\mathbf{X})$  is uniformly most powerful among all unbiased tests (UMPU).
- (iii) The proposed testing problem remains invariant under some group of transformations, and the statistic  $T(\mathbf{X})$  is maximal invariant with respect to that group; then, the test with rejection region (2.2) is uniformly most powerful among all invariant level- $\alpha$  tests (UMPI) for the same problem.

[For proofs of these statements see Sections 1.9, 4.4 and 6.3 of the book by Lehmann and Romano (2005).]

Settings admitting the construction of UMP tests for (one-sided) equivalence are dealt with in Chapter 4. Important special cases of equivalence testing problems which can be solved through reduction by sufficiency are the comparison of two binomial distributions from which paired [→ § 5.2] or

independent samples [ $\rightarrow \S\ 6.6$ ] are available. Somewhat surprisingly, the same approach works in constructing a UMPU test for goodness of fit (rather than lack of fit) of the distribution of a biallelic genetic marker (SNP) with the Hardy-Weinberg model, as will be shown in  $\S\ 9.4$ . Among the practically most useful specific tests of category (iii) are the equivalence versions of the paired [ $\rightarrow \S\ 5.3$ ] and the two-sample  $t$ -tests [ $\rightarrow \S\ 6.1$ ].

---

## 2.3 Large-sample solutions for problems inaccessible for exact constructions

As is true for virtually all areas of statistical inference, the range of noninferiority testing problems accessible to exact methods is too limited for covering the needs arising in practical data analysis. Problems for which no satisfactory exact solution is available, are by no means predominantly of a kind involving nonparametric or other rather complex models like the proportional hazards model for possibly censored survival times. Even such comparatively simple problems as those of testing for one-sided equivalence of binomial proportions with respect to the difference  $\delta = p_1 - p_2$  in the analysis of two independent or paired samples of binary data are cases in point. Fortunately, there is an asymptotic approach to the construction of tests for noninferiority which will enable us to fill a considerable part of the gaps left by exact constructional approaches. It applies to virtually any situation where the hypothesis to be tested refers to a parameter or (in non- or semiparametric models) functional for which an asymptotically normal estimator can be found.

The notation which was introduced in  $\S\ 1.5$  is interpreted here in its broadest sense covering in particular cases where the target parameter  $\theta$  depends on more complex characteristics of the distribution functions under comparison than merely their first two moments or selected quantiles. In other words,  $\theta$  is allowed to have the meaning of an arbitrary functional of the vector of all underlying distribution functions. The dimension of this vector can be any positive integer  $k$  which must not vary with the total sample size  $N$ . For illustration, let us anticipate a small bit of the material presented in  $\S\ 6.2$  referring to the nonparametric two-sample problem with continuous data  $X_1, \dots, X_m$  and  $Y_1, \dots, Y_n$ . In this setting, one wishes to compare  $k = 2$  distribution functions usually denoted  $F$  and  $G$ , and an intuitively appealing choice of a functional which a reasonable measure of distance between them can be based upon, is  $\theta = \int GdF = P[X_i > Y_j]$ . Returning to the problem of testing for equivalence in its most general formulation, suppose further that for each given value of the total sample size  $N$ , there is some real-valued statistic  $T_N$ , say, such that the associated sequence  $(T_N)_{N \in \mathbb{R}}$  is an asymptotically

normal estimator for  $\theta$  in the sense that we have

$$\sqrt{N}(T_N - \theta)/\sigma \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) \quad \text{as } N \rightarrow \infty \quad (2.5)$$

for a suitable positive functional  $\sigma$  of the vector  $(F_1, \dots, F_k)$  of distribution functions under consideration.

Now, the assumed weak-convergence property of  $T_N$  clearly implies that an asymptotically valid test for the noninferiority problem  $H_1 : \theta \leq \theta_o - \varepsilon$  versus  $K_1 : \theta > \theta_o - \varepsilon$  can be based on the rejection region

$$\left\{ (T_N - (\theta_o - \varepsilon))/\tau_N > u_{1-\alpha} \right\} \quad (2.6)$$

with  $u_{1-\alpha}$  denoting the  $(1 - \alpha)$ -quantile of the standard normal distribution. Of course, the asymptotic standard error  $\tau_N$  will not be known in practice. However, it can be replaced with any estimator  $\hat{\tau}_N = \hat{\sigma}_N/\sqrt{N}$  based on a weakly consistent estimator of  $\sigma$ .

---

## 2.4 Objective Bayesian methods

From the Bayesian viewpoint, problems of testing for noninferiority exhibit no peculiarities at all, neither conceptually nor technically. All that is needed as soon as the target parameter  $\theta$  and the common boundary  $\theta_o - \varepsilon$  of the hypotheses has been fixed, is a joint prior distribution  $\pi(\cdot)$ , say, of all unknown parameters contained in the underlying statistical model. The Bayesian test for one-sided equivalence of  $\theta$  with  $\theta_o$  rejects if the posterior probability of the interval  $(\theta_o - \varepsilon, \infty)$  specified by the alternative hypothesis with respect to this prior is computed to be larger than a suitably lower bound which is customarily set equal to the complement of the nominal significance level. In other words, given the prior distribution  $\pi(\cdot)$ , the Bayesian test for noninferiority uses the decision rule\*:

Reject (relevant) inferiority if and only if

$$\pi^{\theta|\mathbf{X}=\mathbf{x}}(\theta_o - \varepsilon, \infty) \geq 1 - \alpha. \quad (2.7)$$

For the sake of eliminating subjectivism entailed by selecting the prior in an arbitrary way, relying on so-called noninformative priors has much to recommend it. By definition, a noninformative or “objective” prior is a (maybe improper) uniform distribution of the model parameters involved or of suitable parametric functions. Although there is still considerable controversy

---

\*Introducing extra symbols for parameters treated as random variables rather than constants seems dispensable for the moment.

about the theoretical foundation of the concept (cf. Berger, 1985, § 3.3; Cox and Hinkley, 1974, § 10.4), there is a useful pragmatic rule for determining a noninformative prior which has been extensively exploited in many applications of Bayesian inference (see in particular the landmark monograph of Box and Tiao, 1973). The rule was first proposed by Jeffreys (1961) and states that  $\pi(\cdot)$  is (approximately) noninformative if it is defined by a density (with respect to Lebesgue measure) taken proportional to the square root of the determinant of the information matrix.

Experience shows that for many problems of hypotheses testing, the Bayesian approach yields procedures with fairly satisfactory frequentist properties as long as a neutral prior is used. However, even then it can and will happen [see, e.g., § 6.6.2] that the resulting test requires adjustment for level or size through decreasing the nominal  $\alpha$  used in the basic decision rule (2.7). One of the major advantages of this approach is its flexibility with respect to the choice of the target parameter in terms of which the equivalence region shall be defined.

---

## 2.5 Improved nonrandomized tests for discrete distributions

The loss of power entailed by dispensing with randomized decisions in performing tests obtained by applying the principles introduced in § 2.2 to discrete distributions can be dismissed as irrelevant for practical applications only as long as the sample sizes are fairly large. In small samples this conservatism can take on extreme proportions, and that is the reason why even in the case of a traditional one-sided null hypothesis, the exact Fisher type test for the two-sample problem with binomial data which is probably still the most frequently used non-asymptotic test for discrete data, has been criticized in the literature almost from the beginning (cf. Pearson, 1947).

In settings relating to single-parameter problems, the only recommendable approach to reducing the conservatism of a nonrandomized test based on a statistic  $T(\mathbf{X})$  following a distribution of the discrete type, is the so-called mid-p-value method. In order to understand how this does work, it is necessary to realize that, in a noninferiority setting, the nonrandomized version of the test based on (2.2) can equivalently be defined by the p-value-based decision rule

Reject  $H_1 : \theta \leq \theta_o - \varepsilon$  if and only if

$$p(\mathbf{x}) \equiv P_{\theta_o - \varepsilon}[T(\mathbf{X}) \geq T(\mathbf{x})] \leq \alpha. \quad (2.8)$$

Replacing  $p(\mathbf{x})$  with  $p_+(\mathbf{x}) \equiv P_{\theta_o - \varepsilon}[T(\mathbf{X}) > T(\mathbf{x})]$  in (2.8) would clearly lead to an anti-conservative testing procedure, and the simple idea behind the

mid-p-value technique (cf. Agresti, 2002, §1.4.5) is to compromise between over- and anti-conservatism by using the average of  $p(\mathbf{x})$  and  $p_+(\mathbf{x})$ . The resulting modification to (2.8) can be written

Reject  $H_1 : \theta \leq \theta_0 - \varepsilon$  if and only if

$$p_{mid}(\mathbf{x}) \equiv \left( P_{\theta_0 - \varepsilon}[T(\mathbf{X}) > T(\mathbf{x})] + \frac{1}{2} P_{\theta_0 - \varepsilon}[T(\mathbf{X}) = T(\mathbf{x})] \right) \leq \alpha. \quad (2.9)$$

It is important to note that there is no mathematical result due to which validity of the test given by (2.9) with respect to the significance level can be generally guaranteed. However, numerical results from simulation studies of a variety of discrete families of distributions show that mid-p-value based tests are typically still conservative but to a markedly lesser extent than their exact counterparts.

Of course, subject to the reservation made above, the mid-p-value approach works also in situations where the critical lower bound  $k$  [recall (2.2)] has to be determined from the conditional distribution of  $T(\mathbf{X})$  given some other statistics  $S$ , say, as has been outlined in §2.2 under (ii). However, in conditional settings there are more promising ways of avoiding unnecessarily large losses in power through relying on nonrandomized testing procedures. Such methods have been most extensively discussed in the literature for the binomial two-sample setting, mostly in connection with the keyword exact Fisher type test (for a still reasonably up-to-date review see Martín, 1998). For all problems which, like the exact Fisher test, relate to multiparameter exponential families, it seems desirable to preserve the basic conditional structure of the test, by reasons which are carefully explained by Little (1989) for that special case. Essentially, two different approaches to constructing improved nonrandomized versions of conditional tests will be used in later chapters of this book:

- (i) In order to describe the first of these approaches let us assume that the statistic  $S$  upon which conditioning has to be performed, has finite sample space  $\{s_1, \dots, s_q\}$ . The construction of an improved nonrandomized conditional test starts with arranging the set  $\{k(s_\nu) | \nu = 1, \dots, q\}$  of all critical bounds to  $T(\mathbf{X})$  in ascending order with respect to their unconditional rejection probabilities maximized over  $H_1$ . Afterwards, the union of all  $q$  conditional rejection regions in the sample space of  $T(\mathbf{X})$  is successively enlarged by adding in that order as many boundary points  $k(s_\nu)$  as possible without raising the size of the test (i.e., the maximum of the rejection probability taken over the null hypothesis) over the significance level  $\alpha$ .
- (ii) The other approach follows a strikingly simple idea: The usual nonrandomized version of the conditional test is carried out at an increased nominal significance level  $\alpha^* > \alpha$  determined iteratively by maximization over the set of all nominal levels which are admissible in the sense of keeping the size of the resulting test below  $\alpha$ .

Sort of a hybrid between both approaches was used by Starks (1979) in order to construct critical regions of an improved nonrandomized sign test for non-continuous distributions. The second approach goes back to Boschloo (1970) who exploited it for constructing a more powerful version of the one-sided exact Fisher type test for two independent binomially distributed samples. The algorithm described by McDonald et al. (1977, 1981) leads to exactly the same critical regions as Boschloo's increased nominal levels. The most important difference between (i) and (ii) is that the first approach may produce rejection regions containing points of the sample space whose conditional p-value is larger than that of some or even several points retained in the acceptance region. Our preference for (ii) is mainly motivated by the fact that it lacks this rather counterintuitive property. Moreover, for the practical implementation of Boschloo's technique one can rely on tables of an extremely simple and compact form since, given the sample size(s), the boundary of the hypotheses and the target significance level, the nonrandomized test modified according to (ii) is uniquely determined by just one additional constant, viz.  $\alpha^*$ .

---

## 2.6 Relationship between tests for noninferiority and two-sided equivalence tests

From the mathematical point of view, the construction of procedures for noninferiority problems entails no specific ideas and techniques beyond those which will be presented in the next chapter as basic tools for the derivation of two-sided equivalence tests. On the contrary, solving a noninferiority problem is generally a good bit easier than constructing its two-sided counterpart, due to the fact, that for the latter, the number of critical constants to be determined is double (2 versus 1 in the case of a continuously distributed test statistic, and 4 versus 2 in constructing exact nonrandomized tests for discrete distributions). This difference in technical complexity will be directly reflected in the subsequent chapters where for the majority of specific settings considered, the exposition of the test for noninferiority is kept much shorter than that of the equivalence testing procedure.

In practice, there is even the possibility of converting any given test for equivalence in the strict sense into a test for noninferiority: all that is necessary for exploiting the basic logical relationship between the hypotheses to be established, is simply to specify an extremely large value for the right-hand equivalence margin  $\varepsilon_2$  set in the equivalence case. As an obvious implication of this remark, it can be stated that the repertoire of noninferiority testing procedures ready for application to real problems and data sets is at least as large as that of procedures for the assessment of two-sided equivalence.

## 2.7 Halving alpha?

In the statistical parts of the guidelines issued by the drug regulation authorities, there is an increasing tendency to extend the verdict against the use of one-sided testing procedures in the confirmatory analysis of clinical trials also to noninferiority studies (see, e.g., EMEA, 1998, p. 27; EMEA, 2005, p. 3). In this book, we will not follow the recommendation to carry out by default one-sided equivalence tests at level 2.5% instead of 5%. Actually, the only rational argument in favor of this modified convention concerning the choice of the significance level is that one wants to prevent inflating the effective type-I error risk through reporting significance decisions as belonging to a one-sided hypothesis formulation although, in truth, the direction of the difference to detect was not fixed *a priori*. Even in a conventional placebo-controlled trial, this option seems to exist only in a theoretical sense since it is hard to imagine that a clinical investigator could really try to declare a significant benefit of placebo as compared to the active treatment of interest, as a positive result. In a noninferiority trial involving an active control, the suspicion that such post-hoc switching of the roles of both treatments involved might really be intended, seems plainly absurd: every expert in the clinical discipline to which the trial relates, will be perfectly sure about which of the two treatments was used as a reference so that a “positive” result of a test for the problem arising from an interchange of hypotheses in (1.2) would mean that the experimental treatment was shown to be relevantly inferior to the control.

Of course, unwillingness to adopt the  $\alpha/2$  doctrine by no means implies that the tests for noninferiority discussed in the subsequent chapters are useless for a researcher feeling like performing them at level 2.5% nevertheless. The only concrete consequence is that all examples will be worked out under the specification  $\alpha = .05$ .

# 3

---

## *General approaches to the construction of tests for equivalence in the strict sense*

---

### 3.1 The principle of confidence interval inclusion

The most popular and frequently used approach to problems of equivalence testing in the strict (i.e., two-sided) sense is still that which starts from interval estimation. The principle of confidence interval inclusion was originally introduced in a rather inconspicuous paper (written in the form of a Letter to the Editor of a pharmaceutical journal) by Westlake (1972). Since that time, it has been discussed in at least three different versions by numerous authors (among many others, see Metzler, 1974; Westlake, 1976, 1979, 1981; Steinijans and Diletti, 1983, 1985; Mau, 1988; Hauschke et al., 1990) dealing almost exclusively with *bioequivalence* assessment ( $\rightarrow$  Ch. 10) as the only worthwhile field of applying such methods. Our exposition of the approach concentrates on that version which is most efficient with regard to the power of the corresponding testing procedures.

Let  $\underline{\theta}(\mathbf{X}; \alpha)$  and  $\bar{\theta}(\mathbf{X}; \alpha)$  denote a lower and upper confidence bound for  $\theta$  at the *same one-sided* confidence level  $1 - \alpha$ . In the symbols used for these confidence bounds (and similarly for test statistics),  $\mathbf{X}$  represents the collection of all observations obtained from the study under current analysis. Then, according to the confidence interval inclusion principle, we get a valid test for equivalence in the sense of (1.1b) [ $\rightarrow$  p. 11] by way of rejecting the null hypothesis  $H : \theta \leq \theta_0 - \varepsilon_1$  or  $\theta \geq \theta_0 + \varepsilon_2$  if and only if the random interval  $(\underline{\theta}(\mathbf{X}; \alpha), \bar{\theta}(\mathbf{X}; \alpha))$  is completely covered by the equivalence interval  $(\theta_0 - \varepsilon_1, \theta_0 + \varepsilon_2)$  specified by our alternative hypothesis  $K$ . Of course, this means that the interval inclusion test decides in favor of equivalence provided both inequalities  $\underline{\theta}(\mathbf{X}; \alpha) > \theta_0 - \varepsilon_1$  and  $\bar{\theta}(\mathbf{X}; \alpha) < \theta_0 + \varepsilon_2$  are satisfied simultaneously.

To be sure, under the assumptions made explicit in the previous paragraph, the two-sided confidence level of the interval  $(\underline{\theta}(\mathbf{X}; \alpha), \bar{\theta}(\mathbf{X}; \alpha))$  is just  $1 - 2\alpha$ . Nevertheless, the equivalence test based upon it has always significance level  $\alpha$  rather than  $2\alpha$ . The proof of this important fact is almost trivial and requires just a few lines: Let  $\theta$  be any point in the parameter space belonging to the left-hand part of  $H$  such that we have  $\theta \leq \theta_0 - \varepsilon_1$ . By

definition, the event that the interval inclusion test rejects can be written  $\{\theta_0 - \varepsilon_1 < \underline{\theta}(\mathbf{X}; \alpha), \bar{\theta}(\mathbf{X}; \alpha) < \theta_0 + \varepsilon_2\}$ . Hence, for  $\theta \leq \theta_0 - \varepsilon_1$ , an error of the first kind can only occur if we find  $\underline{\theta}(\mathbf{X}; \alpha) > \theta$  which means that the corresponding one-sided confidence region  $(\underline{\theta}(\mathbf{X}; \alpha), \infty)$  fails to cover the true value of  $\theta$ . By construction of  $\underline{\theta}(\mathbf{X}; \alpha)$ , this happens in at most  $100\alpha\%$  of the applications of the corresponding confidence procedure so that we have in fact  $P_\theta[\text{type-I error}] \leq \alpha$  for all  $\theta \leq \theta_0 - \varepsilon_1$ . A completely analogous argument shows that for any  $\theta \geq \theta_0 + \varepsilon_2$ , we have  $P_\theta[\text{type-I error}] \leq P_\theta[\bar{\theta}(\mathbf{X}; \alpha) \leq \theta] \leq \alpha$ . For all remaining points in the parameter space, i.e., for  $\theta_0 - \varepsilon_1 < \theta < \theta_0 + \varepsilon_2$ , there is nothing to show because rejecting  $H$  will be a correct decision then. Although the mathematical reasoning required for establishing the validity of interval inclusion tests based on confidence intervals at level  $(1 - 2\alpha)$  is so simple, most of the earlier work on the approach (see Westlake, 1981) used intervals of two-sided confidence level  $1 - \alpha$  yielding unnecessarily conservative testing procedures. (Nonequal tails confidence intervals at level  $1 - \alpha$  for which the associated equivalence tests do not suffer from such a marked conservatism, are constructed in the paper of Hsu et al., 1994.)

Another fact of considerable importance for a proper understanding of the confidence interval inclusion principle was first pointed out by Schuirmann (1987) for the special case of the setting of the two-sample  $t$ -test. Generally, any interval inclusion test for equivalence as defined above is logically equivalent to a combination of two one-sided tests. In fact, it is obvious from its definition that an interval inclusion test decides in favor of equivalence if and only if both the test of  $\theta \leq \theta_0 - \varepsilon_1$  vs.  $\theta > \theta_0 - \varepsilon_1$  based on  $\underline{\theta}(\mathbf{X}; \alpha)$ , and that of  $\theta \geq \theta_0 + \varepsilon_2$  vs.  $\theta < \theta_0 + \varepsilon_2$  using  $\bar{\theta}(\mathbf{X}; \alpha)$  as the test statistic, can reject its null hypothesis [→ “*double one-sided testing procedure*”]. As will be explained in detail in § 7.1, this means that the principle of confidence interval inclusion is just a special case of the so-called intersection-union principle introduced by Berger (1982) in the context of a problem of statistical quality control (see also Berger and Hsu, 1996).

For the setting of the two-sample  $t$ -test with  $\theta = \mu_1 - \mu_2$  and  $\theta_0 = 0$ , the  $(1 - 2\alpha)$ -confidence interval  $(\underline{\theta}(\mathbf{X}; \alpha), \bar{\theta}(\mathbf{X}; \alpha))$  introduced above is symmetric about the pivot  $\hat{\theta} = \bar{X} - \bar{Y}$  rather than 0. In Westlake (1976) it was argued that applying the interval inclusion rule with confidence intervals non-symmetric with respect to 0 is unsuitable whenever we deal with a symmetric equivalence hypothesis specifying that  $|\theta| = |\mu_1 - \mu_2| < \varepsilon$ . Adopting this point of view would imply that the test be based on a  $(1 - \alpha)$ -confidence interval of the form  $(\bar{Y} - \bar{X} - C_\alpha, \bar{X} - \bar{Y} + C_\alpha)$  with  $C_\alpha$  denoting a suitable real-valued random variable such that  $\bar{X} - \bar{Y} + C_\alpha > 0$ . However, on closer examination it becomes obvious that this suggestion relates the symmetry requirement to the wrong object. In fact, assuming two samples from homoskedastic Gaussian distributions, the testing problem  $|\mu_1 - \mu_2| \geq \varepsilon$  vs.  $|\mu_1 - \mu_2| < \varepsilon$  as such turns out to exhibit a clear-cut symmetry structure in that it remains invariant against treating the  $Y$ 's as making up the first rather than the second of the

two samples [formally: against the transformation  $(x_1, \dots, x_m, y_1, \dots, y_n) \mapsto (y_1, \dots, y_n, x_1, \dots, x_m)$ ]. In such a case, the well-accepted principle of invariance (cf. Cox and Hinkley, 1974, § 2.3 (vi); Lehmann and Romano, 2005, Ch. 6) requires that we use a decision rule leading to the same conclusion irrespective of whether or not the samples are labeled in the original manner. It is easy to check that the test based on the symmetric confidence interval  $(\bar{Y} - \bar{X} - C_\alpha, \bar{X} - \bar{Y} + C_\alpha)$  lacks this invariance property. In fact,  $(\bar{Y} - \bar{X} - C_\alpha, \bar{X} - \bar{Y} + C_\alpha)$  is included in the equivalence interval  $(-\varepsilon, \varepsilon)$  if and only if there holds  $\bar{X} - \bar{Y} + C_\alpha < \varepsilon$ , and the latter inequality is obviously not equivalent to  $\bar{Y} - \bar{X} + C_\alpha < \varepsilon$ . On the other hand, the test based on the shortest  $(1 - 2\alpha)$ -confidence interval symmetric about  $\bar{X} - \bar{Y}$  rejects if and only if it happens that  $|\bar{X} - \bar{Y}| < \varepsilon - S(X_1, \dots, X_m, Y_1, \dots, Y_n) \cdot t_{m+n-2;1-\alpha}$  where  $S(X_1, \dots, X_m, Y_1, \dots, Y_n) = \sqrt{1/m + 1/n} \cdot [(\sum_{i=1}^m (X_i - \bar{X})^2 + \sum_{j=1}^n (Y_j - \bar{Y})^2)/(m+n-2)]^{1/2}$  and  $t_{m+n-2;1-\alpha}$  denotes the upper  $100\alpha$  percentage point of a central  $t$ -distribution with  $m+n-2$  degrees of freedom. In view of  $S(X_1, \dots, X_m, Y_1, \dots, Y_n) = S(Y_1, \dots, Y_n, X_1, \dots, X_m)$ , the condition for rejecting nonequivalence is the same as  $|\bar{Y} - \bar{X}| < \varepsilon - S(Y_1, \dots, Y_n, X_1, \dots, X_m) \cdot t_{m+n-2;1-\alpha}$ . In other words, the nonsymmetric confidence interval leads to a symmetric test, and vice versa.

If we assume the common variance of the two normal distributions under comparison to be a known constant with respect to which all observations in both samples have been standardized, then the test of  $H : |\mu_1 - \mu_2| \geq \varepsilon$  vs.  $K : |\mu_1 - \mu_2| < \varepsilon$  based on shortest  $(1 - 2\alpha)$ -confidence intervals rejects if and only if we have

$$|\bar{X} - \bar{Y}| < \varepsilon - u_{1-\alpha} \cdot \sqrt{m^{-1} + n^{-1}}, \quad (3.1)$$

where  $u_{1-\alpha} = \Phi^{-1}(1-\alpha)$ . Trivially, condition (3.1) defines a nonempty region in the sample space only if the bound  $\varepsilon - u_{1-\alpha} \cdot \sqrt{m^{-1} + n^{-1}}$  represents a positive real number which is true only for sufficiently large sample sizes  $m, n$ . For example, in the balanced case  $m = n$  with  $\varepsilon = .25$  and  $\alpha = .05$ , (3.1) corresponds to a test which will never be able to reject nonequivalence unless the number  $n$  of observations available in both groups is at least 87. At first sight, this example seems to be fatal for the interval inclusion principle as a *general* approach to constructing equivalence tests. In fact, it refers to an extremely regular setting since the distribution of the sufficient statistic  $(\bar{X}, \bar{Y})$  belongs to a two-parameter exponential family and the distribution of the test statistic  $(\bar{X}, \bar{Y})$  is absolutely continuous for any parameter constellation  $(\mu_1, \mu_2)$ . Nevertheless, even if the common sample size is as large as 86, the testing procedure given by (3.1) is still poorer than the trivial level- $\alpha$  “test” deciding independently of the data in favor of equivalence whenever in some external random experiment an event of probability  $\alpha$  happens to get realized. On the other hand, as will be shown in § 4.4, this is well compatible with the fact that for sample sizes exceeding the minimum value of 87 sufficiently far, decision rule (3.1) yields a test which is practically indistinguishable from the optimal solution to exactly the same testing problem.

## 3.2 Bayesian tests for two-sided equivalence

The basic ingredients required for a Bayesian treatment of two-sided equivalence testing problems are the same as introduced in § 2.4 with regard to testing for noninferiority. Again, any Bayesian construction of a test starts from a fully specified joint prior distribution  $\pi(\cdot)$  of all unknown parameters appearing in the model behind the data under analysis, and the test rejects if the posterior probability of the region corresponding to the alternative hypothesis turns out to be larger than a suitably lower bound specified  $1 - \alpha$  as default. The only difference relates to the form of the region in the parameter space whose posterior probability has to be determined. In the case of two-sided equivalence testing, this is given by the bounded interval  $(\theta_0 - \varepsilon_1, \theta_0 + \varepsilon_2)$  so that the decision rule (2.7) for a Bayesian test for noninferiority has to be replaced with

Reject nonequivalence if and only if

$$\pi^{\theta|\mathbf{X}=\mathbf{x}}(\theta_0 - \varepsilon_1, \theta_0 + \varepsilon_2) \geq 1 - \alpha. \quad (3.2)$$

As in the noninferiority case, the Bayesian approach will exclusively be used in its “objective” version, i.e., with a noninformative prior chosen according to the rule proposed by Jeffreys (1961).

For illustration, let us consider the setting of the paired  $t$ -test with the intra-subject differences  $D_i = X_i - Y_i$  as the data eventually to be analyzed and the expected value  $\delta$  of the  $D_i$  as the parameter of interest. This is a case in which the posterior distribution of the target parameter with respect to the noninformative “reference prior” admits a simple explicit representation. As shown, e.g., in Lindley (1970, pp. 36–7) and Box and Tiao (1973, § 2.4), given the observed values  $\bar{d}$  and  $s_D$  of the mean  $\bar{D}$  and standard deviation  $s_D$  of the  $D_i$ ,  $n^{1/2}(\delta - \bar{d})/s_D$  has a posterior distribution then which is central  $t$  with the usual number  $n - 1$  of degrees of freedom. Hence, the posterior probability of the equivalence interval specified by the alternative hypothesis whose limits are this time denoted by  $\delta_1, \delta_2$  can be written

$$\pi^{\delta|\mathbf{D}=\mathbf{d}}(\delta_1, \delta_2) = F_{n-1}^T(n^{1/2}(\delta_2 - \bar{d})/s_D) - F_{n-1}^T(n^{1/2}(\delta_1 - \bar{d})/s_D) \quad (3.3)$$

where  $F_{n-1}^T(\cdot)$  stands for the cdf of a central  $t$ -distribution with  $n - 1$  degrees of freedom. Suppose specifically that we obtained  $\bar{d} = .16$ ,  $s_D = 3.99$  in a sample of size  $n = 23^\dagger$  and that  $\delta_1 = -1.75$ ,  $\delta_2 = 1.75$  have been chosen as equivalence limits for  $\delta$ . Then, evaluating the expression on the right-hand side of (3.3) by means of the SAS function `probt` gives for the posterior probability of  $(\delta_1, \delta_2)$ :

<sup>†</sup>The source of these data is a pilot study (described in some more detail in Example 5.3, p. 96) of the temporal stability of the capillary flow in the brain of rabbits.

$$\begin{aligned}\pi^{\delta|\mathbf{D}=\mathbf{d}}(-1.75, 1.75) &= \\ F_{22}^T(23^{1/2}(1.75 - .16)/3.99) - F_{22}^T(23^{1/2}(-1.75 - .16)/3.99) &= \\ F_{22}^T(1.911121) - F_{22}^T(-2.295749) &= .965447 - .015796 = .949651.\end{aligned}$$

Hence, at the nominal level  $\alpha = 0.05$ , for this specific data set the Bayesian test (3.2) with respect to the reference prior distribution of  $(\delta, \sigma_D)$  would not be able to reject nonequivalence in the sense of  $|\delta| \geq 1.75$ .

It is worth noticing that in situations where Bayesian credible intervals coincide with classical confidence intervals, the test obtained by applying decision rule (3.2) is the same as the interval inclusion test with confidence limits at one-sided confidence level  $1 - \alpha/2$  rather than  $1 - \alpha$ . Recalling the correspondence between interval inclusion tests and double one-sided tests for equivalence, this suggests that for the sake of avoiding overconservatism, the double one-sided Bayesian test given by the decision rule

Reject nonequivalence if and only if

$$\pi^{\theta|\mathbf{X}=\mathbf{x}}(\theta_0 - \varepsilon_1, \infty) \geq 1 - \alpha \quad \text{and} \quad \pi^{\theta|\mathbf{X}=\mathbf{x}}(-\infty, \theta_0 + \varepsilon_2) \geq 1 - \alpha \quad (3.4)$$

should generally preferred to the “direct” Bayesian equivalence test (3.2). In the numerical example introduced before, the posterior probabilities of the two infinite intervals to be intersected in order to get the equivalence interval itself, are computed to be  $\pi^{\delta|\mathbf{D}=\mathbf{d}}(-1.75, \infty) = 1 - F_{22}^T(-2.295749) = 1 - .015796 = .984204$ ,  $\pi^{\delta|\mathbf{D}=\mathbf{d}}(-\infty, 1.75) = F_{22}^T(1.911121) = .965447$ . Since with  $\alpha = .05$ , both of these values exceed the critical lower bound  $1 - \alpha$ , the double one-sided Bayesian test (3.4), in contrast to the direct Bayesian decision rule (3.2), leads to a positive result which reflects its improved power.

As any technique of Bayesian inference, the approach to equivalence testing given by (3.2) or (3.4) is attractive not only for its conceptual simplicity, but likewise for its high flexibility with respect to reparametrizations of the problem. Like the prior, the posterior distribution is determined jointly for all unknown parameters of the model under analysis. Hence, it can easily be calculated for any sufficiently regular function of the primary parameters by means of the well-known transformation theorem for probability densities (cf. Bickel and Doksum, 2001, §B.2). In order to illustrate this technique, we keep considering the paired *t*-test setting and demonstrate how to compute the posterior density of the standardized expected difference  $\theta = \delta/\sigma_D$  assuming the same prior distribution of  $(\delta, \sigma_D)$  which underlies formula (3.3). For that purpose, the following results (to be found again in the book of Box and Tiao, 1973, pp. 95–6) are needed:

- 1) The conditional posterior distribution of  $\delta$  given  $\sigma_D$  is  $\mathcal{N}(\bar{d}, \sigma_D^2/n)$  [i.e., Gaussian with mean  $\bar{d}$  and variance  $\sigma_D^2/n$ ].
- 2) The marginal posterior distribution of  $\sigma_D$  is that of  $s_D(n-1)^{1/2}/X_{n-1}$  with  $X_{n-1}^2 \sim \chi_{n-1}^2$  [= central  $\chi^2$ -distribution with  $n-1$  degrees of freedom].

From 1) and 2), it follows that the joint posterior density of  $(\delta, \sigma_D)$  can be written:

$$f(\delta, \sigma_D | \mathbf{D} = \mathbf{d}) = (n^{1/2}/\sigma_D) \varphi(n^{1/2}(\delta - \bar{d})/\sigma_D) (\Gamma((n-1)/2))^{-1} \times \\ ((n-1)s_D^2)^{(n-1)/2} 2^{(3-n)/2} \sigma_D^{-n} \exp \left\{ -((n-1)/2)(s_D/\sigma_D)^2 \right\} \quad (3.5)$$

with  $\Gamma(\cdot)$  and  $\varphi(\cdot)$  denoting the gamma function and the standard Gaussian density, respectively. (3.5) leads to the following expression for the posterior probability of the hypothesis  $\theta_1 < \theta < \theta_2$  with  $\theta = \delta/\sigma_D$  and arbitrarily fixed equivalence limits  $-\infty < \theta_1 < \theta_2 < \infty$ :

$$\pi^{\theta | \mathbf{D} = \mathbf{d}}(\theta_1, \theta_2) = \int_0^\infty \left[ (\Phi(n^{1/2}(\theta_2\sigma_D - \bar{d})/\sigma_D) - \Phi(n^{1/2}(\theta_1\sigma_D - \bar{d})/\sigma_D)) \times \right. \\ (\Gamma((n-1)/2))^{-1} ((n-1)s_D^2)^{(n-1)/2} 2^{(3-n)/2} \sigma_D^{-n} \times \\ \left. \exp \left\{ -((n-1)/2)(s_D/\sigma_D)^2 \right\} \right] d\sigma_D. \quad (3.6)$$

The integral on the right-hand side of equation (3.6) can be evaluated by means of a suitable quadrature rule to any desired degree of numerical accuracy. A numerical method which serves remarkably well with integrals of that type is Gaussian quadrature (cf. Davis and Rabinowitz, 1975, § 2.7). The error entailed by truncating the interval of integration from above at some point  $c_\circ$  can be made smaller than any prespecified  $\varepsilon_\circ > 0$  [e.g.,  $\varepsilon_\circ = 10^{-6}$ ] by choosing  $c_\circ > \sqrt{(n-1)s_D^2/\chi_{n-1; \varepsilon_\circ}^2}$  with  $\chi_{n-1; \varepsilon_\circ}^2$  denoting the  $100\varepsilon_\circ$ -percentage point of a central  $\chi^2$ -distribution with  $n-1$  degrees of freedom. The program `postmlys` to be found in the **WKTSEQ2 Source Code Package** both as a SAS macro and a R function implements this computational approach using a grid of 96 abscissas (for a table of all constants involved in 96-point Gaussian quadrature see Abramowitz and Stegun, 1965, p. 919). In the specific case  $n = 23$ ,  $\bar{d} = .16$ ,  $s_D = 3.99$  used for illustration once more, it gives for the posterior probability of the equivalence region  $\{(\delta, \sigma_D) | -0.5 < \delta/\sigma_D < 0.5\}$  the value  $\pi^{\theta | \mathbf{D} = \mathbf{d}}(-0.5, 0.5) = .981496$ . Hence, with these data the Bayesian test for equivalence with respect to  $\delta/\sigma_D$  leads to the same decision as the optimal frequentist testing procedure to be discussed in § 5.3 (see in particular p. 96).

Clearly, the most crucial issue to be raised in discussing the suitability of the decision rules (3.2) and (3.4) concerns the point that in the corresponding test maintenance of the prespecified significance level in the classical sense cannot generally be taken for granted. Since the latter is still considered an indispensable property of any procedure to be relied upon for confirmatory purposes by the majority of the users of statistical methodology [in particular, the regulatory authorities deciding upon the eventual success of drug development projects], it seems advisable to check Bayesian decision rules from a

frequentist perspective as carefully as possible before adopting them for real applications. Unfortunately, deriving sufficient conditions for the validity of Bayesian tests with regard to the risk of a type-I error is still more difficult a task in the case of equivalence than that of conventional one-sided hypotheses. In fact, the results obtained by Casella and Berger (1987) for the latter cannot be exploited for establishing the validity of Bayesian equivalence tests even if attention is restricted to simple location-parameter problems decomposed into two one-sided problems by replacing (3.2) with (3.4). The reason is that the results proved by these authors require symmetry of the density of the data about the common boundary of null and alternative hypothesis, and it can easily be shown that there exists no proper probability density which is symmetric about two different points [viz.,  $\theta_0 - \varepsilon_1$  and  $\theta_0 + \varepsilon_2$ ].

An abstract concept which proves worthwhile as a starting-point for investigations into possible relationships between Bayesian and frequentist tests for equivalence is that of a confidence distribution as introduced by Cox (1958) and exploited for a systematic discussion of the interval inclusion approach to equivalence testing by Mau (1987, 1988). Roughly speaking, a confidence distribution is defined to be a random probability measure [depending in some mathematically clear-cut way on the data under analysis] such that its quantiles give confidence bounds to the parameter  $\theta$  of interest. Whenever the posterior distribution of  $\theta$  admits a representation as a realization of a random measure of that type, both versions (3.2) and (3.4) of a Bayesian rule for deciding between equivalence and existence of relevant differences coincide with interval inclusion rules as introduced in the previous section and hence share with them the property of being valid with regard to the type-I error risk. Unfortunately, representability of Bayesian posterior as frequentist confidence distributions turns out to be a highly restrictive condition. In fact, from a practical point of view, it provides hardly more than a characterization in abstract terms of situations where the distributions of the data (typically after suitable reductions by sufficiency) belong either to a location family  $(P_\theta)_{\theta \in \mathbb{R}}$ , a scale family  $(P_\sigma)_{\sigma > 0}$ , or a location-scale family  $(P_{\theta,\sigma})_{(\theta,\sigma) \in \mathbb{R} \times \mathbb{R}_+}$ , and the prior is the usual reference prior defined by an improper density proportional to 1 or  $\sigma^{-1}$ , respectively. However, for most models involving only a location or a scale parameter or both, there exist well-established methods of interval estimation and even optimal tests for equivalence are comparatively easy to implement. In view of this, the possibility of establishing the validity of a Bayesian test for equivalence by way of verifying that the posterior distribution involved satisfies the definition of an observed confidence distribution is mainly of theoretical and conceptual interest. Moreover, for purposes of proving the validity of equivalence tests using decision rule (3.4) with a suitable noninformative reference prior, an alternative method exists which avoids detailed comparative inspections of the underlying posterior distributions. It is obtained by exploiting the theory of right-invariant Haar densities and provides, at least potentially, a more versatile tool than the approach via paralleling posterior and confidence distributions, since it covers transforma-

tion families of any kind (for a detailed exposition of this theory see Berger, 1985, § 6.6).

Notwithstanding the considerable difficulties encountered in proving analytically that a Bayesian testing procedure satisfies at the same time basic frequentist criteria, the proportion of Bayesian contributions to the literature on statistical methods for equivalence assessment has been substantial all along. This has especially been true in the field of the analysis of comparative bioavailability trials. A representative selection of influential papers on Bayesian methods of bioequivalence assessment is constituted of Selwyn et al. (1981); Rodda and Davis (1980); Flühler et al. (1983); Mandallaz and Mau (1981); Selwyn and Hall (1984) and Racine-Poon et al. (1987). Furthermore, in later chapters, the reader will become acquainted with a number of instances of Bayesian constructions of tests for equivalence and noninferiority for which ensuring validity is done via direct numerical determination of the maximum rejection probability taken over the respective null hypothesis. In § 10.3, we will even see that the Bayesian approach has the potential of filling gaps in the repertoire of useful frequentist (including asymptotic) solutions to equivalence testing problems of considerable practical relevance.

---

### 3.3 The classical approach to deriving optimal parametric tests for equivalence hypotheses

In the classical theory of hypotheses testing, statistical tests are derived as solutions to optimization problems arising from the endeavor to maximize the power over a class of alternative decision procedures chosen as large as possible. Ideally, this class contains the totality of all tests of the null hypothesis under consideration which maintain the prespecified significance level, and maximization of power can be achieved uniformly over the whole subset of the parameter space specified by the alternative hypothesis. In the majority of situations occurring in practice, the class of testing procedures to be taken into consideration must be restricted by means of supplementary criteria to be imposed in addition to that of validity with respect to the type-I error risk. The most useful of them are based on the principle of sufficiency and invariance, respectively. The basic mathematical results which the construction of such optimal testing procedures for two-sided equivalence hypotheses relies upon, are presented in due rigor and generality in § A.1 of the Appendix. Hence, it suffices to give in the present section a brief outline of the basic techniques available for purposes of carrying out optimal constructions of tests for equivalence.

As is the case in the treatment of one-sided problems [including problems of testing for noninferiority — recall § 2.2] and conventional two-sided testing

problems, the natural starting-point of the classical approach to the construction of tests for equivalence is the consideration of settings where the possible distributions of the data form a one-parameter family of sufficiently regular structure. When dealing with interval hypotheses, the suitable way of making this regularity precise is based on the concept of total positivity whose fundamental importance for the theory of statistical decision procedures has been demonstrated by S. Karlin in a series of frequently cited papers and an encyclopedic monograph (Karlin, 1956, 1957a, b, 1968). For purposes of deriving optimal one-sided tests, the assumption of monotone likelihood ratios (or, equivalently, of strict total positivity of order 2) proves mathematically natural. In contrast, in order to ensure the existence of optimal tests for equivalence, families of distributions are required whose densities are STP<sub>3</sub> (strictly totally positive of order 3 — see Definition A.1.1 in Appendix A) and depend in a continuous manner both on the realized value of the observed random variable and the parameter  $\theta$ . The basic fact is then that in any STP<sub>3</sub> family exhibiting these additional continuity properties, an optimal test for equivalence in the sense made precise in (1.1 a,b) [→ p. 11] can be constructed by means of the generalized Fundamental Lemma of Neyman and Pearson (1936). Interestingly, this construction whose mathematical basis is given by Theorem A.1.5 of Appendix A, leads to tests which are uniformly most powerful (UMP) among all tests at level  $\alpha$  of the same hypothesis whereas for the dual testing problem  $\theta_0 - \varepsilon_1 \leq \theta \leq \theta_0 + \varepsilon_2$  versus  $\theta < \theta_0 - \varepsilon_1$  or  $\theta > \theta_0 + \varepsilon_2$  (usually termed the problem of testing for relevant differences by biostatisticians — see Ch. 11) only a uniformly most powerful unbiased (UMPU) solution exists (Lehmann and Romano, 2005, p. 83).

As suggested by intuition, every reasonably structured test for equivalence has a form which is complementary to that of a traditional two-sided test. Thus, optimal solutions to some given problem of equivalence testing have to be sought for in the class of tests whose rejection region can be written as

$$\{ \mathbf{x} \mid C_1 < T(\mathbf{x}) < C_2 \} , \quad (3.7)$$

which means that the null hypothesis  $H$  of nonequivalence has to be rejected as long as the observed value of a suitable real-valued statistic remains *within some sufficiently short interval*. The limits  $C_1, C_2$  of that critical interval may depend, in addition to the significance level  $\alpha$ , on the value observed for some other statistic  $S$ , say, which will typically be chosen sufficient for a (maybe multidimensional) nuisance parameter. In order to obtain an optimal test,  $C_1$  and  $C_2$  must be determined in such a way that its (conditional) probability takes on exact value  $\alpha$  at both boundaries of the null hypothesis, i.e., under both  $\theta = \theta_0 - \varepsilon_1$  and  $\theta = \theta_0 + \varepsilon_2$ . In cases where the distributions of the test statistic  $T(\mathbf{X})$  are continuous, this implies that  $C_1, C_2$  must simultaneously satisfy the equations

$$\begin{aligned} P_{\theta_1}[C_1 < T(\mathbf{X}) < C_2] &= \alpha = \\ P_{\theta_2}[C_1 < T(\mathbf{X}) < C_2], \quad , C_1, C_2 \in \mathcal{T}, \quad C_1 < C_2 \end{aligned} \quad (3.8)$$

where  $\mathcal{T}$  denotes the range of  $T(\cdot)$  and  $\theta_1 = \theta_o - \varepsilon_1$ ,  $\theta_2 = \theta_o + \varepsilon_2$ . If the distribution functions of  $T(\mathbf{X})$  exhibit jump discontinuities, a solution to (3.8) generally does not exist. Then, one has to deal with the more complicated system

$$\begin{aligned} P_{\theta_1} [C_1 < T(\mathbf{X}) < C_2] + \sum_{\nu=1}^2 \gamma_\nu P_{\theta_1} [T(\mathbf{X}) = C_\nu] &= \alpha = \\ P_{\theta_2} [C_1 < T(\mathbf{X}) < C_2] + \sum_{\nu=1}^2 \gamma_\nu P_{\theta_2} [T(\mathbf{X}) = C_\nu], \\ C_1, C_2 \in \mathcal{T}, \quad C_1 \leq C_2, \quad 0 \leq \gamma_1, \gamma_2 < 1 \end{aligned} \tag{3.9}$$

instead. The existence of a unique solution to (3.9) is guaranteed whenever the family of distributions of  $T(\mathbf{X})$  is STP<sub>3</sub>, and the optimal test entails a randomized decision in favor of equivalence with probability  $\gamma_1$  or  $\gamma_2$  if  $T(\mathbf{X})$  takes on value  $C_1$  or  $C_2$ , respectively. As has been pointed out in the previous chapter in connection with exact tests for noninferiority problems relating to discrete families of distributions, in the overwhelming majority of applications of statistical testing procedures to real data randomized decisions between the hypotheses are clearly undesirable. Accordingly, in the noncontinuous case the optimal test for equivalence is usually modified in the following way: The critical interval  $(C_1, C_2)$  has still to be determined by solving (3.9) but at its boundaries, i.e., both for  $T(\mathbf{X}) = C_1$  and for  $T(\mathbf{X}) = C_2$ , the null hypothesis is always accepted (corresponding to the specification  $\gamma_1 = 0 = \gamma_2$ ). In situations where the distributions of  $T(\mathbf{X})$  are discrete and at the same time conditional on some other statistic sufficient for the nuisance parameter(s) involved in the underlying model [→ §§5.1, 6.6.5, 9.4.2], adopting the technique of raised nominal significance levels introduced by Boschloo (1970) [recall §2.5] will prove a simple and effective device for reducing the conservatism entailed by accepting nonequivalence whenever we observe that  $T(\mathbf{X}) \in \{C_1, C_2\}$ .

In the *continuous case*, numerical determination of the optimal critical constants  $C_1, C_2$  is greatly facilitated by the availability of a general result (proved by Lehmann and Romano, 2005, as Lemma 3.4.2(iv)) on the sign of the difference of the power functions of two tests of the form (3.7) in families with monotone likelihood ratios, which implies the following fact: Let the family of distributions of  $T(\mathbf{X})$  be STP<sub>3</sub> and  $(C'_1, C'_2)$  denote any interval on the real line such that  $P_{\theta_1}[C'_1 < T(\mathbf{X}) < C'_2] = \alpha$  but  $P_{\theta_2}[C'_1 < T(\mathbf{X}) < C'_2] <$  or  $> \alpha$ . Then it follows that the optimal critical interval  $C_1, C_2$  whose endpoints satisfy (3.8) lies to the right or left of  $(C'_1, C'_2)$ , respectively. Clearly, this shows that a simple iteration scheme for approximating the solution of (3.8) to any desired degree of numerical accuracy, consists of the following steps:

- (i) Choose some estimate  $C_1^0$ , say, of the left-hand limit  $C_1$  of the optimal critical interval.

- (ii) Compute the corresponding upper critical bound  $C_2^0$  as  $C_2^0 = F_{\theta_1}^{-1}[\alpha + F_{\theta_1}(C_1^0)]$ , with  $F_\theta$  and  $F_\theta^{-1}$  denoting the distribution and quantile function of  $T(\mathbf{X})$ , respectively, at an arbitrary point  $\theta$  in the parameter space.
- (iii) Compute the size  $\alpha_2^0 = F_{\theta_2}(C_2^0) - F_{\theta_2}(C_1^0)$  of the critical region  $\{\mathbf{x} \mid C_1^0 < T(\mathbf{x}) < C_2^0\}$  at  $\theta = \theta_2$ .
- (iv) Update  $C_1^0$  by replacing it with  $(\tilde{C}_1 + C_1^0)/2$  or  $(C_1^0 + \tilde{\tilde{C}}_1)/2$  depending on whether there holds  $\alpha_2^0 > \alpha$  or  $\alpha_2^0 < \alpha$  and assuming that  $(\tilde{C}_1, \tilde{\tilde{C}}_1)$  has been previously determined as an interval containing both  $C_1^0$  and the exact solution  $C_1$ .
- (v) Repeat steps (i)–(iv) as many times as required in order to ensure that  $|\alpha_2^0 - \alpha| < \text{TOL}$  with, e.g.,  $\text{TOL} = 10^{-4}$ .

In the *discrete case with integer-valued test statistic  $T(\mathbf{X})$* , extensive experience has shown that the following search algorithm (forming the basis of the programs **bi1st**, **powsign**, **bi2st**, **bi2aeq1 – bi2aeq3** and **gofhwex**) works very efficiently on finding the solution of (3.9):

- (i\*) Choose an initial value  $C_1^0$  for the lower critical bound known to be greater or equal to the correct value  $C_1$ .
- (ii\*) Keeping  $C_1^0$  fixed, find the largest integer  $C_2^0 > C_1^0$  such that the probability of observing  $T(\mathbf{X})$  to take on its value in the closed interval  $[C_1^0 + 1, C_2^0 - 1]$  does not exceed  $\alpha$ , neither for  $\theta = \theta_1$  nor for  $\theta = \theta_2$ .
- (iii\*) Treat (3.9) as a system of linear equations in the two unknowns  $\gamma_1, \gamma_2$  and compute its solution  $\gamma_1^0, \gamma_2^0$ , say (which of course depends on  $C_1^0$  and  $C_2^0$ ).
- (iv\*) Verify the condition  $0 \leq \gamma_1^0, \gamma_2^0 < 1$ . If it is satisfied, the solution of the full system (3.9) is found and given as  $(C_1, C_2, \gamma_1, \gamma_2) = (C_1^0, C_2^0, \gamma_1^0, \gamma_2^0)$ ; if not, diminish  $C_1^0$  by 1 and repeat steps (ii\*) and (iii\*).

Although both of the algorithms outlined above are logically straightforward, concrete numerical determination of the optimal critical constants is a much more demanding task for equivalence problems than for one- or ordinary two-sided testing problems. This explains why in the existing literature contributions presenting classical tests for equivalence as fully explicit decision rules ready for being used by the working statistician, have been sparse for several decades. By the time of appearance of the first edition of this book, the following list has been largely exhaustive: Bondy (1969); Wellek

and Michaelis (1991); Wellek (1993a); Mehring (1993); Wellek (1994); Wang (1997).

Perhaps, the main fact making optimal tests for equivalence (as well as those for noninferiority) computationally more complicated than their traditional one- and two-sided analogues, is that the sampling distributions of customary test statistics are needed in their *noncentral* versions in order to determine sizes of rejection regions etc. An additional complication refers to situations entailing elimination of nuisance parameters by means of conditioning: The possibility of reducing the conditional tests primarily obtained to equivalent nonconditional procedures, which is well known (cf. Lehmann and Romano, 2005, § 5.1) to lead not infrequently to considerable simplifications with one- and two-sided problems, generally does not exist in the equivalence case. Nevertheless, in subsequent chapters the reader will become acquainted with numerous specific equivalence testing problems for which an optimal procedure can be made available in a form suitable for routine use by anybody having access to standard computational tools.

The computational effort entailed by performing an optimal test for equivalence in practice, reduces to a minimum in all cases where the following bipartite symmetry condition is satisfied:

$$\theta_1 = -\varepsilon, \theta_2 = \varepsilon \quad \text{for some } \varepsilon > 0; \quad (3.10a)$$

the distribution of  $T(\mathbf{X})$  under  $\theta = \varepsilon$  coincides

$$\text{with that of } -T(\mathbf{X}) \text{ under } \theta = -\varepsilon. \quad (3.10b)$$

In fact, by the result made precise and proved in the Appendix as Lemma A.1.6, (3.10) implies that the critical interval for  $T(\mathbf{X})$  defining an optimal test for equivalence is symmetric about zero, i.e., of the form  $(-C, C)$ . In the continuous case,  $C$  remains as the only critical constant uniquely determined by the single equation

$$P_\varepsilon[|T(\mathbf{X})| < C] = \alpha, \quad C \in \mathcal{T} \cap (0, \infty). \quad (3.11)$$

An alternative formulation of the same fact is as follows: If, in a situation being symmetric in the sense of (3.10), there exists an optimal test for equivalence, then this can be carried out by means of a p-value defined as the lower tail probability of the absolute value of the test statistic under  $\theta = \varepsilon$  or, equivalently, under  $\theta = -\varepsilon$ . Dropping the continuity assumption, one has  $\gamma_1 = \gamma_2 = \gamma$  in addition to  $-C_1 = C_2 = C$  where  $C$  and  $\gamma$  are defined by

$$C = \max \left\{ t \in \mathcal{T} \cap (0, \infty) \mid P_\varepsilon [|T(\mathbf{X})| < t] \leq \alpha \right\}, \quad (3.12a)$$

$$\gamma = \left( \alpha - P_\varepsilon [|T(\mathbf{X})| < C] \right) / P_\varepsilon [|T(\mathbf{X})| = C]. \quad (3.12b)$$

In particular, an optimal test for equivalence allows this kind of reduction to a one-sided test in terms of the absolute value of the respective test statistic,

whenever  $\theta$  plays the role of a location parameter for the distribution of  $T(\mathbf{X})$ , provided  $\theta = 0$  implies  $-T(\mathbf{X}) \stackrel{d}{=} T(\mathbf{X})$  and we are willing to choose the equivalence range for  $\theta$  as a symmetric interval. The most important special case of such a location problem occurs if  $T(\mathbf{X})$  is normally distributed with unknown expected value  $\theta \in \mathbb{R}$  and fixed known variance  $\sigma_0^2 > 0$  which will be treated in full detail in § 4.1.

The greater part of the equivalence assessment procedures described in the subsequent chapters are tests which maximize the power function uniformly over the class of all valid tests for the same problem, of invariant tests maintaining the prespecified level  $\alpha$ , of all unbiased tests at level  $\alpha$ , or of all valid tests for an associated asymptotic problem. Although optimal testing procedures generally deserve preference over less powerful tests of the same hypotheses, in providing a sufficiently rich repertoire of equivalence testing procedures unification of the constructional approach seems neither desirable nor promising. In fact, there are many rather elementary equivalence testing problems encountered very frequently in practice which admit no optimal solution. Perhaps, the best known and most important of them refers to location equivalence of two Gaussian distributions  $\mathcal{N}(\mu_1, \sigma^2), \mathcal{N}(\mu_2, \sigma^2)$  with common unknown variance. Here, even the construction of an unbiased test for  $|\mu_1 - \mu_2| \geq \varepsilon$  vs.  $|\mu_1 - \mu_2| < \varepsilon$  as carried out by Brown et al. (1997) proves mathematically quite complicated and leads to rejection regions which in view of several counterintuitive properties seem hardly suitable for practice (for critical discussions of this procedure see Hauck and Anderson, 1996; Meredith and Heise, 1996; Schuirmann, 1996; Perlman and Wu, 1999). Although the interval inclusion rule with standard equal-tails confidence intervals at two-sided level  $(1 - 2\alpha)$  gives a solution to this specific problem which still can be strongly recommended, it would be clearly misleading to suppose that the principle of confidence interval inclusion is generally easier to implement than the classical approach to constructing equivalence tests. One out of many other important examples showing this not to hold true occurs if, in the setting of the ordinary two-sample  $t$ -test, equivalence of  $\mathcal{N}(\mu_1, \sigma^2)$  and  $\mathcal{N}(\mu_2, \sigma^2)$  is redefined by requiring that we should have  $|\mu_1 - \mu_2|/\sigma < \varepsilon$ : Computing exact confidence limits for the difference between the standardized means of these distributions is certainly at least as complicated as determining the critical constant of the uniformly most powerful invariant (UMPI) test for the same problem as discussed in § 6.1.

### 3.4 Construction of asymptotic tests for equivalence

In § 2.3 it has been pointed out with regard to noninferiority testing that the range of problems accessible to exact methods is much too limited for cover-

ing the needs arising in practice. This statement applies at least as much to the case of two-sided equivalence testing. In this section, we will show that an asymptotic approach which enables one to derive solutions for settings in which exact methods are not in sight, is also available for problems of testing for equivalence in the strict sense. However, as compared to noninferiority problems, equivalence problems turn out to be considerably more complicated to solve also in an asymptotic framework. The theoretical basis of the asymptotic construction of tests for equivalence is given by a result which has been established for the two-sample case by Wellek (1996). In the Appendix [→ § A.3], it is rigorously shown to generalize in a way allowing us to cover virtually any situation where the equivalence hypothesis to be tested refers to a parameter or (in non- or semiparametric models) to a functional for which an asymptotically normal estimator can be found. In what follows, we give an informal yet exhaustive description of the general form of an asymptotic test for equivalence.

As in § 2.3, we start from the basic assumption that for the target parameter (or functional)  $\theta$  in terms of which the equivalence hypothesis under consideration is formulated, there exists an estimator  $(T_N)_{N \in \mathbb{N}}$  such that  $\sqrt{N}(T_N - \theta)$  is asymptotically normal with zero mean and variance  $\sigma^2$ . Again,  $N$  denotes the total number of observations making up the data set under analysis, and  $\sigma^2$  is another functional of the vector  $(F_1, \dots, F_k)$  of underlying distribution functions taking on only positive values.

Now, the basic steps leading to an asymptotically valid test of  $H : \theta \leq \theta_1$  or  $\theta \geq \theta_2$  versus  $K : \theta_1 < \theta < \theta_2$  are as follows:

- (i) Take (2.5) as if allowing to be strengthened to a statement on the exact sampling distribution of  $T_N$ .
- (ii) Treat the standard error  $\sigma/\sqrt{N}$  of  $T_N$  as a known constant  $\tau_N$ , say, and carry out in terms of  $T_N$  the test which is optimal for  $H$  vs.  $K$  when  $\theta$  stands for the expected value of a normal distribution with known variance  $\tau_N^2$  from which a single observation has been taken
- (iii) Derive an estimator  $\hat{\sigma}_N$  which is weakly consistent for  $\sigma$  and plug in  $\hat{\tau}_N = \hat{\sigma}_N/\sqrt{N}$  in the expressions obtained in step (ii) at every place where  $\tau_N$  appears.

The results derived in § 4.1 strongly suggest to use in step (ii) of such a construction the rejection region

$$\left\{ |T_N - (\theta_1 + \theta_2)/2|/\tau_N < C_\alpha((\theta_2 - \theta_1)/2\tau_N) \right\} \quad (3.13)$$

where for any  $\psi > 0$  and  $0 < \alpha < 1$ ,  $C_\alpha(\psi)$  has to be taken equal to the square root of the  $\alpha$ -quantile of a  $\chi^2$ -distribution with a single degree of freedom and noncentrality parameter  $\psi^2$ . Since the computation of quantiles of noncentral  $\chi^2$ -distributions is a process perfectly supported by the most widely used statistical software (see the intrinsic function `cinv` of SAS or the `qchisq` function

of the R Stats Package), the practical implementation of an asymptotic test for equivalence on the basis of (3.13) is an easy exercise as soon as the test statistic  $T_N$  as well as its estimated asymptotic standard error  $\hat{\tau}_N$  has been calculated. What remains to be done is to determine the ordinary distance in units of  $\hat{\tau}_N$  between  $T_N$  and the middle of the equivalence limits specified by the hypothesis, and compare the result with the number  $C_\alpha((\theta_2 - \theta_1)/2\hat{\tau}_N)$  as a critical upper bound. If the observed value of  $|T_N - (\theta_1 + \theta_2)/2|/\hat{\tau}_N$  is found to be smaller than this bound, nonequivalence can be rejected, and has to be accepted otherwise. Of course, replacing the theoretical standard error  $\tau_N$  with its consistent estimator  $\hat{\tau}_N$  in determining the noncentrality parameter of the  $\chi^2$ -distribution to be relied upon, makes  $C_\alpha((\theta_2 - \theta_1)/2\hat{\tau}_N)$  a random critical bound rather than a critical constant. Nevertheless, very mild conditions suffice to guarantee that the approach yields a test for equivalence which is asymptotically valid with respect to the significance level.

Usually, the most laborious step to be taken in deriving an asymptotic solution to a concrete equivalence testing problem on the basis of the theory outlined in this section, is to find an explicit expression for the asymptotic standard error of the test statistic as well as for a suitable estimator of it. Except for cases in which the complications entailed in variance calculation prove prohibitive, the approach is very useful. One of its nicest and most important properties stems from the finite-sample behavior of the corresponding testing procedures which according to the results of extensive simulation studies often turns out to be remarkably good. As one out of not a few noteworthy instances of this, the reader is referred to the log-rank test for equivalence of two survivor functions described in detail in § 6.7. Even if the expected proportion of censored observations contained in the data is about 50% and both sample sizes are as small as 25, the risk of committing a type-I error does not exceed the nominal significance level to a practically relevant extent.

---

## *Equivalence tests for selected one-parameter problems*

---

### 4.1 The one-sample problem with normally distributed observations of known variance

In this section it is assumed throughout that the hypotheses (1.1.a) and (1.1.b) [→ p. 11] refer to the expected value of a single Gaussian distribution whose variance has some fixed known value. In other words, we suppose that the data under analysis can be described by a vector  $(X_1, \dots, X_n)$  of mutually independent random variables having all distribution  $\mathcal{N}(\theta, \sigma_o^2)$  where  $\sigma_o^2 > 0$  denotes a fixed positive constant. It is easy to verify that it entails no loss of generality if we set  $\sigma_o^2 = 1$ ,  $\theta_o - \varepsilon_1 = -\varepsilon$  and  $\theta_o + \varepsilon_2 = \varepsilon$  with arbitrarily fixed  $\varepsilon > 0$ . In fact, if the variance of the primary observations differs from unity and/or the equivalence interval specified by the alternative hypothesis (1.1.b) fails to exhibit symmetry about zero, both of these properties can be ensured by applying the following simple transformation of the observed variables:  $X'_i = (X_i - \theta_o - (\varepsilon_2 - \varepsilon_1)/2)/\sigma_o$ . In view of the relations  $E_{\theta_o - \varepsilon_1}(X'_i) = (\theta_o - \varepsilon_1 - \theta_o - (\varepsilon_2 - \varepsilon_1)/2)/\sigma_o = -(\varepsilon_1 + \varepsilon_2)/2\sigma_o$ ,  $E_{\theta_o + \varepsilon_2}(X'_i) = (\varepsilon_1 + \varepsilon_2)/2\sigma_o$ , it makes no difference whether we use the original sample to test  $\theta \leq \theta_o - \varepsilon_1$  or  $\theta \geq \theta_o + \varepsilon_2$  versus  $\theta_o - \varepsilon_1 < \theta < \theta_o + \varepsilon_2$ , or base a test of the null hypothesis  $\theta' \leq -\varepsilon$  or  $\theta' \geq \varepsilon$  versus the alternative  $-\varepsilon < \theta' < \varepsilon$  on the transformed sample  $(X'_1, \dots, X'_n)$ , provided we define  $\theta' = E(X'_i)$  and  $\varepsilon = (\varepsilon_1 + \varepsilon_2)/2\sigma_o$ . To simplify notation, we drop the distinction between primed and unprimed symbols and assume that we have

$$X_i \sim \mathcal{N}(\theta, 1), \quad i = 1, \dots, n \tag{4.1a}$$

and that the equivalence problem referring to these observations reads

$$H : \theta \leq -\varepsilon \text{ or } \theta \geq \varepsilon \quad \text{versus} \quad K : -\varepsilon < \theta < \varepsilon. \tag{4.1b}$$

Since (4.1a) implies that the sample mean  $\bar{X} = (1/n) \sum_{i=1}^n X_i$  and hence also the statistic  $\sqrt{n}\bar{X}$  is sufficient for the family of the joint distributions of  $(X_1, \dots, X_n)$ , we may argue as follows (using a well-known result to be found, e.g., in Witting, 1985, Theorem 3.30): The distribution of  $\sqrt{n}\bar{X}$  differs from

that of an individual  $X_i$  only by its expected value which is of course given by  $\tilde{\theta} = \sqrt{n}\theta$ . In terms of  $\tilde{\theta}$ , the testing problem put forward above reads

$$\tilde{H} : \tilde{\theta} \leq -\tilde{\varepsilon} \text{ or } \tilde{\theta} \geq \tilde{\varepsilon} \quad \text{versus} \quad \tilde{K} : -\tilde{\varepsilon} < \tilde{\theta} < \tilde{\varepsilon}, \quad (4.2)$$

on the understanding that we choose  $\tilde{\varepsilon} = \sqrt{n}\varepsilon$ . Now, if there is an UMP test, say  $\tilde{\psi}$  at level  $\alpha$  for (4.2) based on a single random variable  $Z$  with

$$Z \sim \mathcal{N}(\tilde{\theta}, 1), \quad (4.3)$$

then a UMP level- $\alpha$  test for the original problem (4.1) is given by  $\phi(x_1, \dots, x_n) = \tilde{\psi}(n^{-1/2} \sum_{i=1}^n x_i)$ ,  $(x_1, \dots, x_n) \in \mathbb{R}^n$ .

By (4.3), the possible distributions of  $Z$  form a one-parameter exponential family in  $\tilde{\theta}$  and  $Z$  (cf. Bickel and Doksum, 2001, § 1.6.1), and by Lemma A.1.2 [→ Appendix, p. 369] this is a specific example of a STP<sub>3</sub> family. Moreover, the density functions involved obviously exhibit the continuity properties required in the assumptions of Theorem A.1.5 [→ Appendix, p. 371], and in view of the symmetry of the standard Gaussian distribution  $\mathcal{N}(0, 1)$ , the distribution of  $Z$  under  $\tilde{\theta} = \tilde{\varepsilon}$  is clearly the same as that of  $-Z$  under  $\tilde{\theta} = -\tilde{\varepsilon}$ . Thus, applying Lemma A.1.6 [→ p. 372] we can conclude that a UMP level- $\alpha$  test for (4.2) is given by

$$\tilde{\psi}(z) = \begin{cases} 1 & \text{for } |z| < C_{\alpha; \tilde{\varepsilon}} \\ 0 & \text{for } |z| \geq C_{\alpha; \tilde{\varepsilon}} \end{cases}, \quad (4.4)$$

where  $C_{\alpha; \tilde{\varepsilon}}$  the unique solution of

$$\Phi(C - \tilde{\varepsilon}) - \Phi(-C - \tilde{\varepsilon}) = \alpha, \quad C > 0. \quad (4.5)$$

The expression on the left-hand side of this equation equals the probability of the event  $\{ |Z_{\tilde{\varepsilon}}| \leq C \}$  with  $Z_{\tilde{\varepsilon}} \sim \mathcal{N}(\tilde{\varepsilon}, 1)$ . But  $Z_{\tilde{\varepsilon}} \sim \mathcal{N}(\tilde{\varepsilon}, 1)$  implies that  $Z_{\tilde{\varepsilon}}^2$  satisfies the definition of a variable whose distribution is noncentral  $\chi^2$  with a single degree of freedom ( $df$ ) and noncentrality parameter  $\tilde{\varepsilon}^2$  (cf. Johnson et al., 1995, § 29.1). Hence, the solution to equation (4.5) determining the critical constant of a UMP level- $\alpha$  test, admits the explicit representation

$$C_{\alpha; \tilde{\varepsilon}} = \sqrt{\chi_{1; \alpha}^2(\tilde{\varepsilon}^2)}, \quad (4.6)$$

where  $\chi_{1; \alpha}^2(\tilde{\varepsilon}^2)$  denotes the  $\alpha$ -quantile of a  $\chi^2$ -distribution with  $df = 1$  and noncentrality parameter  $\tilde{\varepsilon}^2$ . Finally, the desired optimal test for the problem (4.1) we started from, is obtained by applying the decision rule

Reject  $H : \theta \leq -\varepsilon$  or  $\theta \geq \varepsilon$  in favor of  $K : -\varepsilon < \theta < \varepsilon$

$$\text{if and only if it can be verified that } |\bar{X}| < n^{-1/2} C_{\alpha; \sqrt{n}\varepsilon}. \quad (4.7)$$

In (4.7), the critical constant has to be determined simply by plugging in  $\tilde{\varepsilon} = \sqrt{n}\varepsilon$  into (4.5) or (4.6), and constructing a program which computes  $C_{\alpha; \tilde{\varepsilon}}$ .

for any  $\tilde{\varepsilon} > 0$  is extremely easy whenever a programming language is used in terms of which the standard normal cdf  $\Phi(\cdot)$  is predefined either directly or, as in Fortran, via the so-called error integral  $\text{erf}(z) \equiv (2/\sqrt{\pi}) \int_0^z e^{-y^2} dy$ . In fact, as a function of  $C$ ,  $\Phi(C - \tilde{\varepsilon}) - \Phi(-C - \tilde{\varepsilon})$  is continuous and strictly increasing. Hence, as soon as  $\Phi(\cdot)$  can be treated as explicitly given, no more than an elementary interval halving algorithm is needed for computing  $C_{\alpha; \tilde{\varepsilon}}$  by means of (4.5) to any desired degree of numerical accuracy. Of course, the most convenient way of implementing the optimal test for equivalence of  $\mathcal{N}(\theta, 1)$  to  $\mathcal{N}(0, 1)$  on the basis of a single sample  $(X_1, \dots, X_n)$  from  $\mathcal{N}(\theta, 1)$  is to use a software package which provides the inverse noncentral  $\chi^2$  cdf with  $df = 1$  as a predefined function (in SAS, the name of this function is `cinv`; in R, one has to call the function `qchisq`). For the conventional level  $\alpha = .05$  of significance, Table 4.1 gives a tabulation of  $C_{\alpha; \tilde{\varepsilon}}$  on the grid  $\tilde{\varepsilon} = .1(1)4.9$ .

Table 4.1 *Optimal critical constant  $C_{\alpha; \tilde{\varepsilon}}$  [ $\rightarrow$  (4.5), (4.6)] for equivalence testing at the 5% level in the one-sample case with normally distributed data of known variance.*

$\tilde{\varepsilon}$	0.	1.	2.	3.	4.
.0	—	0.10332	0.42519	1.35521	2.35515
.1	0.06302	0.11470	0.49889	1.45517	2.45515
.2	0.06397	0.12859	0.58088	1.55516	2.55515
.3	0.06559	0.14553	0.66944	1.65515	2.65515
.4	0.06793	0.16624	0.76268	1.75515	2.75515
.5	0.07105	0.19157	0.85893	1.85515	2.85515
.6	0.07507	0.22255	0.95696	1.95515	2.95515
.7	0.08011	0.26032	1.05598	2.05515	3.05515
.8	0.08634	0.30608	1.15552	2.15515	3.15515
.9	0.09398	0.36085	1.25530	2.25515	3.25515

Figure 4.1 shows the density functions of the test statistic  $\sqrt{n}\bar{X}$  at both boundaries of the hypotheses and the critical interval of the optimal test at level  $\alpha = .05$  for the case that the sample consists of  $n = 100$  data points and the constant  $\varepsilon$  defining the width of the equivalence range for the expectation  $\theta$  of the underlying distribution has been chosen to be  $\varepsilon = .25$ . As becomes obvious from the graph, the area under both curves over the critical interval  $(-C_{\alpha; \sqrt{n}\varepsilon}, C_{\alpha; \sqrt{n}\varepsilon}) = (-.85893, .85893)$  is exactly equal to  $.05 = \alpha$ . According to Theorem A.1.5, this property characterizes the optimal test.

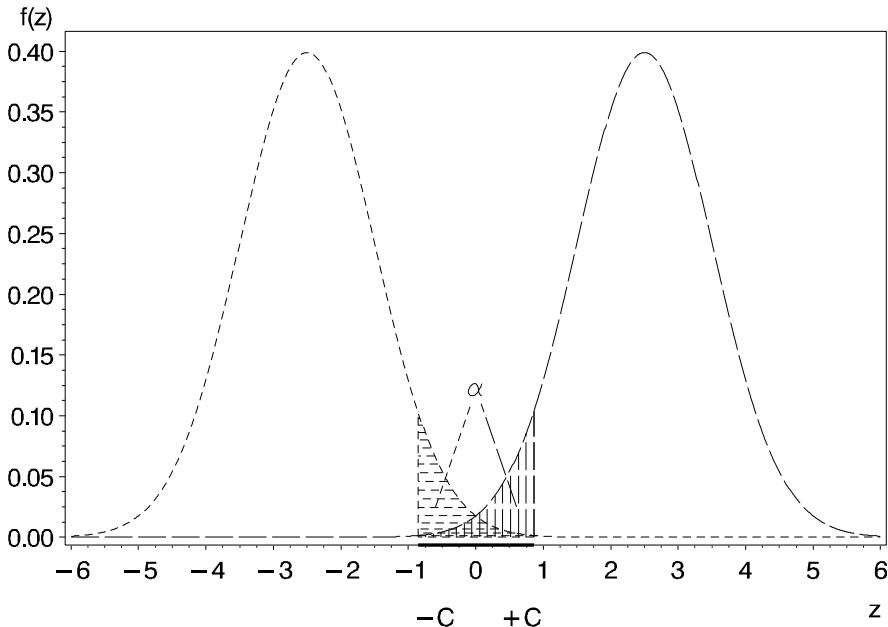


Figure 4.1 *Density of the distributions  $\mathcal{N}(-\sqrt{n}\varepsilon, 1)$  [left] and  $\mathcal{N}(\sqrt{n}\varepsilon, 1)$  [right] and critical interval [horizontal bar centered about 0] for  $Z = \sqrt{n}\bar{X}$  in the UMP test at level  $\alpha = .05$ , with  $\varepsilon = .25$  and  $n = 100$ .*

Figure 4.2 shows, for the same values of  $n$ ,  $\varepsilon$  and  $\alpha$ , the power of the UMP test for (4.1) as a function of the expectation  $\theta$  of the individual observations  $X_1, \dots, X_n$  [recall (4.1a)]. In accordance with the characterization of the form of the power functions of optimal tests for equivalence in symmetric cases given in general terms in Corollary A.1.7 [ $\rightarrow$  pp. 372–3], the curve is strictly unimodal and symmetric about zero. From a practical point of view, perhaps the most striking conclusion to be drawn from this picture is the following: In contrast to tests for traditional one- or two-sided problems, *equivalence tests do not admit the possibility of increasing the power to values arbitrarily close to 100% simply by selecting sufficiently extreme points in the parameter subspace corresponding to the alternative hypothesis under consideration.* The implications of this fact for sample size requirements to be satisfied in equivalence trials are obvious and far-reaching enough.

If we denote for arbitrary values of  $n$ ,  $\varepsilon$  and  $\alpha$  by  $\beta_{n,\varepsilon; \alpha}^*$  the ordinate of the peak of the power curve associated to the test (4.7), there holds the simple explicit formula

$$\beta_{n,\varepsilon; \alpha}^* = 2 \Phi(C_\alpha; \sqrt{n}\varepsilon) - 1 . \quad (4.8)$$

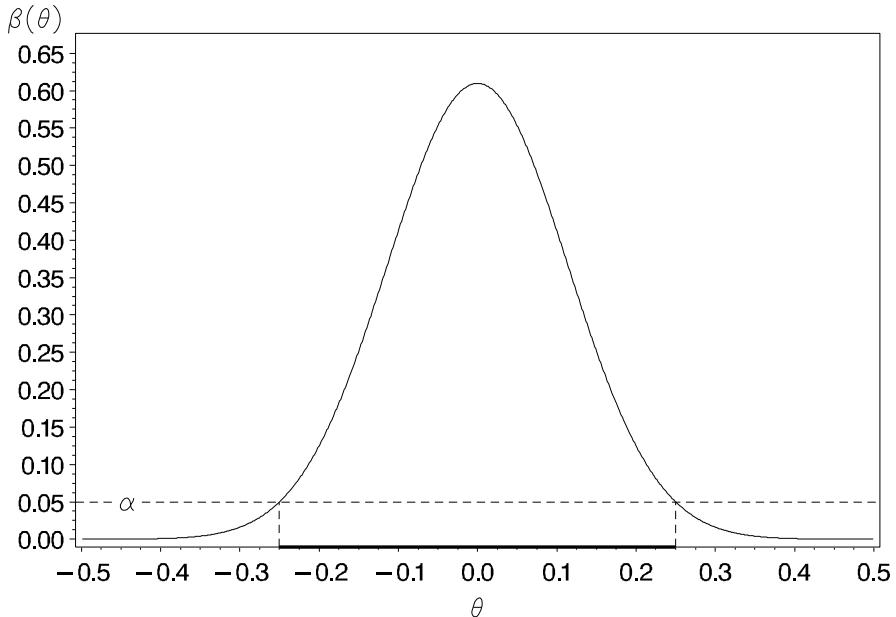


Figure 4.2 *Power function of the UMP test at level  $\alpha = 5\%$  for  $|\theta| \geq .25$  vs.  $|\theta| < .25$  based on  $n = 100$  observations from  $\mathcal{N}(\theta, 1)$ . [Bold-drawn bar on horizontal coordinate axis  $\leftrightarrow$  equivalence interval  $(-\varepsilon, \varepsilon) = (-.25, .25)$  specified by the alternative hypothesis.]*

This implies in particular that the maximum power attainable in any test for the equivalence problem (4.1) likewise depends on  $n$  and  $\varepsilon$  only through  $\sqrt{n}\varepsilon = \tilde{\varepsilon}$ . Table 4.2 shows, for the same selection of values of  $\tilde{\varepsilon}$  covered by the preceding table and again to 5 decimal places, the result of evaluating (4.8). For  $\tilde{\varepsilon} = \sqrt{100} \cdot .25 = 2.50$ , we read the number .60962 which is the precise value of the maximum ordinate of the curve plotted in Figure 4.2. Conversely, Table 4.2 can be used for obtaining a rough estimate of the sample size required for an equivalence trial to be analyzed in a confirmatory manner by means of (4.7). E.g., the smallest  $\tilde{\varepsilon}$  guaranteeing a maximum power of 80% is seen to lie between 2.9 and 3.0. Thus, for a trial in which we aim at establishing equivalence in the sense of  $-.25 < \theta < .25$ , we need a sample of size  $(2.9/.25)^2 = 134.56 < n < 144.00 = (3.0/.25)^2$  in order to detect coincidence of  $\theta$  with zero with a power of 80%, and so on.

Table 4.2 *Maximum power of the UMP test at level  $\alpha = .05$  for the one-sample equivalence problem with Gaussian data of unit variance, for the same grid of values of  $\tilde{\varepsilon} = \sqrt{n}\varepsilon$  as covered by Table 4.1.*

$\tilde{\varepsilon}$	0.	1.	2.	3.	4.
.0	—	.08229	.32930	.82465	.98148
.1	.05025	.09132	.38214	.85438	.98592
.2	.05101	.10232	.43868	.88009	.98939
.3	.05230	.11571	.49679	.90211	.99207
.4	.05416	.13203	.55434	.92077	.99413
.5	.05665	.15192	.60962	.93642	.99570
.6	.05984	.17611	.66141	.94943	.99687
.7	.06385	.20538	.70902	.96014	.99775
.8	.06880	.24046	.75212	.96885	.99840
.9	.07488	.28179	.79063	.97588	.99887

To be sure, as a testing problem per se, the one-sample problem with normally distributed observations of known variance is of limited practical importance since in real applications precise knowledge of the true value of  $\sigma^2$  will rarely be available. However, the testing procedure (4.7) which has been shown above to give a UMP solution to this problem, merits due consideration by the fact that it forms the basis of the construction of asymptotically valid tests for equivalence described in general terms in § 3.4. Recalling the form of the rejection region of such an asymptotic test [→ (3.13)], it is readily checked that the latter is obtained from the results of the present section by formally treating the test statistic  $T_N$  as a single observation from a normal distribution with variance  $\hat{\tau}_N^2 = \hat{\sigma}^2/N$  [→ p. 46 (iii)] and expected value  $\theta = \theta(F_1, \dots, F_k)$ .

With respect to the existing literature, it is worth noticing that in a paper going back as far as 40 years, Bondy (1969) suggested an appropriately modified version of (4.7) as a solution to the problem of testing for equivalence of two Gaussian distributions with known common variance. Bondy's contribution has been almost completely ignored by later authors in the field of equivalence testing (for one of the very few exceptions see Chester, 1986). In particular, although having appeared 3 years earlier than the seminal paper on the confidence interval inclusion approach (Westlake, 1972), Bondy's idea of having recourse to the classical theory of hypotheses testing has never been referenced by any one of the pioneers of the statistical methodology of bioequivalence assessment. One of the reasons why his paper has been left largely unnoticed in the pertinent literature of the years to come might have been that Bondy gave an unnecessarily complicated representation of the critical constant of his test making it appear hardly suited for applications in routine

work. Especially, any hint is missing that the optimal critical constant  $C_{\alpha; \tilde{\varepsilon}} [ \rightarrow (4.6) ]$  can be computed by means of the quantile function of a noncentral  $\chi^2$ -distribution with  $df = 1$ .

---

## 4.2 Test for equivalence of a hazard rate to some given reference value with exponentially distributed survival times

Let  $X$  denote a random variable whose density function (with respect to Lebesgue measure on  $\mathbb{R}$ ) is given by

$$f_X(x; \sigma) = (1/\sigma) \exp\{-x/\sigma\}, \quad x > 0. \quad (4.9)$$

As is well known (cf. Johnson et al., 1994, § 19.4), under this model the scale parameter  $\sigma > 0$  equals the expected value of  $X$ . In the sequel, we mostly use the short-hand notation  $X \sim \mathcal{E}(\sigma)$  in order to indicate that  $X$  is a random variable having density (4.9).

In the present section, it is our purpose to construct an optimal test of  $H: \sigma \leq \sigma_1$  or  $\sigma \geq \sigma_2$  versus  $K: \sigma_1 < \sigma < \sigma_2$  ( $0 < \sigma_1 < \sigma_2 < \infty$ ) based on a sample  $(X_1, \dots, X_n)$  of mutually independent observations from  $\mathcal{E}(\sigma)$ . The interval  $(\sigma_1, \sigma_2)$  specified by that alternative hypothesis  $K$  is supposed to contain a point  $\sigma_0$  serving as the target or reference value of the only parameter appearing in the model under consideration. As is the case in the one-sample problem with normally distributed data of known variance [recall the introductory paragraph of §4.1], it entails no loss of generality to assume that the interval corresponding to  $K$  has a normalized form which this time means that we have  $\sigma_1 = 1/(1 + \varepsilon)$ ,  $\sigma_2 = 1 + \varepsilon$  for some  $\varepsilon > 0$ . In order to induce such a normalization with respect to the form of the equivalence hypothesis, we have simply to apply the following transformation of the individual observations and the parameter, respectively:

$$X'_i = X_i / \sqrt{\sigma_1 \sigma_2}, \quad i = 1, \dots, n; \quad \sigma' = \sigma / \sqrt{\sigma_1 \sigma_2}. \quad (4.10)$$

This follows from the obvious fact that for  $X_i \sim \mathcal{E}(\sigma)$ , the distribution of  $X'_i$  is likewise exponential but with scale parameter  $\sigma'$  instead of  $\sigma$ . Moreover, with  $\varepsilon = \sqrt{\sigma_2/\sigma_1} - 1$ , the condition  $\sigma_1 < \sigma < \sigma_2$  is of course logically equivalent to  $1/(1 + \varepsilon) < \sigma' < (1 + \varepsilon)$ . In view of these relationships, we suppose in analogy to (4.1) that the hypotheses making up our testing problem have been proposed in the form

$$H : \sigma \leq 1/(1 + \varepsilon) \text{ or } \sigma \geq 1 + \varepsilon \text{ versus } K : 1/(1 + \varepsilon) < \sigma < 1 + \varepsilon \quad (4.11a)$$

from the start (with fixed  $\varepsilon > 0$ ), and the data set to be analyzed consists of  $n$  mutually independent observations  $X_1, \dots, X_n$  such that

$$X_i \sim \mathcal{E}(\sigma), \quad i = 1, \dots, n. \quad (4.11b)$$

Now, the distributional assumption (4.11b) is well known to imply that a statistic sufficient for the class of joint distributions of the sample  $(X_1, \dots, X_n)$  is given by  $T = \sum_{i=1}^n X_i$ . The sufficient statistic  $T$  can be shown (see, e.g., Feller, 1971, p. 11) to follow a gamma distribution with shape parameter  $n$  and scale parameter  $\sigma$ , given by the density function

$$f_{\sigma,n}^\Gamma(t) = (1/\Gamma(n))\sigma^{-n}t^{n-1} \exp\{-t/\sigma\}, \quad t > 0. \quad (4.12)$$

Since  $n$  denotes a known constant whereas  $\sigma$  is allowed to vary over the whole positive half-axis, (4.12) is an element of a one-parameter exponential family in  $t$  and  $\theta = -1/\sigma$ . According to Lemma A.1.2 [→ p. 369], each such family is STP<sub>∞</sub> and thus a fortiori STP<sub>3</sub>. Furthermore, it is readily verified that the function  $(t, \sigma) \mapsto f_{\sigma,n}^\Gamma(t)$  is continuous in both of its arguments. In view of Theorem A.1.5 [→ p. 371] and the sufficiency of  $T$ , we are justified to infer from these facts that the decision rule

$$\text{Reject } H \quad \text{if and only if} \quad C_{\alpha; n, \varepsilon}^1 < \sum_{i=1}^n X_i < C_{\alpha; n, \varepsilon}^2 \quad (4.13)$$

yields an UMP level- $\alpha$  test for (4.11a), provided the critical bounds  $C_{\alpha; n, \varepsilon}^1$ ,  $C_{\alpha; n, \varepsilon}^2$  are determined by solving the system

$$\begin{aligned} F_{1/(1+\varepsilon), n}^\Gamma(C_2) - F_{1/(1+\varepsilon), n}^\Gamma(C_1) &= \alpha \\ &= F_{1+\varepsilon, n}^\Gamma(C_2) - F_{1+\varepsilon, n}^\Gamma(C_1), \quad 0 < C_1 < C_2 < \infty, \end{aligned} \quad (4.14)$$

with  $F_{\sigma, n}^\Gamma(\cdot)$  denoting (for any  $\sigma > 0$  and  $n \in \mathbb{N}$ ) the cdf corresponding to the density function (4.12).

Unfortunately, notwithstanding the possibility of symmetrizing the hypotheses in the way having lead to (4.11a), the one-sample setting with exponentially distributed data does not satisfy condition (3.10) for symmetrization of the optimal critical interval. In other words, the system (4.14) does not admit reduction to a single equation involving the right-hand critical bound only. Consequently, implementation of (4.13) requires use of the iteration algorithm described in general terms on pp. 42-3. In order to adapt this computational scheme to the specific setting under discussion, using  $C_1^0 = n$  as a starting value of the left-hand critical bound  $C_1$  to  $T$  is a sensible choice because  $T/n$  is a consistent estimator of  $\sigma$  and the interval of values of  $\sigma$  specified by  $K$  contains unity. Generally, if a UMP test for equivalence is based on a statistic which consistently estimates the parameter to be assessed, the critical region will be a proper subset of the interval corresponding to the alternative hypothesis sharing basic formal properties with the latter. Hence, we can take

it for granted that the solution  $(C_1, C_2)$  to (4.14) gives, after rescaling by the factor  $1/n$ , an interval covering the point  $\sigma = 1$ . Thus, we have  $C_1 < n < C_2$  so that the initial value  $C_1^\circ = n$  is known to lie to the right of the left-hand boundary  $C_1$  of the critical interval to be determined. In accordance with the relationship  $C_1 < n$ , the program supplied in the **WKTSEQ2 Source Code Package** under the name **exp1st** for computing the solution to (4.14), runs through a sequence of preliminary steps diminishing  $C_1^\circ$  by  $\tau = .05 n$  each until it occurs that we have  $F_{1+\varepsilon,n}^\Gamma(C_2^\circ) - F_{1+\varepsilon,n}^\Gamma(C_1^\circ) \leq \alpha$ , with  $C_2^\circ$  computed as  $C_2^\circ = \left( F_{1/(1+\varepsilon),n}^\Gamma \right)^{-1} [\alpha + F_{1/(1+\varepsilon),n}^\Gamma(C_1^\circ)]$ . The point where this condition is met, is chosen as the value of  $\tilde{C}_1$  in the sense of part (iv) of the general description given on pp. 42-3, and  $\tilde{C}_1 + \tau$  as the corresponding upper bound  $\tilde{\tilde{C}}_1$  to  $C_1$ . Finally, successive interval halving is applied to  $(\tilde{C}_1, \tilde{\tilde{C}}_1)$  in order to approximate the exact value of  $C_1$  to the desired degree of accuracy, again as explained on p. 43, (iv).

In addition to the optimal critical constants  $C_{\alpha;n,\varepsilon}^1, C_{\alpha;n,\varepsilon}^2$  [to resume the explicit notation introduced in (4.13)], the program named **exp1st** also computes the power of the test established above against the specific alternative  $\sigma = 1$ . To be sure, this is not the point in the parameter space where the power function  $\beta(\cdot)$ , say, of the UMP level- $\alpha$  test for the equivalence problem made precise in (4.11) takes on its exact maximum. But by the results to be found in § 4.4 we know that the difference between  $\beta(1)$  and  $\max_{\sigma>0} \beta(\sigma)$  is asymptotically negligible (as  $n \rightarrow \infty$ ). Table 4.3 gives the critical constants  $C_{\alpha;n,\varepsilon}^1, C_{\alpha;n,\varepsilon}^2$  for the level  $\alpha = .05$ ,  $\varepsilon \in \{.30, .50\}$ , and sample sizes  $\leq 120$  being a multiple of 10. The corresponding values of the power attained at  $\sigma = 1$  are displayed in Table 4.4.

### Example 4.1

In a retrospective study involving  $n = 80$  patients, it was one of the aims to rule out the possibility that a particular drug used in the long-term treatment of progressive chronic polyarthritis (PCP) changes the mean life span of erythrocytes to a relevant degree. For a reference population, the expected length of the average erythrocytal life cycle was known to be  $3.84 \approx 4$  months. Let the distribution of the intraindividual mean erythrocytal life span be of the form (4.9) and choose the equivalence interval for  $\sigma$  as centered about 4, with length being equal to 2. In the sample recruited for the study a total time of  $\sum_{i=1}^n X_i = 329.7873$  was obtained. Rescaling the originally measured time values as well as the equivalence limits according to (4.10) yields  $T = 329.7873/\sqrt{15} = 85.1507$  and  $\varepsilon = \sqrt{5/3} - 1 = 0.2910$ . Rounding  $\varepsilon$  up to .30 we obtain from Table 4.3 (73.22150, 85.80868) as the critical interval to be used in (4.13). Since the observed value of  $T$  happens to be an inner point of this interval, we can decide in favor of equivalence between study and reference population with respect to mean erythrocytal life span.

Table 4.3 *Critical interval ( $C_{.05; n, \varepsilon}^1, C_{.05; n, \varepsilon}^2$ ) of the UMP equivalence test at the 5% level for the one-sample setting with exponentially distributed data, for  $\varepsilon = .30, .50$  and  $n = 10(10)120$ .*

$n$	$\varepsilon = .30$	$\varepsilon = .50$
10	(9.610923, 10.167879)	(9.302916, 10.179687)
20	(19.227987, 20.331723)	(18.175346, 20.878033)
30	(28.728536, 30.622505)	(26.300336, 32.650207)
40	(38.059577, 41.104285)	(33.955275, 45.339863)
50	(47.173320, 51.843905)	(41.447117, 58.454257)
60	(56.048597, 62.887533)	(48.855774, 71.779509)
70	(64.711420, 74.228599)	(56.204318, 85.244644)
80	(73.221500, 85.808680)	(63.505485, 98.817061)
90	(81.635963, 97.555833)	(70.767971, 112.476582)
100	(89.991186, 109.414666)	(77.998090, 126.208916)
110	(98.306216, 121.350753)	(85.200605, 140.003340)
120	(106.590707, 133.344176)	(92.379217, 153.851535)

Table 4.4 *Power attained at  $\sigma = 1$  when using the critical intervals shown in Table 4.3.*

$n$	$\varepsilon = .30$	$\varepsilon = .50$
10	.070340	.111904
20	.098785	.240904
30	.138156	.442380
40	.191426	.635809
50	.260318	.773372
60	.342885	.862271
70	.432381	.917672
80	.520229	.951433
90	.600305	.971657
100	.670069	.983607
110	.729339	.990588
120	.778984	.994629

**Remark on testing for dispersion equivalence of a single sample of normally distributed data.** Suppose that the distribution which the random sample  $(X_1, \dots, X_n)$  has been drawn from is Gaussian with expectation  $\mu$  and variance  $\sigma^2$  (with both parameters being unknown) and that our main interest is in establishing dispersion equivalence of this distribution to some reference distribution of the same form whose variance is known to equal  $\sigma_o^2$ . For convenience, let us formulate the equivalence hypothesis in this case as  $K : \sigma_1^2 < \sigma^2 < \sigma_2^2$  with  $(\sigma_1^2, \sigma_2^2)$  as a suitably chosen fixed interval covering the reference value  $\sigma_o^2$ . Applying the same transformation as before [recall (4.10)] and redefining  $\varepsilon$  as an abbreviation for  $\sigma_2/\sigma_1 - 1$  rather than  $\sqrt{\sigma_2/\sigma_1} - 1$  leads to the problem of testing

$$H : \sigma^2 \leq 1/(1 + \varepsilon) \text{ or } \sigma^2 \geq 1 + \varepsilon \text{ versus } K : 1/(1 + \varepsilon) < \sigma^2 < 1 + \varepsilon \quad (4.15a)$$

with a sample  $(X_1, \dots, X_n)$  of mutually independent observations satisfying

$$X_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n. \quad (4.15b)$$

Now, it is easy to verify that the problem (4.15) remains invariant under arbitrary common translations  $X_i \mapsto X_i + c$  of all individual observations contained in the sample. Furthermore, it is well known (see Lehmann and Romano, 2005, p. 220) that any test remaining invariant under all transformations of that type, must depend on the data only through the sum of squares  $SQ_X = \sum_{i=1}^n (X_i - \bar{X})^2$  or, equivalently,  $T = (1/2) \cdot SQ_X$ . But under the assumption of (4.15b) the latter statistic has a distribution whose density is precisely as defined in (4.12) with  $\sigma$  and  $n$  replaced by  $\sigma^2$  and  $n' = (n-1)/2$ , respectively (cf. Stuart and Ord, 1994, p. 385). Hence, in the one-sample setting with normally distributed data, a UMP invariant test for dispersion equivalence in the sense of (4.15a) is obtained simply by comparing the observed value of  $T = (1/2) \cdot SQ_X$  with the critical bounds  $C_{\alpha; n', \varepsilon}^1, C_{\alpha; n', \varepsilon}^2$  to be computed in exactly the same way as in the case of exponentially distributed survival times. In particular, the critical interval  $(73.22150, 85.80868)$  computed in the above example for the sum of  $n' = 80$  observations from  $\mathcal{E}(\sigma)$  can also be used for testing for dispersion equivalence of  $\mathcal{N}(\mu, \sigma^2)$  to  $\mathcal{N}(\mu, 1)$  in the sense of  $\sqrt{3/5} < \sigma^2 < \sqrt{5/3}$  with a sample of size  $n = 2n' + 1 = 161$  and  $(1/2) \cdot \sum_{i=1}^n (X_i - \bar{X})^2$  as the test statistic. (Note that the program `exp1st` works also for arbitrary noninteger values of the constant  $n$ ).

### 4.3 Testing for equivalence of a single binomial proportion to a fixed reference success probability

Both specific equivalence testing problems treated in the first two sections of this chapter refer to distributions of the absolutely continuous type. In contrast, the present section deals with the most extreme case of a one-sample

setting with data following a discrete distribution. More precisely speaking we now assume that each of the  $n$  mutually independent observations  $X_1, \dots, X_n$  takes on only two different values, say 0 ( $\leftrightarrow$  “no response”) and 1 ( $\leftrightarrow$  “response”) with probability  $p$  and  $1 - p$ , respectively.

Once more, the (unweighted) sum of all individual observations turns out to be sufficient for the class of the possible distributions of the vector  $(X_1, \dots, X_n)$ . Of course, the distribution of  $T = \sum_{i=1}^n X_i$  is now given by the probability mass function

$$b(t; n, p) = \binom{n}{t} p^t (1-p)^{n-t} , \quad t \in \{0, 1, \dots, n\}. \quad (4.16)$$

Likewise, it is easily seen that the corresponding family  $(b(\cdot; n, p))_{0 < p < 1}$  of densities with respect to the counting measure of the set  $\{0, 1, \dots, n\}$  of possible values of  $T$ , is an exponential family in  $t$  and  $\theta = \log(p/(1-p))$ . In view of these facts we can apply essentially the same arguments as were used in § 4.2 showing that for

$$\begin{aligned} H : 0 < p \leq p_1 \text{ or } p_2 \leq p < 1 \text{ versus } K : p_1 < p < p_2 \\ (0 < p_1 < p_2 < 1, \text{ fixed}) \end{aligned} \quad (4.17)$$

there exists an UMP level- $\alpha$  test defined by the following decision rule:

$$\left\{ \begin{array}{l} \text{Rejection of } H \text{ for } C_\alpha^1(n; p_1, p_2) < T < C_\alpha^2(n; p_1, p_2) \\ \text{Rejection with prob. } \gamma_\alpha^1(n; p_1, p_2) \text{ for } T = C_\alpha^1(n; p_1, p_2) \\ \text{Rejection with prob. } \gamma_\alpha^2(n; p_1, p_2) \text{ for } T = C_\alpha^2(n; p_1, p_2) \\ \text{Acceptance for } T < C_\alpha^1(n; p_1, p_2) \text{ or } T > C_\alpha^2(n; p_1, p_2) \end{array} \right. . \quad (4.18)$$

The optimality property of (4.18) holds on the understanding that the constants  $C_\alpha^\nu(n; p_1, p_2)$ ,  $\gamma_\alpha^\nu(n; p_1, p_2)$ ,  $\nu = 1, 2$ , are determined by solving

$$\begin{aligned} \sum_{t=C_1+1}^{C_2-1} b(t; n, p_1) + \sum_{\nu=1}^2 \gamma_\nu b(C_\nu; n, p_1) = \alpha = \sum_{t=C_1+1}^{C_2-1} b(t; n, p_2) + \\ \sum_{\nu=1}^2 \gamma_\nu b(C_\nu; n, p_2) , \quad 0 \leq C_1 \leq C_2 \leq n, \quad 0 \leq \gamma_1, \gamma_2 < 1. \end{aligned} \quad (4.19)$$

The computer program to be found in the **WKTSHEQ2 Source Code Package** under the name **bi1st** implements the search algorithm described in full generality on p. 43 for this specific case. In addition, the program computes the power of the test (4.18) against the special alternative that the true value of the response probability  $p$  coincides with the midpoint  $(p_1 + p_2)/2$  of the interval corresponding to  $K$ . In view of the undesirability of using external chance mechanisms in real applications of statistical inference, the program also calculates the power against  $p = (p_1 + p_2)/2$  of the nonrandomized version of the test as obtained by adding both boundary points of the rejection interval to the acceptance region.

The sufficient condition (3.10) for symmetrizing the optimal critical interval is satisfied here whenever we are willing to choose  $p_1 = 1/2 - \varepsilon$ ,  $p_2 = 1/2 + \varepsilon$  which seems especially natural in those applications of the test where the  $X_i$  are in fact indicators of the sign of intra-subject differences between paired continuous observations [for details see § 5.1]. Trivially, the distribution of  $\sum_{i=1}^n X_i - n/2$  under  $p = 1/2 + \varepsilon$  is the same as that of  $n/2 - \sum_{i=1}^n X_i$  under  $p = 1/2 - \varepsilon$ . Since, as a test statistic,  $\sum_{i=1}^n X_i - n/2$  is equivalent to  $T = \sum_{i=1}^n X_i$  and, for  $p_1 = 1/2 - \varepsilon$ ,  $p_2 = 1/2 + \varepsilon$ , (4.17) can equivalently be written  $H : |\theta| \geq \varepsilon$  vs.  $K : |\theta| < \varepsilon$  with  $\theta = p - 1/2$ , we may conclude from (3.12) that in the symmetric case, the optimal critical interval for  $T$  is given by

$$\left( C_\alpha^1(n; 1/2 - \varepsilon, 1/2 + \varepsilon), C_\alpha^2(n; 1/2 - \varepsilon, 1/2 + \varepsilon) \right) = \left( n - C_\alpha^*(n; \varepsilon), C_\alpha^*(n; \varepsilon) \right) \quad (4.20a)$$

where

$$C_\alpha^*(n; \varepsilon) = n/2 + \max \left\{ c \left| \sum_{n/2-c+1}^{n/2+c-1} b(t; n, 1/2 + \varepsilon) \leq \alpha, \right. \right. \\ \left. \left. 0 < c \leq n/2 + 1, (c + n/2) \in \mathbb{Z} \right\} . \quad (4.20b)$$

Furthermore, specializing (3.12) to the case  $T(\mathbf{X}) = \sum_{i=1}^n X_i - n/2$ , it is easy to check that the nonrandomized version of the test (4.18) can also be based on a p-value in the usual sense. For that purpose, one has to define the significance probability of the observed value  $t \in \{0, 1, \dots, n\}$  of  $T$  by

$$p_{obs}(t) = \sum_{j=n-\tilde{t}}^{\tilde{t}} b(j; n, 1/2 + \varepsilon) \quad (4.21a)$$

with

$$\tilde{t} = \max\{t, n - t\} . \quad (4.21b)$$

Table 4.5 shows the critical interval  $(C_{.05}^1(n; p_1, p_2), C_{.05}^2(n; p_1, p_2))$  to be used in (4.18) at the 5% level for various choices of the equivalence interval  $(p_1, p_2)$  and a selection of sample sizes ranging from 25 to 125. The power of the corresponding tests (with and without randomization) against the alternative  $p = (p_1 + p_2)/2$  is given in Table 4.6. It should be noticed that these tables cover, for the same values of  $n$ , also the cases  $(p_1, p_2) \in \{(0.10, 0.30), (0.20, 0.40), (0.30, 0.50), (0.05, 0.35)\}$ . Generally, it is not hard to establish the relationships

$$C_\alpha^\nu(n; 1 - p_2, 1 - p_1) = n - C_\alpha^{3-\nu}(n; p_1, p_2), \quad \nu = 1, 2 , \quad (4.22a)$$

$$\beta_{n,\alpha}^{1-p_2,1-p_1}(1-p) = \beta_{n,\alpha}^{p_1,p_2}(p) , \quad \tilde{\beta}_{n,\alpha}^{1-p_2,1-p_1}(1-p) = \tilde{\beta}_{n,\alpha}^{p_1,p_2}(p) , \quad (4.22b)$$

of which the latter refer to the power function of the UMP test (4.18) [ $\rightarrow \beta_{n,\alpha}^{q_1,q_2}(\cdot)$ ] and its nonrandomized version [ $\rightarrow \tilde{\beta}_{n,\alpha}^{q_1,q_2}(\cdot)$ ], respectively.

Table 4.5 *Critical interval ( $C_{.05}^1(n; p_1, p_2), C_{.05}^2(n; p_1, p_2)$ ) for the test (4.18) at the 5% level, for various specifications of  $(p_1, p_2)$  and  $n = 25(25)125$ .*

$(p_1, p_2) =$						
$n$	(.40, .60)	(.50, .70)	(.60, .80)	(.70, .90)	(.35, .65)	(.65, .95)
25	(12, 13)	(15, 16)	(17, 18)	(20, 21)	(12, 13)	(20, 22)
50	(24, 26)	(29, 31)	(35, 36)	(40, 42)	(23, 27)	(38, 45)
75	(36, 39)	(44, 47)	(51, 54)	(59, 63)	(33, 42)	(55, 68)
100	(48, 52)	(58, 63)	(68, 73)	(77, 85)	(43, 57)	(73, 91)
125	(59, 66)	(72, 79)	(84, 93)	(96, 107)	(53, 72)	(90, 115)

Table 4.6 *Power of the UMP tests with critical intervals shown in Table 4.5 against the alternative  $p = (p_1 + p_2)/2$  [italicized values  $\leftrightarrow$  nonrandomized version of the tests].*

$(p_1, p_2) =$						
$n$	(.40, .60)	(.50, .70)	(.60, .80)	(.70, .90)	(.35, .65)	(.65, .95)
25	.0816 . <i>0000</i>	.0847 . <i>0000</i>	.0900 . <i>0000</i>	.1093 . <i>0000</i>	.1551 . <i>0000</i>	.3167 . <i>1867</i>
50	.1367 . <i>1123</i>	.1428 . <i>1146</i>	.1663 . <i>0000</i>	.2451 . <i>1364</i>	.4124 . <i>3282</i>	.7193 . <i>6626</i>
75	.2201 . <i>1824</i>	.2347 . <i>1854</i>	.2918 . <i>1987</i>	.4419 . <i>3288</i>	.6815 . <i>6443</i>	.8940 . <i>8899</i>
100	.3368 . <i>2356</i>	.3602 . <i>3157</i>	.4429 . <i>3367</i>	.6319 . <i>6104</i>	.8392 . <i>8067</i>	.9583 . <i>9418</i>
125	.4685 . <i>4083</i>	.4980 . <i>4149</i>	.5924 . <i>5590</i>	.7611 . <i>7166</i>	.9231 . <i>8930</i>	.9844 . <i>9801</i>

Figure 4.3 shows the complete power function of the UMP test for  $K: .70 < p < .90$  at level  $\alpha = .05$  based on a sample of size  $n = 82$ , along with its counterpart referring to the nonrandomized version of (4.18). An obvious feature evident in both curves is the lack of symmetry about the midpoint of the

equivalence interval specified by the alternative hypothesis. Furthermore, a comparison of both graphs discloses that even in applications with rather large samples the gain in power attained by exhausting the significance level through randomized decisions on the boundaries of the critical interval, is by no means negligible.

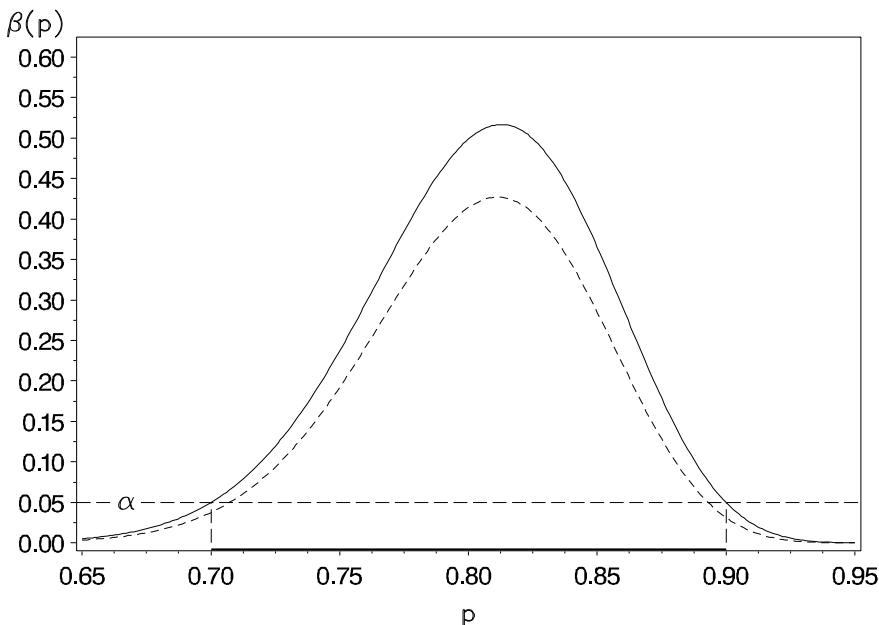


Figure 4.3 Power function of the UMP level- $\alpha$  test of  $p \leq .70$  or  $p \geq .90$  vs.  $.70 < p < .90$  [solid line] and its nonrandomized counterpart [broken line] for  $n = 82$  and  $\alpha = .05$ .

#### Example 4.2

In an observational study of an improved combination regime for the adjuvant chemotherapy of a particular tumor entity, the investigators aimed at showing that a highly toxic drug which had been used in the past as one of the components could be replaced with a much better tolerable substance without inducing a relevant change in patients' prognosis. The sample comprised  $n = 273$  patients suffering from the respective tumor. Each patient had been under follow-up for at least 24 months. From extensive clinical experience the traditionally administered combination containing the toxic ingredient was known to provide a 2-year progression-free survival rate of 73%. The interval

of acceptable deviations from this reference value was specified to be  $(p_1, p_2) = (.65, .75)$ . In the current study a number of  $t = 191$  patients survived the first 24 months since initiation of treatment without exhibiting signs of a progression. The limits of the critical interval of the test (4.18) at level  $\alpha = .05$  are computed (by means of the program mentioned above) to be  $C_{.05}^1(273; .65, .75) = 189$ ,  $C_{.05}^2(273; .65, .75) = 194$  leading to a rejection of the null hypothesis of nonequivalence. Power calculation shows the probability of detecting exact coincidence of the parameter  $p$  with the reference value .73 in the nonrandomized test to be .1216. When randomized decisions are admitted for  $T = 189$  and  $T = 194$ , this value is slightly increased to .1435. If the true value of  $p$  coincides with the midpoint of the equivalence interval instead, the rejection probability amounts to .2082 (nonrandomized version of the test) and .2431 (UMP test at exact level .05), respectively.

---

#### 4.4 Confidence-interval inclusion rules as asymptotically UMP tests for equivalence

At first glance the principle of confidence interval inclusion and the classical approach to the construction of equivalence tests seem to have nothing in common except for yielding procedures which are valid in the sense of maintaining the significance level. Nevertheless, it can be shown that an interval inclusion test based on a pair of confidence bounds at one-sided level  $1 - \alpha$  [recall § 3.1] approximates, under suitable conditions, the optimal test for the same problem with arbitrarily high precision if the sample size is chosen sufficiently large. A rigorous mathematical proof of this result is technically rather complicated and can be found in Wellek (1994, pp. 96-103) or Romano (2005). However, in the one-sample setting with normally distributed observations of unit variance and symmetrically chosen equivalence interval for the parameter  $\theta$  [ $\rightarrow$  (4.1)], it admits a simple heuristic derivation which will be presented on the following pages.

If the values realized in a sample of size  $n$  from  $\mathcal{N}(\theta, 1)$  are  $x_1, \dots, x_n$  with  $\bar{x}_n$  as their ordinary arithmetic mean, the standard method of interval estimation yields the number  $\bar{x}_n - u_{1-\alpha}/\sqrt{n}$  and  $\bar{x}_n + u_{1-\alpha}/\sqrt{n}$  as a lower and upper  $(1 - \alpha)$ -confidence bound for  $\theta$ , respectively. Hence, rejection of the null hypothesis  $|\theta| \geq \varepsilon$  in the interval inclusion test requires in this case that both of the inequalities  $\bar{x}_n - u_{1-\alpha}/\sqrt{n} > -\varepsilon$  and  $\bar{x}_n + u_{1-\alpha}/\sqrt{n} < \varepsilon$  must be satisfied. Clearly, this holds true if and only if  $\sqrt{n}\bar{x}_n$  is a point in the open interval with limits  $\mp(\sqrt{n}\varepsilon - u_{1-\alpha})$ . In other words, the interval inclusion test rejects if and only if we have  $\sqrt{n}|\bar{x}_n| < \sqrt{n}\varepsilon - u_{1-\alpha}$ . Accordingly, its rejection region is exactly of the same form as that of the optimal level- $\alpha$  test [ $\rightarrow$  (4.7)], with the sole difference that the critical constant  $C_{\alpha; \sqrt{n}\varepsilon}$  as defined

by (4.6) is replaced by  $\sqrt{n}\varepsilon - u_{1-\alpha}$ .

Trivially, the equation determining  $C_{\alpha; \sqrt{n}\varepsilon}$  according to (4.5) can be rewritten as

$$\Phi(c - \sqrt{n}\varepsilon) + \Phi(c + \sqrt{n}\varepsilon) = 1 + \alpha , \quad 0 < c < \infty . \quad (4.23)$$

Since  $\varepsilon$  stands for a positive number, one has of course  $\sqrt{n}\varepsilon \rightarrow +\infty$  (as  $n \rightarrow \infty$ ) and consequently  $c - \sqrt{n}\varepsilon \rightarrow -\infty$ ,  $c + \sqrt{n}\varepsilon \rightarrow +\infty$ . In view of  $\Phi(-\infty) = 0$ ,  $\Phi(+\infty) = 1$  this implies  $\Phi(c - \sqrt{n}\varepsilon) \rightarrow 0$ ,  $\Phi(c + \sqrt{n}\varepsilon) \rightarrow 1$ . Furthermore, like every reasonable testing procedure, the UMP level- $\alpha$  test for (4.1b) is consistent (Wellek, 1994, Lemma A.3.3). But convergence of the rejection probability to 1 under any  $\theta \in (-\varepsilon, \varepsilon)$  implies [recall (4.5)] that the area under the density curve of a Gaussian variable with unit variance over the critical interval  $(-C_{\alpha; \sqrt{n}\varepsilon}, C_{\alpha; \sqrt{n}\varepsilon})$  tends to 1, which can only be true if the length of that interval and hence  $C_{\alpha; \sqrt{n}\varepsilon}$  itself increases to  $\infty$  as  $n \rightarrow \infty$ . In view of this, the values of  $c$  to be taken into consideration when solving equation (4.23) are very large positive numbers if  $n$  is large, and for fixed  $c \gg 0$ ,  $\Phi(c + \sqrt{n}\varepsilon)$  will come very close to its limit, i.e., to 1, already for values of  $n$  which are still too small for carrying  $\Phi(c - \sqrt{n}\varepsilon)$  into a small neighborhood of zero. According to this intuitive notion the solution to (4.23), i.e., the optimal critical constant  $C_{\alpha; \sqrt{n}\varepsilon}$  satisfies the approximate equation

$$\Phi(C_{\alpha; \sqrt{n}\varepsilon} - \sqrt{n}\varepsilon) \approx \alpha . \quad (4.24)$$

On the other hand, in view of the definition of  $u_\gamma$  and the symmetry of the standard normal distribution we clearly have  $\alpha = \Phi(-u_{1-\alpha})$ . Hence, (4.24) holds if and only if

$$\Phi(C_{\alpha; \sqrt{n}\varepsilon} - \sqrt{n}\varepsilon) \approx \Phi(-u_{1-\alpha}) . \quad (4.25)$$

Since the standard normal cdf has a continuous inverse we are justified to drop  $\Phi(\cdot)$  on both sides of (4.25) and write for “sufficiently large” sample sizes  $n$

$$C_{\alpha; \sqrt{n}\varepsilon} \approx \sqrt{n}\varepsilon - u_{1-\alpha} . \quad (4.26)$$

Finally, each of the two tests under comparison is uniquely determined by one of the critical bounds  $C_{\alpha; \sqrt{n}\varepsilon}$  and  $\sqrt{n}\varepsilon - u_{1-\alpha}$  being approximately equal according to (4.26). In view of this, we have given a heuristic proof of the assertion that approximate identity holds between both tests with respect to all probabilistic properties, in particular the power against any specific alternative  $\theta \in (-\varepsilon, \varepsilon)$ . In fact, according to the result proved as Theorem A.3.5 in Wellek (1994), still more can be said: *Asymptotically*, the interval inclusion test is uniformly most powerful for the equivalence problem (4.1) even if tests are admitted which maintain the nominal level only in the limit and/or the arbitrarily selected alternative  $\theta \in K$  is allowed to converge to one of the boundary points of the equivalence interval rather than kept fixed for all values of  $n$ .

To be sure, demonstration of asymptotic optimality of a test per se is mainly of theoretical interest. The actual relevance of such a result can only be assessed if detailed numerical investigations have been performed providing in particular precise information on the order of magnitude of the sample sizes required to ensure that the loss of power entailed by replacing the optimal finite-sample test with the asymptotic procedure under consideration be practically negligible. Since both the UMP level- $\alpha$  test and the test based on the principle of confidence interval inclusion depends on  $n$  and  $\varepsilon$  only through  $\tilde{\varepsilon} = \sqrt{n}\varepsilon$ , it suffices to compare the procedures for various values of this latter quantity.

Tables 4.7 and 4.8 are the direct counterparts of Tables 4.1 and 4.2 for the interval inclusion test at the same level  $\alpha = .05$ . Inspecting the entries in Table 4.8 as compared to the corresponding values shown in Table 4.2 leads to the following major conclusions:

- (i) As long as  $\tilde{\varepsilon} = \sqrt{n}\varepsilon$  fails to exceed the value 1.7 the interval inclusion test is very unsatisfactory since its maximum power falls below the significance level. In contrast, for  $\tilde{\varepsilon} = 1.7$  the power of the UMP test against  $\theta = 0$  amounts to 21% and is thus more than four times larger than  $\alpha$ .
- (ii) The cases where the interval inclusion test does better than the trivial procedure rejecting  $H$  with constant probability  $\alpha$  (independently of the data), but is still markedly inferior to the UMP test at the same level are characterized by the condition  $1.7 < \tilde{\varepsilon} \leq 2.3$ .
- (iii) If  $\tilde{\varepsilon}$  is increased beyond the right-hand limit of this interval (by weakening the equivalence condition under the hypothesis and/or increasing the sample size), the difference between both tests disappears very quickly (except of course for some remainder being negligible for all practical purposes).

Numerical power comparisons for various other one-sample problems whose results are not reported in detail here, fully confirm these findings. Summarily speaking, the gain from applying the UMP instead of the corresponding interval inclusion test is very modest whenever the power attained in the former has the order of magnitude usually required in a study to be analyzed by means of a conventional one- or two-sided test. However, in equivalence testing power values exceeding 50% can only be obtained if either the equivalence range specified by the alternative hypothesis is chosen extremely wide or the sample size requirements are beyond the scope of feasibility for most, if not all, applications.

Table 4.7 Critical constant  $\tilde{\varepsilon} - u_{.95}$  for the interval inclusion test at level  $\alpha = .05$  in the one-sample setting with data from  $\mathcal{N}(\theta, 1)$ .

$\tilde{\varepsilon}$	0.	1.	2.	3.	4.
.0	—	-.64485	0.35515	1.35515	2.35515
.1	-1.5449	-.54485	0.45515	1.45515	2.45515
.2	-1.4449	-.44485	0.55515	1.55515	2.55515
.3	-1.3449	-.34485	0.65515	1.65515	2.65515
.4	-1.2449	-.24485	0.75515	1.75515	2.75515
.5	-1.1449	-.14485	0.85515	1.85515	2.85515
.6	-1.0449	-.04485	0.95515	1.95515	2.95515
.7	-.94485	0.05515	1.05515	2.05515	3.05515
.8	-.84485	0.15515	1.15515	2.15515	3.15515
.9	-.74485	0.25515	1.25515	2.25515	3.25515

Table 4.8 Power of the test with critical region  $\{\sqrt{n}|\bar{X}_n| < \tilde{\varepsilon} - u_{.95}\}$  against  $\theta = 0$  for the same grid of values of  $\tilde{\varepsilon} = \sqrt{n}\varepsilon$  as covered by Table 4.7.

$\tilde{\varepsilon}$	0.	1.	2.	3.	4.
.0	—	.00000	.27752	.82463	.98148
.1	.00000	.00000	.35100	.85437	.98592
.2	.00000	.00000	.42121	.88009	.98939
.3	.00000	.00000	.48763	.90211	.99207
.4	.00000	.00000	.54984	.92077	.99413
.5	.00000	.00000	.60753	.93642	.99570
.6	.00000	.00000	.66050	.94943	.99687
.7	.00000	.04398	.70864	.96014	.99775
.8	.00000	.12329	.75197	.96885	.99840
.9	.00000	.20139	.79057	.97588	.99887

Clearly, the intuitive notion of equivalence suggests to tolerate only small deviations of the parameter  $\theta$  from its target value  $\theta_0$ . If we adopt this point of view there is no choice but to either refrain from launching a suitable study at all, or to be content with power values falling distinctly below 50%. The latter case is the one where replacing an interval inclusion procedure with an optimal test typically yields the most marked gains in power, which can amount to up to 30% in these instances. In addition to such considerations concerning power and efficiency, there is another major reason (which has been alluded to already in § 3.3) why there is a real need for carrying out classical optimal

constructions of tests for equivalence in spite of the option of solving testing problems of that kind by means of the interval inclusion principle: The range of settings for which well-established interval estimation procedures exist, is by far too narrow to provide a clinical or experimental investigator with a sufficiently rich arsenal of specific equivalence testing methods.

---

## 4.5 Noninferiority analogues of the tests derived in this chapter

For all three concrete one-sample settings considered in this chapter, derivation of a noninferiority test is fairly straightforward.

In the case of a single sample from  $\mathcal{N}(\theta, \sigma_o^2)$  with known variance  $\sigma_o^2 > 0$ , the UMP level- $\alpha$  test of  $H_1 : \theta \leq -\varepsilon$  versus  $K_1 : \theta > -\varepsilon$  rejects if it turns out that  $\bar{X} > \sigma_o u_{1-\alpha}/\sqrt{n} - \varepsilon$ . The power of this test against an arbitrarily chosen alternative  $\theta_a > -\varepsilon$  can be computed by evaluating the expression  $1 - \Phi(u_{1-\alpha} - \sqrt{n}(\varepsilon + \theta_a)/\sigma_o)$ . Thus, for fixed significance level  $\alpha$ , the power of the test for noninferiority of a normal distribution with unknown expected value  $\theta$  to a Gaussian distribution of the same known variance  $\sigma_o^2$  being centered at zero, depends on  $n$ ,  $\varepsilon$ ,  $\theta_a$  and  $\sigma_o$  only through  $\tilde{\varepsilon} = \sqrt{n}(\varepsilon + \theta_a)/\sigma_o$ , and it is tempting to look at the noninferiority analogue of Table 4.2 which is shown below as Table 4.9. Comparing the entries in these two tables we can concretely assess the price which has to be paid for establishing with the same degree of certainty in terms of the type-I error risk a hypothesis which is considerably more precise than that stated in the noninferiority problem as the alternative: The power attainable with the same

Table 4.9 *Analogue of Table 4.2 for the noninferiority case.*

$\tilde{\varepsilon}$	0.	1.	2.	3.	4.
.0	—	.25951	.63876	.91231	.99074
.1	.06119	.29293	.67550	.92719	.99296
.2	.07425	.32821	.71060	.94004	.99469
.3	.08934	.36510	.74381	.95105	.99604
.4	.10659	.40328	.77492	.96038	.99707
.5	.12613	.44241	.80376	.96821	.99785
.6	.14805	.48211	.83025	.97472	.99844
.7	.17237	.52199	.85432	.98007	.99888
.8	.19910	.56165	.87598	.98442	.99920
.9	.22818	.60069	.89529	.98794	.99943

sample size in the two-sided equivalence testing scenario is substantially lower unless the power of the test for noninferiority is close to unity.

Let us suppose next that the individual observations  $X_i$  making up the sample under analysis follow a standard exponential distribution  $\mathcal{E}(\sigma)$  with unknown scale parameter  $\sigma > 0$  and the problem of testing  $H_1 : \sigma \leq 1/(1+\varepsilon)$  versus  $H_1 : \sigma > 1/(1+\varepsilon)$  has been proposed. Then, an UMP test at level  $\alpha$  is carried out through deciding in favor of noninferiority for values of  $\sum_{i=1}^n X_i$  found to be larger than the upper  $100\alpha$  percentage point  $\gamma_\alpha(n, \varepsilon)$ , say, of the gamma distribution with parameters  $n$  and  $1/(1+\varepsilon)$ . In SAS, the critical constant  $\gamma_\alpha(n, \varepsilon)$  is readily obtained by calling the intrinsic function `gaminv` with arguments  $1 - \alpha$  and  $n$  (in that order) and dividing the result by  $1 + \varepsilon$ . In the R system, the function `qgamma` is available for the same purpose. Alternatively, the test can be carried out by computing the p-value associated with the realized value  $t$  of  $T = \sum_{i=1}^n X_i$ . In SAS, this just requires to write down the assignment statement `p=1-probgam((1+eps)*t,n)`. Revisiting Example 4.1 where we had  $n = 80$ ,  $\varepsilon = .3$  and  $t = 85.1507$ , this yields  $p = .00094$  so that noninferiority can be rejected *a fortiori*.

In the binomial case, the decision rule of the UMP level- $\alpha$  test of  $H_1 : p \leq p_o - \varepsilon$  versus  $K_1 : p > p_o - \varepsilon$  is obtained by simplifying (4.18) to

$$\begin{cases} \text{Rejection of } H_1 \text{ for } T > C_\alpha(n; p_o, \varepsilon) \\ \text{Rejection with prob. } \gamma_\alpha(n; p_o, \varepsilon) \text{ for } T = C_\alpha(n; p_o, \varepsilon) , \\ \text{Acceptance for } T < C_\alpha(n; p_o, \varepsilon) \end{cases} \quad (4.27)$$

where

$$C_\alpha(n; p_o, \varepsilon) = \max \left\{ k \in \mathbb{N}_0 \mid \sum_{j=k}^n b(j; n, p_o - \varepsilon) > \alpha \right\}, \quad (4.28)$$

$$\gamma_\alpha(n; p_o, \varepsilon) = \left( \alpha - \sum_{j=C_\alpha(n; p_o, \varepsilon)+1}^n b(j; n, p_o - \varepsilon) \right) / b(C_\alpha(n; p_o, \varepsilon); n, p_o - \varepsilon). \quad (4.29)$$

The power of the exact randomized UMP test (4.27) against any alternative  $p > p_o - \varepsilon$  is given by

$$\beta_{\alpha,n}(p_o, \varepsilon; p) = \sum_{j=C_\alpha(n; p_o, \varepsilon)+1}^n b(j; n, p) + \gamma_\alpha(n; p_o, \varepsilon) b(C_\alpha(n; p_o, \varepsilon); n, p), \quad (4.30)$$

which reduces to

$$\beta_{\alpha,n}^{(0)}(p_o, \varepsilon; p) = \sum_{j=C_\alpha(n; p_o, \varepsilon)+1}^n b(j; n, p) \quad (4.31)$$

when randomization is omitted. The corresponding functions of  $p$  are plotted in Figure 4.4 for the same specifications as underly Figure 4.3, except of course for setting the right-hand boundary of the hypothetical equivalence range equal to unity. For  $\alpha = .05$ ,  $n = 82$ ,  $p_o = .80$  and  $\varepsilon = .10$  the critical constants are computed to be  $C_\alpha(n; p_o, \varepsilon) = 64$  and  $\gamma_\alpha(n; p_o, \varepsilon) = .35671$  using either the above formulae (4.28) and (4.29), or running the program `bi1st` with  $p_2 = 1 - 10^{-15}$ .

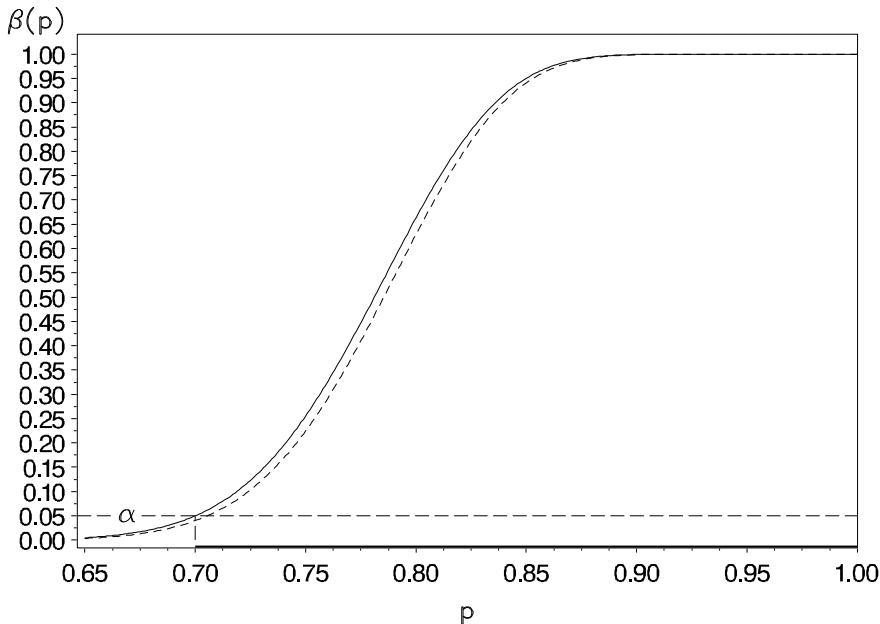


Figure 4.4 *Power function of the UMP level- $\alpha$  test of  $p \leq .70$  vs.  $p > .70$  [solid line] and its nonrandomized counterpart [broken line] for  $n = 82$  and  $\alpha = .05$ .*

# 5

---

## *Equivalence tests for designs with paired observations*

---

### 5.1 Sign test for equivalence

The assumptions which must be satisfied in order to ensure that the sign test for equivalence be an exactly valid procedure exhibiting some basic optimality property are as weak as in the case of its traditional one- and two-sided analogue. This means that the data have to be given by  $n$  mutually independent pairs  $(X_1, Y_1), \dots, (X_n, Y_n)$  of random variables following all the same bivariate distribution which may be of arbitrary form. In particular, the distribution of the  $(X_i, Y_i)$  is allowed to be discrete. Within each such random pair, the first and second component gives the result of applying the two treatments, say  $A$  and  $B$ , under comparison to the respective observational unit. Except for treatment, the observations  $X_i$  and  $Y_i$  are taken under completely homogeneous conditions so that each intraindividual difference  $D_i \equiv X_i - Y_i$  can be interpreted as an observed treatment effect. For definiteness, we assume further that large values of the  $X$ 's and  $Y$ 's are favorable, which implies that  $A$  has been more efficient than  $B$  in the  $i$ th individual if and only if the value taken on by  $D_i$  turns out to be positive.

Now, let us agree on defining equivalence of both treatments by means of the condition that, except for practically irrelevant differences, the probability  $p_+ \equiv P[D_i > 0]$  of observing  $A$  to turn out superior to  $B$  in a randomly selected individual, coincides with the probability  $p_- \equiv P[D_i < 0]$  of observing  $A$  to yield a weaker effect than  $B$ . In view of our aim to construct an optimal testing procedure, it proves advisable that we relate this condition to the logarithmic scale which leads to considering the statement  $-\varepsilon_1 < \log p_+ - \log p_- < \varepsilon_2$  (with sufficiently small positive real numbers  $\varepsilon_1, \varepsilon_2$ ) as the equivalence hypothesis of interest.

In fact, the mathematically natural choice of a target parameter for the inferential problem in mind is given by

$$\theta = \log(p_+/p_-) . \quad (5.1)$$

This is readily seen by examining the joint distribution of the usual counting statistics  $N_+ = \#\{i | D_i > 0\}$ ,  $N_0 = \#\{i | D_i = 0\}$ ,  $N_- = \#\{i | D_i < 0\}$

whose probability mass function admits the representation

$$f(n_+, n_0, n_-; n, \theta, \zeta) = \frac{n!}{n_+! n_0! n_-!} \exp\{n_+ \theta + n_0 \zeta\} (1 + e^\theta + e^\zeta)^{-n}, \quad (5.2)$$

where

$$\theta = \log(p_+/p_-), \quad \zeta = \log(p_0/p_-). \quad (5.3)$$

Furthermore, the argument made explicit in Lehmann and Romano (2005, pp. 137-8) shows that a test which is uniformly most powerful unbiased (UMPU) in the class of those level- $\alpha$  tests of

$$H : \theta \leq -\varepsilon_1 \vee \theta \geq \varepsilon_2 \text{ versus } K : -\varepsilon_1 < \theta < \varepsilon_2 \quad (5.4)$$

which depend on the data only through  $(N_+, N_0, N_-)$ , uniformly maximizes the power also within the class of *all* unbiased tests at the same level for (5.4).

By (5.2), the class of all possible joint distributions of  $(N_+, N_0, N_-)$ , i.e., of all trinomial distributions based on a sample of size  $n$ , is a two-parameter exponential family in  $(\theta, \zeta)$  and  $(N_+, N_0)$ . Hence, we can apply Theorem A.2.2 [→ Appendix, p. 375] with  $k = 1$ ,  $T = N_+$ ,  $S = N_0$  which implies that a UMPU level- $\alpha$  test for the problem (5.4) exists and is of the following form: Given the observed number of ties between  $A$  and  $B$ , i.e., of zeroes contained in the sample  $(D_1, \dots, D_n)$ , is  $n_0 \in \{0, 1, \dots, n\}$ , the null hypothesis  $H$  of nonequivalence has to be rejected if the number  $N_+$  of positive signs satisfies the double inequality  $C_{\alpha|n_0}^1 + 1 \leq N_+ \leq C_{\alpha|n_0}^2 - 1$ . On the boundaries of the corresponding critical interval, i.e., for  $N_+ = C_{\alpha|n_0}^\nu$ , a randomized decision in favor of  $K$  has to be taken with probability  $\gamma_{\alpha|n_0}^\nu$  ( $\nu = 1, 2$ ). The complete set  $C_{\alpha|n_0}^\nu$ ,  $\gamma_{\alpha|n_0}^\nu$ ,  $\nu = 1, 2$ , of (conditional) critical constants has to be determined by solving the equations

$$\sum_{n_+=C_{\alpha|n_0}^1+1}^{C_{\alpha|n_0}^2-1} P_{-\varepsilon_1}[N_+ = n_+ | N_0 = n_0] + \sum_{\nu=1}^2 \gamma_{\alpha|n_0}^\nu P_{-\varepsilon_1}[N_+ = C_{\alpha|n_0}^\nu | N_0 = n_0] = \alpha, \quad (5.5a)$$

$$\sum_{n_+=C_{\alpha|n_0}^1+1}^{C_{\alpha|n_0}^2-1} P_{\varepsilon_2}[N_+ = n_+ | N_0 = n_0] + \sum_{\nu=1}^2 \gamma_{\alpha|n_0}^\nu P_{\varepsilon_2}[N_+ = C_{\alpha|n_0}^\nu | N_0 = n_0] = \alpha. \quad (5.5b)$$

Since the conditional distribution of  $N_+$  given  $N_0 = n_0$  is well known (cf. Johnson et al., 1997, p. 35) to be binomial with parameters  $n - n_0$  and  $p_+/(p_+ + p_-)$

$= e^\theta / (1 + e^\theta)$  for each  $n_0 \in \{0, 1, \dots, n\}$  and  $\theta \in \mathbb{R}$ , the system (5.5) differs from (4.19) only by changes in notation. More precisely speaking, (5.5) turns into the system of equations to be solved in testing for equivalence of a single binomial proportion to  $1/2$  simply by making the substitution  $n \leftrightarrow n - n_0$ ,  $p_1 \leftrightarrow e^{-\varepsilon_1} / (1 + e^{-\varepsilon_1})$ ,  $p_2 \leftrightarrow e^{\varepsilon_2} / (1 + e^{\varepsilon_2})$ .

Thus, it eventually follows that an UMPU level- $\alpha$  test for the nonparametric equivalence problem proposed at the beginning of this section is obtained by treating the observed number  $n_0$  of pairs with tied components as a fixed quantity and carrying out the test derived in § 4.3 with  $n - n_0$  as nominal sample size and  $(e^{-\varepsilon_1} / (1 + e^{-\varepsilon_1}), e^{\varepsilon_2} / (1 + e^{\varepsilon_2}))$  as the equivalence range for the binomial parameter. Accordingly, the practical implementation of the decision rule to be used in the sign test for equivalence entails no new computational steps. In particular, in the symmetric case  $\varepsilon_1 = \varepsilon_2 = \varepsilon$ , the nonrandomized version of the test can be carried out by means of the p-value as defined in (4.21) (with  $\varepsilon$  and  $n$  replaced with  $e^\varepsilon / (1 + e^\varepsilon) - 1/2$  and  $n - n_0$ , respectively), and even the table of critical bounds to the test statistic presented before for the binomial one-sample problem can be used for some values of  $n - n_0$  and some specifications of  $\varepsilon$ . The only true extension brought into play when using the binomial one-sample test of § 4.3 as a conditional procedure for testing the hypotheses (5.4) concerns the computation of power values and minimum sample sizes required for guaranteeing specified lower bounds to the power. As is generally the case with conditional testing procedures, the power function of the sign test for equivalence is of practical interest only in its nonconditional form, which implies that the respective computations involve averaging with respect to the distribution of the conditioning statistic  $N_0$ . Obviously, irrespective of the form of the hypotheses under consideration, the nonconditional power function of any test based on the statistics  $N_+$  and  $N_0$  is a function of two arguments, viz.  $p_+$  and  $p_0$ . A program which enables one to compute the rejection probability of both the exact and the nonrandomized version of the UMPU level- $\alpha$  test for (5.4) against the specific alternative  $\theta = 0$  (or, equivalently,  $p_+ = p_-$  which in turn is equivalent to  $p_+ = (1 - p_0)/2$  for arbitrary rates  $p_0$  of zeroes among the intra-subject differences  $D_i$ ), can be found in the **WKSHEQ2 Source Code Package** under the program name **powsign** (again, both SAS, and R code is provided).

Clearly, the problem of inducing considerable conservatism by not allowing randomized decisions in favor of the alternative hypothesis which has been discussed from a general perspective in § 2.5 (for the noninferiority case) and § 3.3, affects the sign test for (two-sided) equivalence as well. The sign test is among the settings where mitigating this conservatism by increasing the nominal significance level as far as possible without rising the nonconditional rejection probability under any null parameter constellation above the target level, provides a very satisfactory compromise between maximization of power and suitability of the resulting testing procedure for real applications. For sample sizes up to 150 being multiples of 25 and cases where the equivalence limits to the target parameter  $\theta$  are chosen in accordance with Table 1.1

(note that in presence of ties,  $p_+/(1 - p_0) = p_+/(p_+ + p_-)$  is the analogue of  $p_+$  as to be considered in the continuous case), Table 5.1 shows maximally increased nominal levels  $\alpha^*$  for use in the nonrandomized sign test for equivalence at target level  $\alpha = .05$ .

Table 5.1 *Nominal significance levels and sizes of improved nonrandomized sign tests for equivalence at target level 5% for  $\varepsilon_1 = \varepsilon_2 = \varepsilon = \log(6/4), \log(7/3)$  and  $n = 25(25)150$ . [Number in (): size of the nonrandomized test at nominal level  $\alpha = .05$ ;  $p_0^*$  = value of the nuisance parameter maximizing the rejection probability of  $H$ .]*

$\varepsilon$	$n$	$\alpha^*$	$p_0^*$	Size	
.405465	25	.11714	.0675	.04746	(.00000)
"	50	.08429	.0000	.04046	(.04046)
"	75	.06699	.0113	.04832	(.04014)
"	100	.05802	.0671	.04413	(.03935)
"	125	.05750	.0296	.04743	(.04018)
"	150	.05709	.0408	.04899	(.04366)
.847298	25	.07060	.0000	.03825	(.03825)
"	50	.06580	.2214	.04831	(.04771)
"	75	.06000	.0206	.04928	(.04133)
"	100	.05305	.0494	.04545	(.04101)
"	125	.06000	.3580	.04983	(.04157)
"	150	.05963	.4851	.04888	(.04704)

### Example 5.1

In a study based on a sample of size  $n = 50$ , the number of patients doing better under treatment  $A$  as compared to  $B$  was observed to be  $n_+ = 17$ . In  $n_0 = 13$  patients, no difference between the responses to  $A$  and  $B$  could be detected. The equivalence interval for  $\theta = \log(p_+/p_-)$  was chosen symmetric about 0 specifying  $\varepsilon_1 = \varepsilon_2 = \varepsilon = \log(7/3) = .847298$  which is the same as requiring of the parameter of the conditional binomial distribution of  $N_+$  given  $N_0 = n_0$  to fall in the interval  $(3/10, 7/10)$ . With these data and specifications of constants, the sign test for equivalence has to be carried out as a binomial test for equivalence of  $p = e^\theta/(1 + e^\theta)$  to  $1/2$  in a sample of size  $n' = n - n_0 = 37$  and equivalence limits  $.5 \mp .2$  to  $p$ . At the 5% level, the rejection region of this test is computed (by means of the program **bi1st** referred to in § 4.3) to be the set  $\{16 < N_+ < 21\}$ . Since the observed value of the test statistic  $N_+$  is an element of this set, we are justified to reject nonequivalence in the sense of  $|\log(p_+/p_-)| \geq .847298$ .

For assessing the power of the test let us assume that the true proportion  $p_0$  of patients responding equally well to both treatments coincides with the observed relative frequency  $13/50 = .26$  of ties and that in the subpopulation of patients providing nonindifferent responses, superiority of  $A$  to  $B$  is exactly as frequent as inferiority. Running the enclosed program `powsign` we find the power against this specific alternative to be 61.53% (UMPU test) and 54.19% (nonrandomized version), respectively. Replacing the target significance level .05 with the raised nominal level  $\alpha^* = .0658$  as displayed in the pertinent line of the above table, the power of the nonrandomized test increases to 60.91%. Thus, under the present circumstances, the loss of efficiency entailed by discarding randomization can be compensated almost completely without changing the basic structure of the test.

### *Noninferiority version of the sign test*

The modifications to be performed in order to construct a UMPU test for noninferiority rather than two-sided equivalence based on the sign statistic, are straightforward. Conditional on the value  $n_0$  observed in the sample for the number of tied observations, the decision between  $H_1 : \theta \leq -\varepsilon$  and  $K_1 : \theta > -\varepsilon$  has to be taken according to the rule (4.27) substituting  $p_0 - \varepsilon$  and  $n$  with  $e^{-\varepsilon}/(1 + e^{-\varepsilon})$  and  $n - n_0$ , respectively. Conditional on  $\{N_0 = n_0\}$ , the power against an arbitrarily selected alternative  $(\theta, p_0)$  is obtained by making the analogous substitutions  $1/2 \rightarrow p_0$ ,  $1/2 - e^{-\varepsilon}/(1 + e^{-\varepsilon}) \rightarrow \varepsilon$ ,  $n - n_0 \rightarrow n$ ,  $e^\theta/(1 + e^\theta) \rightarrow p$  on the right-hand side of Equation (4.30). Again, averaging the result with respect to the distribution of  $N_0$  which is given by  $\mathcal{B}(n, p_0)$  yields the nonconditional power against that alternative.

The problem of constructing an improved nonrandomized version of the test can likewise be addressed in the same manner as in the case of two-sided equivalence, i.e., by determining largest admissible nominal levels  $\alpha^*$ . Keeping everything else fixed, except for replacing  $(-\varepsilon, \varepsilon)$  by  $(-\varepsilon, \infty)$  as the  $\theta$ -range specified by the hypothesis to be established, produces Table 5.2 as a noninferiority analogue of Table 5.1. In terms of  $\alpha^*$ , the results differ between both hypotheses formulations the more the smaller the sample size has been chosen. Given  $n$ , the more liberal of the two specifications of the equivalence margin is associated with smaller changes in the resulting value of the admissible nominal significance level when interest is in establishing noninferiority instead of equivalence in the two-sided sense.

Table 5.2 *Nominal significance levels and sizes of improved nonrandomized sign tests for noninferiority at target level 5% for  $\varepsilon = \log(6/4), \log(7/3)$  and  $n = 25(25)150$ . [Number in (): size of the nonrandomized test at nominal level  $\alpha = .05$ ;  $p_0^*$  = value of the nuisance parameter maximizing the rejection probability on  $H$ .]*

$\varepsilon$	$n$	$\alpha^*$	$p_0^*$	Size	
.405465	25	.07780	.2940	.04606	(.03423)
"	50	.05734	.1710	.04391	(.03931)
"	75	.06000	.0480	.04825	(.03981)
"	100	.05875	.0210	.04879	(.04195)
"	125	.05500	.0170	.04819	(.04205)
"	150	.05688	.0960	.04847	(.04278)
.847298	25	.07424	.1610	.04874	(.04409)
"	50	.06584	.2255	.04857	(.04753)
"	75	.06000	.0205	.04928	(.04114)
"	100	.05305	.0495	.04546	(.04101)
"	125	.06000	.3580	.04983	(.04157)
"	150	.05963	.4850	.04888	(.04672)

## 5.2 Equivalence tests for the McNemar setting

The test commonly named after McNemar (1947) deals with a setting which is just a special case of that underlying the previous section. Specialization results from the fact that all primary observations  $X_i$  and  $Y_i$  are now assumed to be binary response-status indicators taking on values 1 ( $\leftrightarrow$  success) and 0 ( $\leftrightarrow$  failure), respectively. (Of course, any other way of coding the two alternative elementary outcomes leads to the same inferences.) This obviously implies that the set  $(X_1, Y_1), \dots, (X_n, Y_n)$  of bivariate data points make up a sample of size  $n$  from a quadrinomial distribution with parameters  $p_{00}, \dots, p_{11}$ , say, given by  $p_{00} = P[X_i = 0, Y_i = 0]$ ,  $p_{01} = P[X_i = 0, Y_i = 1]$ ,  $p_{10} = P[X_i = 1, Y_i = 0]$ ,  $p_{11} = P[X_i = 1, Y_i = 1]$ . In view of the sufficiency of the respective frequencies for the corresponding family of joint distributions of all  $n$  random pairs, there is no loss of efficiency in reducing the full data set as directly observed, to the entries in a  $2 \times 2$  contingency table of the form displayed in Table 5.3.

Table 5.3 *Outline of a  $2 \times 2$  contingency table for use in the analysis of a trial yielding paired binary data.*

Treatment ment	Treatment B			
	A	0	1	$\Sigma$
0	$N_{00}$ ( $p_{00}$ )	$N_{01}$ ( $p_{01}$ )	$N_{0\cdot}$ ( $p_{0\cdot}$ )	
1	$N_{10}$ ( $p_{10}$ )	$N_{11}$ ( $p_{11}$ )	$N_{1\cdot}$ ( $p_{1\cdot}$ )	
$\Sigma$	$N_{\cdot 0}$ ( $p_{\cdot 0}$ )	$N_{\cdot 1}$ ( $p_{\cdot 1}$ )	$n$ (1.00)	

As is the case in the “classical” McNemar test concerning an ordinary one- or two-sided problem, in constructing a test for equivalence of the two treatments  $A$  and  $B$  based on a contingency table of that structure, the difference rather than the ratio of the probabilities  $p_{10}$ ,  $p_{01}$  of the two possible kinds of discordant pairs of primary observations will be regarded as the target parameter. Evidently, establishing equivalence or noninferiority with respect to  $p_{10}/p_{01}$  leads not to a new testing problem but can be done simply by applying the appropriate version of the test obtained in § 5.1 with  $N_+ = N_{10}$ ,  $N_- = N_{01}$  and the analogous replacements among the  $p$ ’s.

Although the parametrization through  $p_{10}/p_{01}$  is not only mathematically the most convenient one but also provides an intuitively reasonable measure of dissimilarity of the effects of the treatments to be compared in a matched-pair design using a binary outcome criterion, there is considerable practical interest in alternative hypotheses formulations. Conceptually, a closely related formulation is that used in the papers by Lachenbruch and Lynch (1998) and Nam and Blackwelder (2002). These authors propose to define equivalence of the treatments under study by the condition that the true value of the ratio  $p_{1\cdot}/p_{\cdot 1} \equiv (p_{10} + p_{11})/(p_{01} + p_{11})$  of both marginal binomial parameters must be sufficiently close to unity. Replacing  $p_{10}/p_{01}$  with  $p_{1\cdot}/p_{\cdot 1}$  makes the construction of exact tests technically very complicated, and the procedures which can be found in the literature referenced above are both asymptotic in nature.

Of course, exactly the same objection raised in § 1.6 against basing equivalence assessment of binomial distributions from which independent samples have been taken, on the difference of the two population parameters, likewise applies to the case of correlated proportions. Notwithstanding this fact, in McNemar’s setting, the still by far most popular choice of the target parameter for measuring the dissimilarity between the underlying theoretical distribution is  $\delta = p_{10} - p_{01}$ . Accordingly, the testing problems to be dealt

with in the better part of the present section are given by

$$H : \delta \leq -\delta_1 \text{ or } \delta \geq \delta_2 \text{ versus } K : -\delta_1 < \delta < \delta_2 \quad (5.6a)$$

and

$$H_1 : \delta \leq -\delta_0 \text{ versus } K_1 : \delta > -\delta_0 \quad (5.6b)$$

where

$$\delta = p_{10} - p_{01}, \quad \delta_k \in (0, 1) \quad \text{for } k \in \{0, 1, 2\}. \quad (5.7)$$

In the journal literature which has become available about tests for these problems (see Lu and Bean, 1995; Morikawa et al., 1996; Tango, 1998; Hsueh et al., 2001; Sidik, 2003; Nam, 2006), there is a preference in favor of using test statistics which are based on maximum likelihood estimators restricted by the assumption that  $\delta$  falls on the common boundary of the hypotheses. The first to propose the use of restricted maximum likelihood estimation in a comparable context (testing for one-sided equivalence of binomial distributions in a parallel-group design) were Farrington and Manning (1990). In contrast, the tests which will be presented in Subsections 5.2.1 and 5.2.2 are based on Wald type statistics computed from the likelihood maximized over the whole parameter space. In the derivation of these procedures, it will be convenient to introduce an extra symbol, say  $\eta$ , for denoting the probability of observing in a randomly selected experimental unit a constellation belonging to the off-diagonal cells of Table 5.3. In other words, we define

$$\eta = p_{10} + p_{01}. \quad (5.8)$$

### 5.2.1 Large-sample solution

A large-sample solution to the testing problem (5.6a) is obtained by applying the theory outlined in §3.4. In fact, standard results on the limiting distribution of arbitrary linear functions of the absolute frequencies appearing as entries in a contingency table with random marginal sums (see, e.g., Rao, 1973, p. 383, 6a.1(i)) imply that the statistic  $\sqrt{n}(\hat{\delta}_n - \delta)/\sqrt{\eta - \delta^2}$ , with  $\hat{\delta}_n = (N_{10} - N_{01})/n$ , is asymptotically a standard normal variate. Furthermore,  $\sqrt{\eta - \delta^2}$  is obviously a continuous function of the basic multinomial parameters  $p_{00}, \dots, p_{11}$  so that plugging in the natural estimators  $\hat{\eta}_n = (N_{10} + N_{01})/n$ ,  $\hat{\delta}_n = (N_{10} - N_{01})/n$  yields a consistent estimator of the standard error of the numerator of that statistic (as follows from another well-known result — see Rao, 1973, p. 345 — on the asymptotic properties of the multinomial family). Hence, on the understanding that both extreme constellations  $\eta = 0$  and  $|\delta| = 1$  under which the variance of  $\sqrt{n}(\hat{\delta}_n - \delta)$  vanishes can be ruled out, the testing problem (5.6a) satisfies all conditions of using the approach of §3.4 with the following specifications:  $N = n$ ,  $T_N = \hat{\delta}_n$ ,  $\theta = \delta$ , and  $\hat{\tau}_N = [(\hat{\eta}_n - \hat{\delta}_n^2)/n]^{1/2}$ . In terms of the off-diagonal frequencies  $N_{10}$  and  $N_{01}$ ,

the decision rule of the corresponding asymptotically valid test for the problem (5.6a) admits the following representation:

$$\text{Reject nonequivalence iff } \frac{n^{1/2}|(N_{10} - N_{01}) - n(\delta_2 - \delta_1)/2|}{[n(N_{10} + N_{01}) - (N_{10} - N_{01})^2]^{1/2}} < C_\alpha \left( \frac{n^{3/2}(\delta_1 + \delta_2)/2}{[n(N_{10} + N_{01}) - (N_{10} - N_{01})^2]^{1/2}} \right). \quad (5.9)$$

For definiteness, let us recall that the critical upper bound appearing in the above rule has to be determined as the square root of the lower  $100\alpha$  percentage point of a  $\chi^2$ -distribution with a single degree of freedom and a (random) noncentrality parameter which equals the square of the expression appearing in the argument to the functional symbol  $C_\alpha$ , i.e.,  $n^3(\delta_1 + \delta_2)^2/4[n(N_{10} + N_{01}) - (N_{10} - N_{01})^2]$ .

### Example 5.2

We illustrate the asymptotic testing procedure given by (5.9), by analyzing a contingency table compiled from data contained in the reference database of the CSE (Common Standards for Quantitative Electrocardiography) project launched in the early 1980s. One of the major objectives of this project was the establishment of tools for a comparative assessment of the diagnostic accuracy of various computer programs for the automatic classification of electrocardiograms (ECGs). For that purpose, a diagnostic “gold standard” was set up by classifying many hundred subjects with well-documented ECG recordings into clear-cut diagnostic categories by means of carefully validated clinical criteria irrespective of the ECG. In the version underlying the report of Willems et al. (1990) the database contained  $n = 72$  ECGs recorded from patients suffering from left ventricular hypertrophy (LVH), a pathological condition of the heart notoriously difficult to recognize by means of noninvasive methods. Table 5.4 summarizes the results of the assessment of these ECGs by a panel of human experts (*CR* [“Combined Referee”]) on the one hand, and one specific computer program denoted by *F* on the other. In this table, 0 stands for classification into a wrong category (different from LVH) whereas 1 indicates that the LVH was recognized from the patient’s ECG. For the general purposes of the CSE project, it was in particular of interest to assess the equivalence of the computer program to the human experts with respect to the capability of diagnosing LVH. Since the difference between the respective probabilities of detecting an existing hypertrophy of the left ventricle equals to  $\delta$  as defined in (5.7), this amounts to raising a testing problem of the form (5.6a). Finally, let us choose the equivalence interval once more in a symmetric manner setting  $\delta_1 = \delta_2 = .20$ , and use the conventional value  $\alpha = .05$  for the nominal significance level.

Table 5.4 *Comparison between the computer program F and the “Combined Referee” (CR) in the diagnostic classification of the ECGs of 72 patients with clinically proven hypertrophy of the left ventricle of the heart. [Data from Willems et al. (1990, p. 113).]*

		F		
C		0	1	$\Sigma$
R				
0		47	5	52
		(.6528)	(.0694)	(.7222)
1		4	16	20
		(.0556)	(.2222)	(.2778)
$\Sigma$		51	21	72
		(.7083)	(.2917)	(1.000)

With  $\delta_1 = \delta_2$  and the observed frequencies displayed in the above table, the expression on the left-hand side of the critical inequality (5.9) becomes

$$\begin{aligned} & \frac{n^{1/2}|(N_{10} - N_{01}) - n(\delta_2 - \delta_1)/2|}{[n(N_{10} + N_{01}) - (N_{10} - N_{01})^2]^{1/2}} \\ &= \frac{\sqrt{72}|4 - 5|}{\sqrt{72(4+5) - (4-5)^2}} = \sqrt{\frac{72}{647}} = \sqrt{.111283} = .3336. \end{aligned}$$

On the other hand, we get for the noncentrality parameter of the  $\chi^2$ -squared distribution providing an upper critical bound to our test statistic:

$$\begin{aligned} \hat{\psi}^2 &= \left[ \frac{n^{3/2}(\delta_1 + \delta_1)/2}{[n(N_{10} + N_{01}) - (N_{10} - N_{01})^2]^{1/2}} \right]^2 \\ &= \frac{72^3 \cdot .20^2}{72 \cdot 9 - (-1)^2} = 23.075611. \end{aligned}$$

But the .05-quantile of a  $\chi^2$ -distribution with a single degree of freedom and that noncentrality parameter is computed (e.g., by means of the SAS intrinsic function `cinv`) to be 9.978361. Consequently, as the critical upper bound we obtain

$$C_\alpha \left( \frac{n^{3/2}(\delta_1 + \delta_2)/2}{[n(N_{10} + N_{01}) - (N_{10} - N_{01})^2]^{1/2}} \right) = \sqrt{9.978361} = 3.1589.$$

Since the latter is much larger than the observed value of the test statistic  $n^{1/2}|(N_{10} - N_{01}) - n(\delta_2 - \delta_1)/2|/[n(N_{10} + N_{01}) - (N_{10} - N_{01})^2]^{1/2}$ , it follows that at the 5% level, the data of Table 5.4 allow a decision in favor of equivalence of *F* and *CR* in the sense of  $-.20 < \delta < .20$ .

### 5.2.2 Corrected finite-sample version of the large-sample test

In view of the purely asymptotic nature of the approach leading to (5.9), the validity of the procedure in finite samples can obviously not be taken for granted without further study. The clearest and most reliable answer to the question of possible discrepancies between actual and nominal level of an asymptotic testing procedure is obtained by determining the rejection probability under arbitrary null parameter constellations by means of exact computational methods. For that purpose, it is convenient to introduce the following transformation of variables and parameters, respectively:

$$U = N_{10}, \quad V = N_{10} + N_{01}; \quad \pi = p_{10}/\eta, \quad (5.10)$$

with  $\eta$  defined as in (5.8). In terms of  $(U, V)$ , the critical region of the test (5.9) can be rewritten

$$\left\{ (u, v) \in \mathbb{N}_0^2 \mid u \leq v \leq n, T_n^{\delta_1, \delta_2}(u, v) < c_{\alpha; n}^{\delta_1, \delta_2}(u, v) \right\} \quad (5.11)$$

where

$$T_n^{\delta_1, \delta_2}(u, v) = \frac{\sqrt{n}|(2u - v) - n(\delta_2 - \delta_1)/2|}{[nv - (2u - v)^2]^{1/2}},$$

$$c_{\alpha; n}^{\delta_1, \delta_2}(u, v) = C_\alpha \left( \frac{n^{3/2}(\delta_1 + \delta_2)/2}{[nv - (2u - v)^2]^{1/2}} \right).$$

The advantage of representing the critical region in this form stems from the fact that the conditional distribution of  $U$  given  $\{V = v\}$  is binomial with parameters  $(v, \pi)$  for any  $v = 0, 1, \dots, n$ , and the marginal distribution of  $V$  is likewise binomial, with parameters  $(n, \eta)$ . Hence, the rejection probability of the test (5.9) under any  $(\pi, \eta)$  admits the representation

$$Q_{\alpha, n}^{\delta_1, \delta_2}(\pi, \eta) = \sum_{v=0}^n \left\{ \sum_{u \in \mathcal{U}_{\alpha, n}^{\delta_1, \delta_2}(v)} b(u; v, \pi) \right\} b(v; n, \eta), \quad (5.12)$$

where  $\mathcal{U}_{\alpha, n}^{\delta_1, \delta_2}(v)$  denotes the  $v$ -section of the set (5.11) given by

$$\mathcal{U}_{\alpha, n}^{\delta_1, \delta_2}(v) = \left\{ u \in \mathbb{N}_0 \mid u \leq v, T_n^{\delta_1, \delta_2}(u, v) < c_{\alpha; n}^{\delta_1, \delta_2}(u, v) \right\}, \quad (5.13)$$

which is typically\* an (maybe empty) interval of nonnegative integer numbers  $\leq v$ . Furthermore, in (5.12), both symbols of the type  $b(k; m, p)$  stand for the respective binomial point masses.

---

\*By “typically,” we mean in the present context that in constructing the respective sets explicitly by direct computation, we never found an exception but no mathematical proof of the corresponding general assertion is available. In view of this, both computer programs implementing evaluations of (5.12) (`mcnemasc`, `mcnempow`), although working on the understanding that it is a generally valid fact, check upon it during each individual run. In case of finding an exception, either program would stop and produce an error message.

Now, in order to determine the actual size of the asymptotic McNemar test for equivalence, we have to maximize  $Q_{\alpha,n}^{\delta_1,\delta_2}(\pi, \eta)$  over the set of all  $(\pi, \eta) \in (0, 1)^2$  satisfying the above [ $\rightarrow$  (5.6a)] null hypothesis  $H$  of nonequivalence. As an immediate consequence of the definition of the parametric functions  $\pi$ ,  $\delta$  and  $\eta$ , one has  $\pi = (\delta + \eta)/2\eta$  so that for any fixed  $\eta > 0$  the  $\eta$ -section of  $H$  equals the complement of the interval  $(1/2 - \delta_1/2\eta, 1/2 + \delta_2/2\eta)$ . Hence, for the size of the test we obtain the formula

$$\begin{aligned} \text{SIZE}_{\alpha}(\delta_1, \delta_2; n) = \sup \left\{ Q_{\alpha,n}^{\delta_1,\delta_2}(\pi, \eta) \mid \begin{array}{l} 0 < \eta < 1, \\ 0 < \pi \leq 1/2 - \delta_1/2\eta \text{ or } 1/2 + \delta_2/2\eta \leq \pi < 1 \end{array} \right\}. \end{aligned} \quad (5.14)$$

Qualitatively speaking, the results to be obtained by evaluating the right-hand side of (5.14) numerically admit the conclusion that the finite-sample behavior of the asymptotic test given by (5.9) is surprisingly poor (in the subsequent sections and chapters, the reader will find numerous applications of the theory of §3.4 to much more complex models yielding very satisfactory solutions to the respective equivalence testing problem). In fact, the convergence of the exact size of the test to its limiting value, i.e., the target significance level  $\alpha$ , turns out to be extremely slow, with a general tendency towards anticonservatism. For instance, in the specific case of Example 5.2, the exact size of the critical region is seen to be as large as .0721 although the size of the sample distinctly exceeds that available for many equivalence studies conducted in practice. Even if the number of observations in the sample is raised to 150 and thus more than doubled, the smallest significance level maintained by the asymptotic test is still .0638 rather than .05.

Clearly, these results, together with many other examples pointing in the same direction, strongly suggest to replace the asymptotic solution to the problem of testing for equivalence in McNemar's setting by a finite-sample version corrected with respect to the significance level. Provided repeated evaluation of (5.14) with sufficiently high numerical precision causes no serious problem with respect to computing time, such a correction is easily carried out: All that is left to do is to determine by means of an iteration process the largest nominal significance level  $\alpha^*$  such that  $\text{SIZE}_{\alpha^*}(\delta_1, \delta_2; n) \leq \alpha$ . Of course, it is tempting to speed up this procedure by restricting at each step the search for the maximum probability of an incorrect decision in favor of equivalence to the boundary of the null hypothesis  $H$  of (5.6a). Since, in terms of  $(\pi, \eta)$ , the latter is given as the set  $\{(1/2 - \delta_1/2\eta, \eta) \mid \delta_1 < \eta < 1\} \cup \{(1/2 + \delta_2/2\eta, \eta) \mid \delta_2 < \eta < 1\}$ , this leads to determining for varied values of  $\alpha$

$$\begin{aligned} \text{SIZE}_{\alpha}^B(\delta_1, \delta_2; n) = \max \left\{ \sup_{\eta > \delta_1} Q_{\alpha,n}^{\delta_1,\delta_2}(1/2 - \delta_1/2\eta, \eta), \right. \\ \left. \sup_{\eta > \delta_2} Q_{\alpha,n}^{\delta_1,\delta_2}(1/2 + \delta_2/2\eta, \eta) \right\} \end{aligned} \quad (5.15)$$

instead of  $SIZE_\alpha(\delta_1, \delta_2; n)$  as obtained by searching through the whole set of parameter values compatible with  $H$ . Although, in general we can only state that there holds the inequality  $SIZE_\alpha^B(\delta_1, \delta_2; n) \leq SIZE_\alpha(\delta_1, \delta_2; n)$ , taking for granted that the sets  $\mathcal{U}_{\alpha, n}^{\delta_1, \delta_2}(v) [\rightarrow (5.13)]$  are intervals allows to restrict the search for the maximum probability of a type-I error to the common boundary of the hypotheses (5.6a) whenever the equivalence range for  $\delta$  is chosen as a symmetric interval. In fact,  $\delta_1 = \delta_2$  clearly implies that each such interval in the sample space of  $U|V=v$  must be symmetric about  $v/2$  from which it can be concluded by means of elementary arguments (starting from the well-known relationship between the binomial cdf and the incomplete beta integral — see, e.g., Feller, 1968, p. 173) that, as a function of  $\pi$ , the  $v$ th inner sum of (5.12) is nondecreasing on  $(0, 1/2]$  and nonincreasing on  $[1/2, 1)$ . The program `mcnemasc` [ $\rightarrow$  Appendix B] which for cases with  $\delta_1 = \delta_2 = \delta_0$  serves the purpose of determining a nominal significance level  $\alpha^*$  such that the exact size of the respective test for (5.6a) approaches the target level from below as closely as possible, exploits this possibility of simplifying the search for the maximum of the rejection probability (5.12) over the null hypothesis.

For two symmetric choices of the equivalence range  $(-\delta_1, \delta_2)$  for  $\delta = p_{10} - p_{01}$  and selected sample sizes ranging from 50 to 200, corrected nominal levels computed in this way are shown in Table 5.5, together with the exact size of the corrected test and that of the original large-sample procedure. All results

Table 5.5 *Nominal significance level and exact size for the corrected McNemar test for equivalence maintaining the 5% level in finite samples of size  $n = 50(25)200$ , for  $\delta_1 = \delta_2 = \delta_0 = .2, .4$ . [Number in (): size of the noncorrected test at nominal level  $\alpha = .05$ .]*

$\delta_0$	$n$	$\alpha^*$	$SIZE_{\alpha^*}(\delta_0, \delta_0; n)$	
.20	50	.024929	.04801	(.10338)
"	75	.016502	.02671	(.09275)
"	100	.036807	.04884	(.08444)
"	125	.037268	.04766	(.07047)
"	150	.032418	.04137	(.06376)
"	175	.027880	.03581	(.07501)
"	200	.039334	.04919	(.06396)
.40	50	.029388	.03532	(.05651)
"	75	.039807	.04201	(.06257)
"	100	.043173	.04737	(.06150)
"	125	.043206	.04601	(.05981)
"	150	.041632	.04358	(.05608)
"	175	.039291	.04076	(.05363)
"	200	.046252	.04920	(.05327)

presented in this table have been obtained by means of `mcnemasc` setting the fineness of the grid of  $\eta$ -values searched through for the maximum rejection probability under nonequivalence [recall (5.14)] equal to .0005 throughout. Strikingly, neither the difference between the target significance level  $\alpha$  and the nominal level to be used in the corrected test, nor between  $\alpha$  and the size attainable under the correction decreases monotonically with  $n$ , and even with samples comprising several hundred observations it happens that the nominal level has to be halved.

Once the corrected nominal level has been determined, computing the exact power of the corresponding finite-sample test for the equivalence problem (5.6a) against arbitrary specific alternatives  $\delta \in (-\delta_1, \delta_2)$  is comparatively easy since it requires only a single evaluation of the double sum on the right-hand side of (5.12). However, given the alternative of interest in terms of  $\delta$ , the power depends in a rather complicated way on the nuisance parameter  $\eta$ . As becomes obvious from the numerical example shown in Table 5.6, the power is in particular neither monotonic nor concave in  $\eta$ . Generally, the program `mcnempow` allows the computation of the exact rejection probability of the test (5.9) at any nominal significance level  $\tilde{\alpha}$  under arbitrary parameter constellations  $(p_{10}, p_{01})$ .

Table 5.6 *Exact power of the corrected version of (5.9) against the alternative  $\delta = 0$  [ $\Leftrightarrow \pi = 1/2$ ] for  $n = 50$ ,  $\delta_1 = \delta_2 = .20$ ,  $\alpha^* = .024902$  and various values of the nuisance parameter  $\eta$ .*

$\eta$	.0002	.002	.005	.01	.02
$Q_{\alpha^*, 50}^{.20, .20}(1/2, \eta)$	.00995	.09525	.22169	.39499	.63581
	.20	.40	.60	.80	1.00
	.73720	.29702	.11515	.06281	.11228

### Example 5.2 (continued)

For  $n = 72$  and  $\delta_1 = \delta_2 = .20$ , the smallest nominal level  $\alpha^*$  to be used in (5.9) in order to maintain the target significance level  $\alpha = .05$ , is obtained (by running the program `mcnemasc`) to be  $\alpha^* = .027148$ . Since the noncentrality parameter of the  $\chi^2$ -distribution through which the critical upper bound to the test statistic  $n^{1/2}|(N_{10} - N_{01})|/[n(N_{10} + N_{01}) - (N_{10} - N_{01})^2]^{1/2}$  has to be determined, has been estimated by means of the data of the present example to be  $\hat{\psi}^2 = 23.075611$ , recomputing the critical bound with  $\alpha^*$  instead of  $\alpha$  gives:  $C_{\alpha^*}(\hat{\psi}) = \sqrt{\text{cinv}(.027148, 1, 23.075611)} = \sqrt{8.290026} = 2.8792$ , after another application of the SAS intrinsic function `cinv`. This bound is

still much larger than the value which has been observed for the test statistic so that the decision keeps the same as if the nominal level was set equal to the target significance level. Finally, let us assume that the true rates of discordant pairs of both kinds in the underlying population coincide with the corresponding relative frequencies of the observed  $2 \times 2$  table so that we have  $p_{10} = 4/9$ ,  $p_{01} = 5/9$ . Running program `mcnempow`, the power of the corrected test against this “observed alternative” is found to be only slightly larger than the target significance level, namely .0599, as compared to .1794 which turns out to be the power of the noncorrected asymptotic test (5.9) against the same alternative.

### 5.2.3 Modifications for the noninferiority case

The noninferiority analogue of the asymptotic test given by (5.9) rejects the null hypothesis  $H_1 : \delta \equiv p_{10} - p_{01} \leq -\delta_0$  if and only if the observed value of  $(N_{10}, N_{10} + N_{01}) \equiv (U, V)$  turns out to fall in the critical region

$$\left\{ (u, v) \in \mathbb{N}_0^2 \mid u \leq v \leq n, T_n^{\delta_0}(u, v) > u_{1-\alpha} \right\}, \quad (5.16)$$

where

$$T_n^{\delta_0}(u, v) = \frac{\sqrt{n}((2u - v) + n\delta_0)}{[nv - (2u - v)^2]^{1/2}}$$

and  $u_{1-\alpha}$  denotes the  $(1 - \alpha)$ -quantile of the standard normal distribution [recall (2.6)]. In order to exactly control the significance level in finite samples, the test has likewise to be carried out at an corrected nominal level  $\alpha^*$  which can be determined by similar numerical methods as were described above for the case of two-sided equivalence. The starting point of the computational procedure is the following expression for the exact rejection probability of the test given by (5.16) under an arbitrary parameter constellation  $(\pi, \eta) \equiv (p_{10}/(p_{10} + p_{01}), p_{10} + p_{01})$ :

$$Q_{\alpha, n}^{\delta_0}(\pi, \eta) = \sum_{v=0}^n \left\{ \sum_{u \in \mathcal{U}_{\alpha, n}^{\delta_0}(v)} b(u; v, \pi) \right\} b(v; n, \eta). \quad (5.17)$$

The only difference between (5.12) and (5.17) refers to the inner summation region which is now given by

$$\mathcal{U}_{\alpha, n}^{\delta_0}(v) = \left\{ u \in \mathbb{N}_0 \mid u \leq v, \frac{\sqrt{n}((2u - v) + n\delta_0)}{[nv - (2u - v)^2]^{1/2}} > u_{1-\alpha} \right\}. \quad (5.18)$$

In contrast to the  $v$ -sections  $\mathcal{U}_{\alpha, n}^{\delta_1, \delta_2}(v)$  of the critical region (5.11) of the asymptotic test for the two-sided version of the equivalence problem under

discussion, the sets  $\mathcal{U}_{\alpha,n}^{\delta_0}(v)$  can be shown analytically to exhibit the appropriate form. A sufficient condition which ensures that for each  $v = 1, \dots, n$ , there exists some integer  $k_v \in \{0, 1, \dots, v\}$  such that

$$\mathcal{U}_{\alpha,n}^{\delta_0}(v) = \{u \in \mathbb{N}_0 \mid k_v \leq u \leq v\} \quad (5.19)$$

is that we have

$$n > \frac{u_{1-\alpha}^2}{4\delta_0(1 + \delta_0)}. \quad (5.20)$$

Since in a binomial distribution with parameter  $\pi \in (0, 1)$ , the probability of any set of the form (5.19) is well known to be an increasing function of  $\pi$  (cf. Johnson et al., 1992, p. 117), it follows that as long as the very mild condition (5.20) is satisfied, the size of the test with critical region (5.16) can be computed from

$$SIZE_\alpha(\delta_0; n) = \sup_{\delta_0 < \eta < 1} Q_{\alpha,n}^{\delta_0}(1/2 - \delta_0/2\eta, \eta). \quad (5.21)$$

In other words, the facts stated above ensure that in the noninferiority case, there is no need to make a distinction between the supremum of the probability of a false rejection of the null hypothesis taken over  $H_1$  as a whole on the one hand and its boundary on the other. The practical use of (5.21) is the same as that of (5.15): It serves as the basis of an iteration algorithm for finding the largest nominal level  $\alpha^*$  to be used in the asymptotic test in order to ensure that its exact rejection probability remains below the target significance level  $\alpha$  under any parameter configuration belonging to the null hypothesis. The algorithm is implemented in the program named `mcnasc_ni` whose source code can likewise be downloaded from the **WKTSEQ2 Source Code Package**.

Table 5.7 whose entries were computed by means of this tool, gives corrected nominal significance levels to be used in order to maintain the target level  $\alpha = .05$  in the asymptotic McNemar test for noninferiority for two choices of  $\delta_0$  ( $\delta_0 = .05$  and  $\delta_0 = .10$ ) and different sample sizes ranging from 50 through 200. The power of the level-corrected test for noninferiority against a selection of alternatives specifying  $\delta = 0$  and the same range of values of  $\delta_0$  and  $n$  as covered by the previous table is shown in Table 5.8.

Table 5.7 *Nominal significance level and exact size for the corrected McNemar test for noninferiority maintaining the 5% level in finite samples of size  $n \in \{50, 60, 80, 100, 150, 200\}$  for  $\delta_0 = .05$ , and  $\delta_0 = .10$ . [Number in (): size of the noncorrected test at nominal level  $\alpha = .05$ .]*

$\delta_0$	$n$	$\alpha^*$	$SIZE_{\alpha^*}(\delta_0; n)$	
.05	50	.032178	.048617	(.087571)
"	60	.016943	.031673	(.069655)
"	80	.029199	.048647	(.079308)
"	100	.036621	.047380	(.069718)
"	150	.025684	.046480	(.073553)
"	200	.028320	.049189	(.073399)
.10	50	.036523	.045439	(.066661)
"	60	.029199	.048548	(.065415)
"	80	.028516	.048663	(.078042)
"	100	.027881	.046130	(.069658)
"	150	.031738	.044867	(.067092)
"	200	.034180	.046787	(.062220)

Table 5.8 *Power of the level-corrected asymptotic McNemar test for noninferiority against different alternatives with  $\delta = 0$  and the same choices of  $\delta_0$  and  $n$  as in Table 5.7.*

$\delta_0$	$n$	$\eta =$							
		.0002	.002	.02	.20	.30	.50	.80	
.05	50	.004963	.04651	.26770	.16990	.12349	.09482	.07760	
"	60	.005947	.05502	.28939	.11339	.08736	.06334	.05163	
"	80	.007968	.07670	.47530	.20673	.14768	.10803	.08476	
"	100	.009951	.09519	.58899	.25922	.19765	.14377	.10986	
"	150	.014889	.13936	.76559	.29294	.21168	.14135	.10545	
"	200	.019802	.18135	.86281	.37670	.27483	.18483	.13320	
.10	50	.004988	.04879	.38991	.43374	.32404	.21837	.16642	
"	60	.005982	.05826	.45150	.46770	.32046	.22279	.15709	
"	80	.007968	.07692	.55199	.54714	.39558	.26195	.18390	
"	100	.009951	.09521	.63393	.63072	.47378	.30981	.21766	
"	150	.014889	.13936	.77855	.81395	.65072	.45046	.31403	
"	200	.019802	.18135	.86602	.90715	.77786	.57018	.40633	

An alternative approach to the problem of testing with a sample of paired binary data for one-sided equivalence with respect to the difference of both success probabilities is through applying the objective Bayesian methodology outlined in general terms in § 2.4. Collapsing the two cells on the main diagonal of the underlying contingency table [→ Tab. 5.3] into a single category leads to a trimomial model described by the random triplet  $(N_{10}, N_{01}, \tilde{N}_0)$  with parameter vector  $(p_{10}, p_{01}, \tilde{p}_0)$  where  $\tilde{N}_0$  stands for the total count  $N_{00} + N_{11}$  of concordant pairs and  $\tilde{p}_0$  for the probability  $p_{00} + p_{11}$  of observing a pair of that kind. The arguments by which this reduction can be justified are manifold, and there is no reason to review them here (for a recent detailed discussion see Lloyd, 2008a, § 2.2). According to Jeffreys' (1961) rule, the standard noninformative prior of  $(p_{10}, p_{01})$  is the two-dimensional Dirichlet distribution given by the density

$$g(p_{10}, p_{01}) = \frac{1}{2\pi} p_{10}^{-1/2} p_{01}^{-1/2} (1 - p_{10} - p_{01})^{-1/2}, \quad 0 < p_{10} + p_{01} < 1. \quad (5.22)$$

The Bayesian test of  $H_1 : p_{10} - p_{01} \leq -\delta_0$  vs.  $K_1 : p_{10} - p_{01} > -\delta_0$  based on (5.22) is fairly easy to compute and admits exact calculation of its size and power against any specific alternative  $(p_{10}, p_{01})$  with  $\delta \equiv p_{10} - p_{01} > -\delta_0$ . Under the model  $(N_{10}, N_{01}, \tilde{N}_0) \sim \mathcal{M}(n; p_{10}, p_{01}, \tilde{p}_0)$ , the joint posterior distribution associated with this prior is well known (see, e.g., Berger, 1985, pp. 287, 561) to be again Dirichlet, but with data-dependent parameters  $(n_{10} + 1/2, n_{01} + 1/2, \tilde{n}_0 + 1/2)$ . The corresponding density function is

$$g(p_{10}, p_{01} | n_{10}, n_{01}, \tilde{n}_0) = \frac{\Gamma(n + 3/2)}{\Gamma(n_{10} + 1/2)\Gamma(n_{01} + 1/2)\Gamma(\tilde{n}_0 + 1/2)} \cdot p_{10}^{n_{10}-1/2} p_{01}^{n_{01}-1/2} (1 - p_{10} - p_{01})^{\tilde{n}_0-1/2}. \quad (5.23)$$

Making use of some of the basic properties of Dirichlet distributions (see Kotz et al., 2000, § 49.1), the posterior probability of the subset  $\mathcal{K}_1 \equiv \{(p_{10}, p_{01}) | 0 < p_{10} + p_{01} < 1, p_{10} - p_{01} > -\delta_0\}$  of the parameter space specified by the non-inferiority hypothesis  $K_1$  to be established can be written

$$\begin{aligned} P[\mathcal{K}_1 | n_{10}, n_{01}, \tilde{n}_0] &= I_{\delta_0}(n_{01} + 1/2, n - n_{01} + 1) + \\ &\int_{\delta_0}^{(1+\delta_0)/2} \left[ B(n_{01} + 1/2, n - n_{01} + 1/2) p_{01}^{n_{01}-1/2} (1 - p_{01})^{n-n_{01}-1/2} \right. \\ &\quad \left. \cdot \left( 1 - I_{(p_{01}-\delta_0)/(1-p_{01})}(n_{10} + 1/2, \tilde{n}_0 + 1/2) \right) \right] dp_{01}, \end{aligned} \quad (5.24)$$

where

$$B(a, b) \equiv \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)}, \quad a, b > 0,$$

$$I_u(a, b) \equiv \text{cdf at } u \in (0, 1) \text{ of a beta distribution with parameters } a, b > 0.$$

The integral on the right-hand side of this equation is just one-dimensional and can thus be evaluated by means of a standard quadrature procedure with comparably low computational effort. Again, all numerical results to be presented about the test based on (5.24) were obtained by means of 96-point Gauss-Legendre integration. In order to ensure a high degree of numerical accuracy, the range of integration was partitioned into 10 subintervals over each of which integration was done separately using this rule. Even in this high-accuracy version, the algorithm is fast enough for allowing us to search through the whole sample space of  $(N_{10}, N_{01}, \tilde{n}_0)$  checking each of its points for inclusion in the critical region

$$\left\{ (n_{10}, n_{01}, \tilde{n}_0) \in \mathbb{N}_0^3 \mid n_{10} + n_{01} + \tilde{n}_0 = n, P[\mathcal{K}_1 \mid n_{10}, n_{01}, \tilde{n}_0] \geq 1 - \alpha \right\}. \quad (5.25)$$

As before, it is more convenient to represent this region as a subset of the sample space of the statistic  $(U, V)$  introduced in (5.10) yielding

$$\left\{ (u, v) \in \mathbb{N}_0^2 \mid u \leq v, P[\mathcal{K}_1 \mid u, v - u, n - v] \geq 1 - \alpha \right\}. \quad (5.26)$$

The remaining steps to be carried out in order to determine a largest nominal significance level  $\alpha^*$  which is admissible in the sense of generating a rejection region whose probability maximized over the null hypothesis  $H_1$  remain below the target level  $\alpha$ , are the same as before for the Wald type test: Again, irrespective of the choice of the nominal significance level in (5.26) and for any  $v = 0, \dots, n$ , the set of all  $u$  satisfying the critical inequality  $P[\mathcal{K}_1 \mid u, v - u, n - v] \geq 1 - \alpha$  turns out to be of the form (5.19). Consequently, the size of the critical region (5.26) can be computed in the same way as before [recall (5.21)], and the corrected nominal level  $\alpha^*$  is readily found by bisection. A tool for implementing the complete computational procedure which underlies the objective Bayesian solution of the problem of constructing an exactly valid test for noninferiority in terms of  $\delta$  for the matched-pair design with binary data, has been made available under the program name `mcnby_ni`. The program is designed to output exact values of the power against specific null alternatives, i.e., parameter constellations with  $\delta = 0$  in addition. The selection of numerical results shown in Tables 5.9 and 5.10 have been computed by means of this tool for the same constellations covered by the preceding pair of tables.

For a brief illustration of the level-corrected asymptotic and the objective Bayesian test for noninferiority, let us reanalyze Table 5.4 using a noninferiority margin of  $\delta_0 = .05$ . With this specification, the Wald type statistic  $T_n^{\delta_0}(u, v)$  is found to have value 0.86734 which is clearly insignificant. Evaluating the integral on the right-hand side of (5.24) for  $n = 72$ ,  $\delta_0 = .05$ ,  $n_{10} = 4$ ,  $n_{01} = 5$  through running the program `mcnby_ni_pp` which serves the purpose of computing the posterior probability of  $\mathcal{K}_1$  for individual points in the sample space, yields  $P[\mathcal{K}_1 \mid n_{10}, n_{01}, \tilde{n}_0] = .812488 < .95$  so that the null hypotheses  $H_1 : \delta < .05$  of substantial inferiority of the “combined referee” to

the computer-aided diagnostic procedure has to be accepted by both tests.

Table 5.9 *Nominal significance level and exact size for the objective Bayes solution to the problem of testing for noninferiority in terms of  $\delta$  with a sample of paired binary variables of size  $n \in \{50, 60, 80, 100, 150, 200\}$  for  $\delta_0 = .05$  and  $\delta_0 = .10$ . [Number in (): size of the noncorrected test when the nominal level is chosen equal to the target significance level  $\alpha = .05$ .]*

$\delta_0$	$n$	$\alpha^*$	$SIZE_{\alpha^*}(\delta_0; n)$	
.05	50	.0365381	.0486168	(.0578281)
"	60	.0395752	.0432548	(.0696555)
"	80	.0299854	.0330561	(.0688144)
"	100	.0441748	.0497734	(.0573983)
"	150	.0244629	.0347813	(.0627488)
"	200	.0320557	.0393598	(.0672558)
.10	50	.0421094	.0481380	(.1056926)
"	60	.0204492	.0258535	(.0630695)
"	80	.0387744	.0463279	(.0866751)
"	100	.0285938	.0364573	(.0626736)
"	150	.0342871	.0396826	(.0628134)
"	200	.0346631	.0435619	(.0637456)

Table 5.10 *Power of the objective Bayesian test for one-sided  $\delta$ -equivalence of binomial proportions with paired data against different alternatives with  $\delta = 0$  and the same choices of  $\delta_0$  and  $n$  as in Table 5.9.*

$\delta_0$	$n$	$\eta =$							
		.0002	.002	.02	.20	.30	.50	.80	
.05	50	.004963	.046513	.267703	.170682	.133745	.101893	.084922	
"	60	.005947	.055016	.289644	.191543	.145117	.115914	.097100	
"	80	.007968	.076695	.454858	.189916	.143495	.108030	.084760	
"	100	.019755	.177215	.724317	.287333	.213761	.162102	.125306	
"	150	.029557	.258976	.879858	.274944	.199468	.136064	.099494	
"	200	.039214	.329947	.968756	.391696	.291472	.199715	.144822	
.10	50	.009939	.094122	.575710	.436718	.347309	.240782	.178784	
"	60	.011912	.111577	.625577	.386520	.279759	.183838	.120505	
"	80	.015874	.147991	.796701	.604435	.459459	.320297	.223178	
"	100	.019803	.181433	.865856	.627323	.474800	.323541	.217856	
"	150	.029557	.259404	.951697	.818218	.666396	.471277	.330827	
"	200	.039214	.329948	.982412	.906756	.784520	.576421	.411560	

### Discussion

Comparing the objective Bayesian with the level-corrected asymptotic test in terms of the power achieved under the specific alternatives appearing in Tables 5.8 and 5.10 does not lead to a clear-cut preference: As holds typically true for any exactly valid nonrandomized test of a hypothesis relating to a family of discrete distributions, both tests fail to be unbiased in that their power converges to zero as the probability  $\eta$  of observing a discordant pair [i.e., finding that  $(X_i, Y_i)$  equals either  $(1, 0)$  or  $(0, 1)$ ] becomes arbitrarily small. Furthermore, for both tests the power against fixed alternatives considered as a function of the sample size  $n$ , fails to be monotonically increasing although the exceptions do not seem to be too serious an issue from a practical perspective. Looking at the rowwise maxima of the values shown in the tables, the Bayesian test comes out to be clearly superior. However, these maxima are attained against the alternative  $p_{10} = p_{01} = .01$ , and settings where the true proportion of discordant pairs is as low as 2% will rarely if ever occur in real applications. For the other, much more realistic alternatives, the powers of both tests do not differ substantially from each other.

The journal literature contains several proposals to address the problem of constructing exact nonconditional tests for one-sided equivalence of correlated binomial proportions from a conceptually different perspective: Instead of adjusting the nominal level as far as necessary for ensuring that the supremum of the probability of the resulting rejecting region taken over the nuisance parameter  $\eta$  does not exceed the target value of  $\alpha$ , maximized p-values are compared to the latter. In the work of Sidik (2003) (for the special case  $\delta_0 = 0$  of the traditional one-sided testing problem see also Berger and Sidik, 2003), the technique introduced by Berger and Boos (1994) is adopted, which is to say that p-value maximization is done over a confidence interval rather than the whole range of  $\eta$ , adding the risk of noncoverage to the result as a correction term. The test proposed by Lloyd and Moldovan (2008) is based on so-called E+M p-values (for a general discussion and other applications of this concept see Lloyd, 2008a, b) which are obtained in two steps: Suppose that some “starting p-value” has been selected whose computation from the data  $(U, V)$  involves no unknown parameters, e.g.,  $p(u, v) = 1 - \Phi(T_n^{\delta_0}(u, v))$  with  $T_n^{\delta_0}(u, v)$  defined as at the beginning of this subsection. In a first step, the probability of the event  $\{p(U, V) \leq p(u, v)\}$  that this p-value falls below its observed value, is evaluated replacing the nuisance parameter  $\eta$  with a suitable point estimator  $\tilde{\eta}$  (which might depend on the noninferiority margin  $\delta_0$ ). Denoting the result by  $\tilde{P}(u, v)$ , the second step consists of finding the maximum, say  $P^*(u, v)$ , of the exact probability of  $\{\tilde{P}(U, V) \leq \tilde{P}(u, v)\}$  taken over the whole boundary of the null hypothesis  $H_1$ , i.e., all values of  $\eta$  being compatible with  $\delta = -\delta_0$ . Finally, the rejection region consists of all values  $(u, v)$  in the sample space of  $(U, V)$  satisfying  $P^*(u, v) \leq \alpha$ .

Obviously, all approaches involving maximized p-values are computationally much more demanding than those based on corrected nominal significance

levels since the maximization process applies to the individual points in the sample space rather than the rejection region as a whole. The return in terms of power which the user of these tests can expect for this extra computational effort, is only moderate: For  $\delta_0 = .05$ ,  $n = 50$  and  $\eta \in \{.20, .30, .50\}$ , Sidik (2003) obtained the power values .175, .154 and .119 which are not much different from the respective entries in Tables 5.8 and 5.10. Sensible comparisons with the E+M test proposed by Lloyd and Moldovan (2008) are difficult since these authors provided data about power only in graphical, not in numerical form and for the choice  $\alpha = .10$  which is very unusual in real applications. Roughly speaking, the most substantial differences will occur under small values of the probability  $\eta$  of obtaining a discordant pair, reflecting the fact that the power of the E+M test converges to unity rather than zero as  $\eta$  approaches 0. This is made possible by including the point  $(u, v) = (0, 0)$  in the rejection region implying that the power approaches unity even for  $n = 1$ . Since attaining maximum power with minimum sample size is clearly counter to common statistical sense, Lloyd and Moldovan's test cannot be recommended for real applications even if computational burden is not an issue.

---

### 5.3 Paired $t$ -test for equivalence

In this section, we assume throughout that the setting described in § 5.1 has to be considered in the special version that the distribution of the intraindividual differences  $D_i$  between the measurements taken under both treatments is Gaussian with (unknown) parameters  $\delta = E(D_i)$  and  $\sigma_D^2 = \text{Var}(D_i)$ . In the special case

$$D_i \sim \mathcal{N}(\delta, \sigma_D^2) \quad \forall i = 1, \dots, n, \quad (5.27)$$

the hypothesis of equivalence with respect to the probability of a positive sign [cf. (5.1), (5.4)] reduces to a statement about the *standardized* expected value  $\delta/\sigma_D$ . In fact, under the parametric model (5.27), we have  $p_0 = 0$  and hence  $p_- = 1 - p_+$ . Consequently, we may write:  $-\varepsilon_1 < \log(p_+/p_-) < \varepsilon_2 \Leftrightarrow -\varepsilon_1 < \log(p_+/(1 - p_+)) < \varepsilon_2 \Leftrightarrow e^{-\varepsilon_1} < p_+/(1 - p_+) < e^{\varepsilon_2} \Leftrightarrow e^{-\varepsilon_1}/(1 + e^{-\varepsilon_1}) < p_+ < e^{\varepsilon_2}/(1 + e^{\varepsilon_2})$ . On the other hand, under (5.27),  $p_+$  admits the representation  $p_+ = P[D_i > 0] = 1 - P[(D_i - \delta)/\sigma_D < -\delta/\sigma_D] = 1 - \Phi(-\delta/\sigma_D) = \Phi(\delta/\sigma_D)$ . Since the standard normal distribution function  $\Phi$  is strictly increasing and continuous, we eventually see that under (5.27), the hypothesis  $-\varepsilon_1 < \log(p_+/p_-) < \varepsilon_2$  to be established by means of the sign test for equivalence is logically equivalent to  $\Phi^{-1}(e^{-\varepsilon_1}/(1 + e^{-\varepsilon_1})) < \delta/\sigma_D < \Phi^{-1}(e^{\varepsilon_2}/(1 + e^{\varepsilon_2}))$ . Hence, setting for brevity  $\Phi^{-1}(e^{-\varepsilon_1}/(1 + e^{-\varepsilon_1})) = \theta_1$ ,  $\Phi^{-1}(e^{\varepsilon_2}/(1 + e^{\varepsilon_2})) = \theta_2$ , the testing problem which in its nonparametric form was dealt with in § 5.1, can now be written

$$H : \delta/\sigma_D \leq \theta_1 \vee \delta/\sigma_D \geq \theta_2 \quad \text{vs.} \quad K : \theta_1 < \delta/\sigma_D < \theta_2. \quad (5.28)$$

Evidently, this problem does not change if the observations  $D_i$  are rescaled by multiplying each of them by the same positive constant  $c$ . Put in more technical terms, this means that the parametric equivalence testing problem (5.28) remains invariant under all transformations of the sample space  $\mathbb{R}^n$  of  $(D_1, \dots, D_n)$  taking the form of  $(d_1, \dots, d_n) \mapsto (cd_1, \dots, cd_n)$ , with arbitrary  $c > 0$ . As is shown in Lehmann and Romano (2005, § 6.4), the construction of a test of the ordinary one-sided null hypothesis about  $\delta/\sigma_D$  which is uniformly most powerful among all tests being invariant against all transformations of that type, conveniently starts from reducing the data  $(D_1, \dots, D_n)$  to the sufficient statistics  $(\bar{D}, S_D)$  before applying the principle of invariance. The same two-step reduction can be carried out in the present context which implies that a uniformly most powerful invariant (UMPI) level  $\alpha$  test for (5.28) is obtained in the following way: After reducing the primary data set  $(D_1, \dots, D_n)$  to the usual one-sample  $t$ -statistic

$$T = \sqrt{n}\bar{D}/S_D, \quad (5.29)$$

a UMP level- $\alpha$  test based on  $T$  is carried out for

$$\tilde{H} : \tilde{\theta} \leq \sqrt{n}\theta_1 \vee \tilde{\theta} \geq \sqrt{n}\theta_2 \quad \text{vs.} \quad \tilde{K} : \sqrt{n}\theta_1 < \tilde{\theta} < \sqrt{n}\theta_2, \quad (\widetilde{5.28})$$

where  $\tilde{\theta}$  denotes the parameter of a noncentral  $t$ -distribution with  $n-1$  degrees of freedom.

Denoting the density function of that distribution by  $g_{\tilde{\theta}}(\cdot)$ , we know from a result derived in Karlin (1968, §4(iii)) [cf. also Lemma A.1.3 of Appendix A to this book] that the family  $(g_{\tilde{\theta}})_{\tilde{\theta} \in \mathbb{R}}$  is strictly positive of any order and hence *a fortiori* STP<sub>3</sub>. Furthermore, for arbitrary  $(\tilde{\theta}, t)$ ,  $g_{\tilde{\theta}}(t)$  admits the representation

$$g_{\tilde{\theta}}(t) = e^{-\tilde{\theta}^2/2}(n-1+t^2)^{-n/2} \sum_{j=0}^{\infty} c_j [h(t, \tilde{\theta})]^j, \quad (5.30)$$

where

$$h(t, \tilde{\theta}) = t\tilde{\theta}\sqrt{2}/\sqrt{n-1+t^2} \quad (5.31)$$

and the power series  $\sum_{j=0}^{\infty} c_j x^j$  converges for any  $x \in \mathbb{R}$  (see Johnson et al., 1995, p. 516). This ensures continuity of  $g_{\tilde{\theta}}(t)$  in both of its arguments so that the family  $(g_{\tilde{\theta}})_{\tilde{\theta} \in \mathbb{R}}$  of noncentral  $t$ -densities with an arbitrary number  $(n-1)$  of degrees of freedom satisfies all the conditions of Theorem A.1.5 [ $\rightarrow$  Appendix, p. 371]. Hence, we may conclude that an UMP level- $\alpha$  test for (5.28) exists and is given by the critical region

$$\left\{ \tilde{C}_{\alpha; n-1}^1(\theta_1, \theta_2) < T < \tilde{C}_{\alpha; n-1}^2(\theta_1, \theta_2) \right\}, \quad (5.32)$$

where the bounds  $\tilde{C}_{\alpha; n-1}^{\nu}(\theta_1, \theta_2)$  are uniquely determined by the equations

$$G_{\tilde{\theta}_1}(C_2) - G_{\tilde{\theta}_1}(C_1) = \alpha = G_{\tilde{\theta}_2}(C_2) - G_{\tilde{\theta}_2}(C_1), \quad -\infty < C_1 < C_2 < \infty. \quad (5.33)$$

In (5.33),  $G_{\tilde{\theta}}(\cdot)$  denotes for any  $\tilde{\theta} \in \mathbb{R}$  the cumulative distribution function associated to  $g_{\tilde{\theta}}(\cdot)$ , and  $\tilde{\theta}_\nu$  has to be set equal to  $\sqrt{n}\theta_\nu$ , both for  $\nu = 1$  and  $\nu = 2$ . Both in view of the form of the statistic  $T$  and the distribution which the computation of the critical constants has to be based upon, the term paired  $t$ -test seems fully appropriate for the testing procedure given by (5.32).

A glance at (5.30) and (5.31) shows that we have  $g_{-\tilde{\theta}}(-t) = g_{\tilde{\theta}}(t) \forall (\tilde{\theta}, t) \in \mathbb{R}^2$ . Hence, for any  $\tilde{\theta} > 0$ , the distribution of  $-T$  under  $-\tilde{\theta}$  coincides with that of  $T$  under  $\tilde{\theta}$  so that for a *symmetric choice of the equivalence interval*, i.e., for  $\theta_1 = -\varepsilon$ ,  $\theta_2 = \varepsilon$  and correspondingly for  $\tilde{\theta}_1 = -\tilde{\varepsilon}$ ,  $\tilde{\theta}_2 = \tilde{\varepsilon}$  with  $\varepsilon > 0$ ,  $\tilde{\varepsilon} = \sqrt{n}\varepsilon$ , the conditions of Lemma A.1.6 are satisfied as well. Applying this result we see that in the symmetric case, (5.32) and (5.33) can be simplified to

$$\left\{ |T| < \tilde{C}_{\alpha; n-1}(\varepsilon) \right\} \quad (5.34)$$

and

$$G_{\tilde{\varepsilon}}(C) - G_{\tilde{\varepsilon}}(-C) = \alpha, \quad 0 < C < \infty, \quad (5.35)$$

respectively, where  $\tilde{C}_{\alpha; n-1}(\varepsilon)$  stands for the solution of the latter equation. Furthermore, it is seen at once that the expression on the left-hand side of (5.35) is equal to the probability of the event  $\{|T| < C\} = \{T^2 < C^2\}$  under  $\tilde{\theta} = \tilde{\varepsilon}$ . Since it is a well known fact (cf. Johnson et al., 1995, loc. cit.) that squaring a random variable following a  $t$ -distribution with  $n - 1$  degrees of freedom and noncentrality parameter  $\tilde{\varepsilon}$  yields a variable which is noncentral  $F$  with  $1, n - 1$  degrees of freedom and  $\tilde{\varepsilon}^2$  as the noncentrality parameter, it follows that the solution to (5.35) admits the explicit representation

$$\tilde{C}_{\alpha; n-1}(\varepsilon) = \left[ F_{1, n-1; \alpha}(\tilde{\varepsilon}^2) \right]^{1/2}, \quad (5.35')$$

where  $F_{1, n-1; \cdot}(\tilde{\varepsilon}^2)$  denotes the quantile function of that  $F$ -distribution.

Due to the relationship (5.35'), when the equivalence hypothesis is formulated symmetrically setting  $-\theta_1 = \theta_2 = \varepsilon > 0$ , the practical implementation of the paired  $t$ -test for equivalence providing an UMPI solution to (5.28), involves only slightly more effort than that of the ordinary one- or two-sided  $t$ -test for paired observations. In fact, from the viewpoint of any user having access to up-to-date statistical software, the noncentral versions of the basic sampling distribution functions as well as their inverses, can be considered as explicitly given. In particular, the SAS system provides an intrinsic function named `finv` returning specific quantiles of  $F$ -distributions with arbitrary numbers of degrees of freedom and values of the noncentrality parameter with very high numerical accuracy (for details about the syntax to be used in calling this routine and the underlying algorithm see the SAS language reference dictionary). In the Stats Package of R, a program serving the same purpose can be run calling the function named `qf`. Table 5.11 enables the reader to perform the test in the symmetric case without any computational tools at

the 5% level, for sample sizes up to 100 being multiples of 10 and equivalence margin  $\varepsilon = .25(.25)1.00$ . The corresponding maximum rejection probabilities  $\tilde{\beta}_{\alpha; n-1}(\varepsilon)$  attained at  $\delta = 0$  ( $\Leftrightarrow P[D_i > 0] = 1/2$ ) are shown in Table 5.12. The latter are readily computed by means of the central *t*-distribution function  $G_0(\cdot)$ , say, using the formula

$$\tilde{\beta}_{\alpha; n-1}(\varepsilon) = 2G_0(\tilde{C}_{\alpha; n-1}(\varepsilon)) - 1. \quad (5.36)$$

Table 5.11 *Critical constant  $\tilde{C}_{.05; n-1}(\varepsilon)$  of the one-sample *t*-test for equivalence at level  $\alpha = 5\%$  in the symmetric case  $(\theta_1, \theta_2) = (-\varepsilon, \varepsilon)$ , for  $\varepsilon = .25(.25)1.00$  and  $n = 10(10)100$ .*

$n$	$\varepsilon =$			
	.25	.50	.75	1.00
10	0.08811	0.22188	0.73424	1.46265
20	0.11855	0.61357	1.67161	2.70944
30	0.16079	1.08722	2.40334	3.68005
40	0.21752	1.50339	3.02359	4.50470
50	0.29164	1.87104	3.57213	5.23475
60	0.38432	2.20387	4.06944	5.89697
70	0.49319	2.51030	4.52773	6.50742
80	0.61250	2.79581	4.95499	7.07666
90	0.73573	3.06421	5.35682	7.61210
100	0.85817	3.31826	5.73729	8.11913

Table 5.12 *Power attained at  $\theta = \delta/\sigma_D = 0$  when using the critical constants shown in Table 5.11 for the one-sample *t*-statistic.*

$n$	$\varepsilon =$			
	.25	.50	.75	1.00
10	0.06828	0.17064	0.51851	0.82241
20	0.09313	0.45323	0.88901	0.98610
30	0.12662	0.71411	0.97713	0.99905
40	0.17106	0.85921	0.99560	0.99994
50	0.22821	0.93268	0.99919	1.00000
60	0.29787	0.96855	0.99986	1.00000
70	0.37656	0.98559	0.99998	1.00000
80	0.45803	0.99350	1.00000	1.00000
90	0.53617	0.99711	1.00000	1.00000
100	0.60713	0.99873	1.00000	1.00000

*Example 5.3*

In an experimental study of the effects of increased intracranial pressure on the cortical microflow of rabbits, a preliminary test had to be done to ensure that the measurements would exhibit sufficient stability during a pre-treatment period of 15 minutes' duration. In a sample of  $n = 23$  animals, at the beginning and the end of that period the mean flow [ml/min/100g body mass] was observed to be  $\bar{X} = 52.65$  and  $\bar{Y} = 52.49$ , respectively, corresponding to a mean change of  $\bar{D} = 0.16$ ; the standard deviation of the intraindividual changes was computed to be  $S_D = 3.99$  (data from Ungersböck and Kempski, 1992). The equivalence limits for the standardized expected change of the microflow were chosen to be  $\mp\varepsilon$  with  $\varepsilon = .50$ .

For the values of  $n$  and  $\varepsilon$  applying to the present example, the noncentrality parameter of the  $F$ -distribution to be referred to for computing the critical upper bound to the absolute value of the  $t$ -statistic, is  $\psi^2 = 23 \cdot 0.50^2 = 5.75$ . In order to compute the critical bound using SAS, one has to write in a data step just the single statement `csquare=finv(.05,1,22,5.75)` which yields the output: `csquare=.576779`. From this, the critical constant itself is  $\tilde{C}_{.05;22}(.50) = \sqrt{.576779} = .7595$ .

On the other hand, for a sample of size  $n = 23$ , the  $t$ -value corresponding to the observed point  $(0.16, 3.99)$  in the space of the sufficient statistic  $(\bar{D}, S_D)$ , is .1923. Thus, we have  $|T| < \tilde{C}_{.05;22}(.50)$  indeed which implies that the observed value of  $(\bar{D}, S_D)$  belongs to the rejection region of our test. Accordingly, we can conclude that at the 5% level, the experimental data of the present example contain sufficient evidence in support of the hypothesis that the cortical microflow of rabbits does not change to a relevant extent over a time interval of 15 minutes during which no active treatment is administered. The power of the UMPI test at level  $\alpha = .05$  for (5.28) with  $\theta_1 = -.50$ ,  $\theta_2 = .50$  and sample size  $n = 23$  is computed by means of formula (5.36) at 54.44%.

If the equivalence hypothesis  $K$  of (5.28) is formulated in a nonsymmetric way, computation of the critical constants  $\tilde{C}_{\alpha; n-1}^\nu(\theta_1, \theta_2)$ ,  $\nu = 1, 2$ , of the one-sample  $t$ -test for equivalence becomes considerably more complicated. It requires an iterative algorithm, then, for solving the system (5.33) of equations numerically. An iteration procedure for accomplishing this task can be obtained by suitable modifications to the algorithm described in § 4.2 for the one-sample equivalence testing problem with exponentially distributed data. It suffices to replace the gamma distribution function  $F_{\sigma,n}^\Gamma(\cdot)$  being the basic building block there, by the noncentral  $t$ -distribution function  $G_{\tilde{\theta}_\nu}(\cdot)$  (for  $\nu = 1, 2$ ) and redefine the initial value  $C_1^\circ$  of  $C_1$ . Extensive experience has shown that  $C_1^\circ = (\tilde{\theta}_1 + \tilde{\theta}_2)/2 = \sqrt{n}(\theta_1 + \theta_2)/2$  is a sensible choice. The program `tt1st` which is again made available both as a SAS macro and a R function in the `WKTSEQ2 Source Code Package` implements this computational procedure for determining the critical constants of the one-sample  $t$ -test for equivalence in cases of an arbitrary choice of the equivalence range for  $\delta/\sigma_D$ .

Before moving on to a description of the noninferiority analogue of the paired *t*-test for equivalence, a remark on the possible role of the test discussed on the preceding pages in bioequivalence assessment seems in place: Under a set of conditions satisfied by the majority of comparative bioavailability trials conducted in practice (lognormality of the intraindividual bioavailability ratios; negligibility of period effects), the procedure has much to recommend it as an alternative to the conventional interval inclusion approach to so-called average bioequivalence [a more elaborate argument for this view will be given in Ch. 10].

### *Paired *t*-test for noninferiority*

In the noninferiority version of the testing problem considered in this section, the hypotheses read

$$H_1 : \delta/\sigma_D \leq -\varepsilon \quad \text{versus} \quad K_1 : \delta/\sigma_D \geq -\varepsilon, \quad (5.37)$$

where, as before,  $\varepsilon$  stands for a prespecified positive number. The same kind of argument which lead us to use the critical region (5.32) for testing for two-sided equivalence with respect to  $\delta/\sigma_D$  shows that in order to perform a UMPI level- $\alpha$  test for (5.37), we have to reject the null hypothesis  $H_1$  if and only if it turns out that

$$T > t_{n-1;1-\alpha}(-\sqrt{n}\varepsilon). \quad (5.38)$$

According to this critical inequality, the standard paired-sample *t*-statistic has to be compared with the  $(1 - \alpha)$ -quantile of a noncentral *t*-distribution with density obtained by making the substitution  $\tilde{\theta} = -\sqrt{n}\varepsilon$  in (5.30-1). Computation of these quantiles is as easy as of the noncentral *F*-quantiles to be used in the equivalence version of the test with symmetric choice of the margins. In SAS, the intrinsic function named `tinv` serves this purpose, and again, this has a direct counterpart in R, namely the function `qt`. The rejection probability under an arbitrary parameter constellation  $(\delta, \sigma_D)$  is given by  $1 - G_{\sqrt{n}\theta}(t_{n-1;1-\alpha}(-\sqrt{n}\varepsilon))$  with  $\theta = \delta/\sigma_D$  and  $G_{\tilde{\theta}}(\cdot)$  denoting the noncentral *t*-distribution function with generic nc-parameter  $\tilde{\theta}$  and  $df = n-1$ . In particular, the power against the null alternative  $\theta = 0 \Leftrightarrow \delta = 0$  is simply obtained by evaluating the central *t*-distribution function  $G_0(\cdot)$  at  $t = -t_{n-1;1-\alpha}(-\sqrt{n}\varepsilon)$ .

Making use of these tools and relations, it is easy to generate tables of critical constants and power values. For the same selection of sample sizes as considered in the two-sided case and equivalence margins  $\varepsilon \in \{k/10 \mid k = 1, \dots, 5\}$ , Table 5.13 shows the critical bound to which the *t*-statistic has to be compared in the UMPI test of (5.37) at level  $\alpha = .05$ . The power against  $\theta = 0$  for the same sample sizes and specifications of the equivalence margin can be read from Table 5.14.

The major differences between the one- and the two-sided versions of the paired *t*-test for equivalence become apparent when the entries in the last

Table 5.13 *Critical constant  $t_{n-1, .95}(-\sqrt{n}\varepsilon)$  of the one-sample t-test for noninferiority at level  $\alpha = 5\%$  for  $\varepsilon = .10(.10).50$  and  $n = 10(10)100$ .*

$n$	$\varepsilon =$				
	.10	.20	.30	.40	.50
10	1.45767	1.09339	0.74018	0.39776	0.06570
20	1.24639	0.77315	0.30929	-0.14545	-0.59145
30	1.12433	0.55857	0.00170	-0.54652	-1.08647
40	1.03000	0.38394	-0.25346	-0.88244	-1.50337
50	0.95008	0.23227	-0.47699	-1.17798	-1.87104
60	0.87932	0.09613	-0.67861	-1.44515	-2.20387
70	0.81507	-0.02856	-0.86382	-1.69094	-2.51030
80	0.75577	-0.14435	-1.03613	-1.91983	-2.79581
90	0.70039	-0.25293	-1.19794	-2.13492	-3.06421
100	0.64823	-0.35552	-1.35099	-2.33845	-3.31826

Table 5.14 *Power against the null alternative  $\theta = 0$  attained in the paired t-test for noninferiority for the sample sizes and equivalence-margin specifications covered by Table 5.13.*

$n$	$\varepsilon =$				
	.10	.20	.30	.40	.50
10	0.08946	0.15131	0.23903	0.35004	0.47453
20	0.11388	0.22447	0.38023	0.55706	0.71940
30	0.13505	0.29037	0.49933	0.70556	0.85689
40	0.15468	0.35156	0.59938	0.80853	0.92960
50	0.17337	0.40865	0.68226	0.87775	0.96634
60	0.19140	0.46187	0.74998	0.92315	0.98428
70	0.20892	0.51135	0.80466	0.95232	0.99279
80	0.22602	0.55720	0.84835	0.97076	0.99675
90	0.24276	0.59955	0.88294	0.98224	0.99856
100	0.25917	0.63852	0.91011	0.98931	0.99937

column of Table 5.14 are compared to those in the middle column of Table 5.12. Under the same choice of the left-hand endpoint of the equivalence range specified by the alternative hypothesis, the noninferiority test is, uniformly in  $n$ , more powerful than the test for equivalence in the two-sided sense. Of course, this statement is by no means at variance with the fact, that for  $n \geq 50$ , the lower critical bound to the t-statistic is the same for both tests

(except for negligible differences less than  $.5 \times 10^{-6}$  in absolute value): Due to the existence of a finite right-hand critical bound, the rejection region of the equivalence test is necessarily a proper subset of that of the test for noninferiority.

---

## 5.4 Signed rank test for equivalence

The testing procedure to be derived in the present section is a natural nonparametric competitor to the paired  $t$ -test for equivalence as discussed in § 5.3. It will be obtained by modifying the most frequently applied nonparametric test for paired data in such a way that the resulting test for equivalence is both interpretable in terms of a sensible measure of distance between distribution functions on  $\mathbb{R}$  and fairly easy to implement in practice. The assumptions which the construction to be described on the pages to follow starts from, are as weak as those underlying the sign test for equivalence with continuously distributed intraindividual differences  $D_i$ . This means that, in contrast to the “classical” signed rank test of the hypothesis of symmetry of the distribution of the  $D_i$ , the distribution function  $F$  of the  $D_i$  is not assumed symmetric, neither under the null nor the alternative hypothesis. Basically, this possibility of enlarging the class of admissible distributions is a consequence of the fact that the procedure we propose to term signed rank test for equivalence, is asymptotic in nature. However, we will be able to show by means of the results of suitable simulation experiments that under surprisingly mild restrictions on the order of magnitude of the sample size  $n$ , the discrepancies between the nominal level of significance and the actual size of the test are practically negligible even if the distribution of the  $D_i$  exhibits gross asymmetry.

As is well known from textbooks on nonparametric statistics (see, e.g., Lehmann, 1975, p. 129, (3.16)) for continuously distributed  $D_i$ , the test statistic  $V_s$ , say, of the ordinary signed rank test admits the representation

$$V_s = \sum_{i=1}^n I_{(0,\infty)}(D_i) + \sum_{i=1}^{n-1} \sum_{j=i+1}^n I_{(0,\infty)}(D_i + D_j) , \quad (5.39)$$

with  $I_{(0,\infty)}(\cdot)$  as the indicator of a positive sign of any real number defined formally by

$$I_{(0,\infty)}(x) = \begin{cases} 1 & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases} . \quad (5.40)$$

From (5.39) it is obvious that the finite-sample expectation of the signed rank statistic is a weighted sum of the two parameters (functionals)  $p_+ = P[D_i > 0]$  and

$$q_+ = P[D_i + D_j > 0] . \quad (5.41)$$

On the other hand, it seems reasonable to formulate the equivalence hypothesis to be established by means of the test to be constructed in the present section, in terms of the mean of the asymptotic distribution of the customary signed rank statistic. In view of this, it is preferable to drop the first term on the right-hand side of (5.39) and base the construction on the statistic

$$U_+ = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^n I_{(0,\infty)}(D_i + D_j) \quad (5.42)$$

instead of  $V_s$ . In fact,  $U_+$  is an unbiased estimator of the probability  $q_+$  of getting a positive sign of the so-called Walsh-average within any pair of  $D$ 's and is asymptotically equivalent to  $V_s$ , in the sense that the difference between the standardized versions of both statistics vanishes in probability for  $n \rightarrow \infty$  under any specification of the true distribution of  $D_i$  (cf. Randles and Wolfe, 1979, pp. 84-5).

Therefore, as an asymptotic procedure, the conventional signed rank test is a test of a hypothesis on  $q_+$  and accordingly, we refer throughout this section to the following equivalence testing problem:

$$H : 0 < q_+ \leq q'_+ \vee q''_+ \leq q_+ < 1 \text{ versus } K : q'_+ < q_+ < q''_+ , \quad (5.43)$$

where the limits  $q'_+$ ,  $q''_+$  of the equivalence range can be chosen as arbitrary fixed numbers satisfying  $0 < q'_+ < q''_+ < 1$ . The exact finite-sample variance of  $U_+$  admits the representation

$$\text{Var}[U_+] = \binom{n}{2}^{-1} \{2(n-2)[q_{1(2,3)}^+ - q_+^2] + q_+(1-q_+)\} , \quad (5.44)$$

with

$$q_{1(2,3)}^+ = P[D_i + D_j > 0, D_i + D_k > 0] \quad (5.45)$$

(cf. Randles and Wolfe, 1979, (3.1.21)). For  $q_{1(2,3)}^+$ , there exists a natural estimator as well which exhibits the form of a  $U$ -statistic and is given by

$$\begin{aligned} \hat{q}_{1(2,3)}^+ &= \binom{n}{3}^{-1} \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{k=j+1}^n (1/3) \left[ I_{(0,\infty)}(D_i + D_j) \cdot \right. \\ &\quad I_{(0,\infty)}(D_i + D_k) + I_{(0,\infty)}(D_i + D_j) I_{(0,\infty)}(D_j + D_k) + \\ &\quad \left. I_{(0,\infty)}(D_i + D_k) I_{(0,\infty)}(D_j + D_k) \right]. \end{aligned} \quad (5.46)$$

From the general theory of  $U$ -statistic estimators, it is in particular known that  $\hat{q}_{1(2,3)}^+$  as defined by (5.46) is consistent for  $q_{1(2,3)}^+$  (see once more Randles and Wolfe, 1979, Corollary 3.2.5) which implies that the variance of  $U_+$  can in turn be estimated consistently by

$$\hat{\sigma}^2[U_+] = \binom{n}{2}^{-1} \{2(n-2)[\hat{q}_{1(2,3)}^+ - U_+^2] + U_+(1-U_+)\} . \quad (5.47)$$

Since  $(U_+ - q_+)/\sqrt{\text{Var}[U_+]}$  is asymptotically standard normal for any distribution of the  $D_i$  under which  $\text{Var}[U_+]$  does not vanish, we may apply the result presented in § 3.4 yielding the decision rule

Reject nonequivalence  $\Leftrightarrow$

$$\left| U_+ - (1/2)(q'_+ + q''_+) \right| / \hat{\sigma}[U_+] < C_{\pm R}(\alpha; q'_+, q''_+), \quad (5.48)$$

with

$$C_{\pm R}(\alpha; q'_+, q''_+) = \left\{ \begin{array}{l} \text{100}\alpha-\text{percentage point of the } \chi^2 \text{- distri-} \\ \text{bution with } df = 1 \text{ and } \lambda_{nc}^2 = (q''_+ - q'_+)^2 / 4\hat{\sigma}^2[U_+] \end{array} \right\}^{1/2}. \quad (5.49)$$

Although the quantile function of any noncentral  $\chi^2$ -distribution is predefined in well-known software packages for use in statistics, the practical implementation of the signed rank test for equivalence based on (5.48–9) entails considerably more effort than that of the conventional one- and two-sided Wilcoxon test for paired observations. The reason is that in addition to  $U_+$  or  $V_s$ , the quantity  $\hat{q}_{1(2,3)}^+$  not appearing in the “ordinary” signed rank test, is needed. Hence, at the URL associated with this book, the reader may find another SAS macro [program name: **sgnrk**] designed for carrying out all computational steps to be taken in an application of (5.48). As input data, it requires in addition to the size  $n$  of the sample, the nominal level  $\alpha$  of significance and the limits  $q'_+, q''_+$  of the equivalence range for the target parameter  $q_+$ , only the name of a raw-data file containing just the observed values of the  $D_i$ .

#### *Example 5.4*

In a pilot study preceding a large-scale multicenter trial of the antihypertensive efficacy of various drugs, the aim was to establish comparability of the measurements of blood pressure (BP) taken with two different automatic devices (denoted by  $A$  and  $B$  in the sequel). In each of a total of 20 volunteers participating in the pilot study, 12 repeated BP measurements were taken during the same session, 6 using each of the two devices. The temporal spacing of all individual measurements was strictly uniform, with an interval of 2 minutes between consecutive readings and altering between  $A$  and  $B$ . The 6 values produced by each device within one such series were collapsed to a single one by simple averaging.

For the analysis of the data of this pilot study, the comparison of both measurement devices with respect to the diastolic BP (DBP) was of primary interest. Correspondingly, Table 5.15 shows for each subject the averages of 6 DBP values obtained by the first device ( $A$ ) as compared to  $B$ . The limits of the equivalence range for  $q_+$  were set equal to  $q'_+ = .2398$ ,  $q''_+ = .7602$ ,

Table 5.15 *Results of a pilot study on the comparability of two different devices A ( $\leftrightarrow X_i$ ) and B ( $\leftrightarrow Y_i$ ) for measuring diastolic blood pressure.*

$i$	$X_i$	$Y_i$	$D_i$	$i$	$X_i$	$Y_i$	$D_i$
1	62.167	62.667	-0.500	11	74.500	76.667	-2.167
2	85.667	85.333	0.333	12	91.667	93.500	-1.833
3	80.667	80.000	0.667	13	73.667	69.167	4.500
4	55.167	53.833	1.333	14	63.833	71.333	-7.500
5	92.000	93.500	1.500	15	80.333	77.667	2.667
6	91.000	93.000	-2.000	16	61.167	57.833	3.333
7	107.833	108.833	-1.000	17	63.167	67.333	-4.167
8	93.667	93.833	-0.167	18	73.167	67.500	5.667
9	101.167	100.000	1.667	19	103.333	101.000	2.333
10	80.500	79.667	0.833	20	87.333	89.833	-2.500

which can be motivated as follows: In the specific case that the  $D_i$  are normal with expectation  $\delta$  and variance  $\sigma_D^2$ , it is easily verified that there holds the relationship  $q_+ = \Phi(\sqrt{2}\delta/\sigma_D)$ , and this suggests to transform the interval  $(-.5, .5)$  making up a reasonable equivalence range in terms of  $\delta/\sigma_D$  [recall Tab. 1.1 (iv)], to the set  $\{q_+ \mid \Phi(-\sqrt{2}\delta/\sigma_D) < q_+ < \Phi(\sqrt{2}\delta/\sigma_D)\} = (.2398, .7602)$ .

Now, in order to test for equivalence of the measuring devices  $A$  and  $B$ , we apply the above decision rule (5.48) with these value of  $q'_+$  and  $q''_+$ ,  $\alpha = .05$ , and the  $n = 20$  intra-subject differences shown in the above table as the observations eventually to be analyzed. Running the program `sgnrk` we find  $U_+ = .55263$ ,  $\hat{\sigma}[U_+] = .12071$ ,  $C_{\pm R}(.05; .2398, .7602) = 0.54351$ , and it follows that  $|U_+ - (1/2)(q'_+ + q''_+)|/\hat{\sigma}[U_+] = |.55263 - .50000|/.12071 = .43600 < C_{\pm R}(.05; .2398, .7602)$ . Hence, we can reject the null hypothesis of nonequivalence of the two measuring devices under comparison.

For the remaining part of this section, we restrict consideration to the well-known semiparametric shift model assuming that the density and cumulative distribution function of the intraindividual differences  $D_i$  is given by  $z \mapsto f_\circ(z - \vartheta)$  and  $z \mapsto F_\circ(z - \vartheta)$  with fixed (though unspecified) baseline  $f_\circ : \mathbb{R} \rightarrow [0, \infty)$  and  $F_\circ : \mathbb{R} \rightarrow [0, 1]$ , respectively. Even under this comparatively simple submodel, exact computation of rejection probabilities of the test (5.48) is unfeasible so that for purposes of investigating the finite-sample power function of the signed rank test for equivalence we have recourse to simulation methods. A question of particular interest to be answered by this way concerns the possible dependence of the power against some specific alternative in terms of the target functional  $q_+$ , on the form of the baseline distribution function  $F_\circ$ .

Let us denote by  $\beta(F_\circ; q_+)$  the probability that the test rejects its null hypothesis if it is performed with data  $D_1, \dots, D_n$  from a distribution which

belongs to the location parameter family generated by baseline cdf  $F_\circ$  such that the true value of  $P[D_i + D_j > 0]$  is  $q_+$ . Clearly, in a simulation experiment carried out in order to determine  $\beta(F_\circ; q_+)$ , we have to generate sets of values of random variables of the form  $D_i = D_i^\circ + \vartheta(q_+)$  where  $D_i^\circ$  has distribution function  $F_\circ$  for all  $i = 1, \dots, n$ , and  $\vartheta(q_+)$  is the unique solution of the equation

$$\int_{-\infty}^{\infty} [1 - F_\circ(-2\vartheta - z)] dF_\circ(z) = q_+, \quad \vartheta \in \mathbb{R}. \quad (5.50)$$

Given some specific baseline cdf  $F_\circ$ , the expression on the left-hand side of (5.50) is a function, say  $w_+(\vartheta)$ , of the location parameter which admits an explicit representation only for some of the most common location families. Table 5.16 gives a summary of the pertinent formulae. Making use of these relationships, for the models covered by the table, the solution  $\vartheta(q_+)$  to (5.50) can either be written down explicitly (Gaussian, Cauchy, uniform, exponential distribution) or computed numerically by means of a simple interval halving algorithm (Laplace, logistic distribution). Furthermore, it follows immediately from a general result of advanced mathematical analysis (see Pratt, 1960, Corollary 3) that  $w_+(\vartheta)$  is a continuous function of  $\vartheta$  in each location parameter family generated by a continuously differentiable baseline cdf  $F_\circ(\cdot)$ . Hence, to each specification of the equivalence range  $(q'_+, q''_+)$  for the target functional  $q_+$ , there corresponds a unique interval  $(\vartheta_1, \vartheta_2)$ , say, of values of

Table 5.16 *Explicit representation of  $w_+(\vartheta) = P[(D_i^\circ + \vartheta) + (D_j^\circ + \vartheta) > 0]$  for independent random variables  $D_i^\circ, D_j^\circ$  with common density  $f_\circ(\cdot)$ .*

Location Family	$f_\circ(z)$	$w_+(\vartheta) = \int_{-\infty}^{\infty} [1 - F_\circ(-2\vartheta - z)] dF_\circ(z)$
Gaussian	$(2\pi)^{-1/2} e^{-z^2/2}$	$\Phi(\sqrt{2}\vartheta)$
Cauchy	$\pi^{-1}(1+z^2)^{-1}$	$1/2 + \pi^{-1} \arctan(\vartheta)$
Uniform*	$I_{(-1/2, 1/2)}(z)$	$1/2 + 2\vartheta(1-\vartheta)$ for $0 \leq \vartheta < 1/2$ $1$ for $\vartheta \geq 1/2$
Exponential†	$e^{-z} I_{(0, \infty)}(z)$	$1 - G_{4;0}((-4\vartheta) \vee 0)$
Laplace*	$e^{- z-\vartheta }$	$1 - (1/2)e^{-2\vartheta}(1+\vartheta)$ for $\vartheta \geq 0$
Logistic	$e^{-z}(1+e^{-z})^{-2}$	$e^{2\vartheta}(e^{2\vartheta}-1-2\vartheta)(e^{2\vartheta}-1)^{-2}$

\* For  $\vartheta < 0$ , the relationship  $w_+(-\vartheta) = 1 - w_+(\vartheta)$  can be made use of being valid for arbitrary location families with symmetric baseline density  $f_\circ(\cdot)$ .

† In the formula given for  $w_+(\vartheta)$ ,  $G_{4;0}(\cdot)$  is to denote the cdf of a central  $\chi^2$ -distribution with  $df = 4$ . Furthermore,  $(-4\vartheta) \vee 0$  stands for  $\max\{-4\vartheta, 0\}$ .

the parameter indexing the specific location model in mind. In Table 5.17, one finds the thus associated equivalence interval for  $\vartheta$  for both specifications of  $(q'_+, q''_+)$  used in the simulation study and all location families appearing in Table 5.16.

The simulation study whose results are presented below in Table 5.18, was performed to provide sufficiently complete data both on the level properties and the maximum power of the signed rank test for equivalence. Therefore, for each specification of  $F_\circ$  and  $(\vartheta_1, \vartheta_2)$  appearing in Table 5.17, 100,000 samples of the respective size  $n \in \{20, 30, 50\}$  were drawn from the distributions given by  $F_\circ(\cdot - \vartheta_1)$ ,  $F_\circ(\cdot - \vartheta_2)$  and  $F_\circ(\cdot - \vartheta_\circ)$  where  $\vartheta_\circ$  denotes that point in the parameter space which satisfies  $w_+(\vartheta_\circ) = 1/2$ . The case that  $\vartheta$  coincides with

Table 5.17 *Equivalence range for  $\vartheta$  corresponding to  $(q'_+, q''_+) = (.3618, .6382)$  and  $(q'_+, q''_+) = (.2398, .7602)$ , respectively, in the location parameter families covered by Table 5.16.*

Family of distribution	Equivalence Range $(\vartheta_1, \vartheta_2)$ corresponding to $(q'_+, q''_+) =$	
	$(0.3618, 0.6382)$	$(0.2398, 0.7602)$
Gaussian	$(-0.2500, 0.2500)$	$(-0.5000, 0.5000)$
Cauchy	$(-0.4637, 0.4637)$	$(-1.0662, 1.0662)$
Uniform	$(-0.0747, 0.0747)$	$(-0.1537, 0.1537)$
Exponential	$(-1.0853, -0.6340)$	$(-1.3748, -0.4668)$
Laplace	$(-0.2885, 0.2885)$	$(-0.6035, 0.6035)$
Logistic	$(-0.4245, 0.4245)$	$(-0.8563, 0.8563)$

$\vartheta_\circ$  or, equivalently,  $q_+$  with  $1/2$ , occurs in particular if for each  $i = 1, \dots, n$ , the random pair  $(X_i, Y_i)$  giving rise to the intraindividual difference  $D_i$  is continuous and exchangeable in the sense of having the same distribution as  $(Y_i, X_i)$ . Of course, exchangeability is a natural assumption whenever both treatments under comparison are actually identical and assessed by means of an endpoint criterion admitting numerically precise measurement. To be sure, exchangeability of a continuously distributed random pair  $(X_i, Y_i)$  implies still more than the validity of  $P[D_i + D_j > 0] = 1/2$ , namely symmetry about zero of the distribution of the associated intrapair difference  $D_i = X_i - Y_i$ .

Going through the rejection probabilities shown in Table 5.18, one notices in particular (see the entries in line 4) that in the setting of the above Example 5.4 with normally distributed intra-subject differences, the nominal level of significance is strictly maintained. Furthermore, it can be seen that the power attained against the alternative of no difference between both measuring devices comes out as low as 37.5%. On the one hand, this value is exceeded by that of the optimal parametric test for the same setting by about 8%.

Table 5.18 *Simulated rejection probability of the signed rank test for equivalence at level  $\alpha = .05$  at both boundaries of the null hypothesis (columns 4 and 5) and at  $q_+ = 1/2$  (rightmost column) for the distributional shapes and equivalence ranges covered by Table 5.17. [Italicized values: power of the paired t-test for equivalence against the alternative  $q_+ = 1/2$ .]*

Location Family	$(q'_+, q''_+)$	$n$	$\beta(F_\circ; q'_+)$	$\beta(F_\circ; q''_+)$	$\beta(F_\circ; 1/2)$
Gaussian	(.3618, .6382)	20	.04398	.04299	.07982 <i>.09313</i>
	"	30	.04869	.04912	.11893 <i>.12662</i>
	"	50	.04982	.04895	.20635 <i>.22821</i>
	(.2398, .7602)	20	.04287	.04145	.37516 <i>.45323</i>
	"	30	.04133	.03977	.63069 <i>.71411</i>
	"	50	.03785	.03858	.89987 <i>.93268</i>
	Cauchy	(.3618, .6382)	20	.04441	.04460 <i>.07785</i>
	"	30	.05150	.05041	.11613
	"	50	.04952	.04970	.20805
	(.2398, .7602)	20	.04578	.04647	.37366
	"	30	.04215	.04195	.63091
	"	50	.04126	.03885	.89866
Uniform	(.3618, .6382)	20	.04058	.04293	.07828
	"	30	.04713	.04750	.11698
	"	50	.04761	.04822	.21146
	(.2398, .7602)	20	.03996	.03980	.37255
	"	30	.03822	.03901	.63042
	"	50	.03871	.03804	.89656
	Exponential	(.3618, .6382)	20	.04587	.04048 <i>.07311</i>
	"	30	.05315	.04661	.11383
	"	50	.05135	.04580	.19046
	(.2398, .7602)	20	.04589	.03932	.34001
	"	30	.04129	.03656	.59280
	"	50	.03919	.03602	.87640
Laplace	(.3618, .6382)	20	.04543	.04338	.07807
	"	30	.05004	.04990	.12144
	"	50	.04990	.04850	.20773
	(.2398, .7602)	20	.04302	.04435	.37238
	"	30	.04070	.04035	.63089
	"	50	.03959	.03941	.89908
	Logistic	(.3618, .6382)	20	.04497	.04314 <i>.07831</i>
	"	30	.05021	.05028	.11682
	"	50	.05037	.04954	.20822
	(.2398, .7602)	20	.04322	.04284	.37418
	"	30	.04010	.03995	.63520
	"	50	.03900	.03829	.89718

On the other, the power of the sign test for equivalence with limits  $\Phi(\mp.5) = .5 \mp .1915$  to  $p_+$ , against  $p_+ = 1/2$  falls short of that of the signed rank test by almost 15% even if the former is performed in its randomized version.

From a more general perspective, the simulation results shown in Table 5.18 admit the following conclusions:

- (i) Even for sample sizes as small as 20 and highly skewed distributions ( $\rightarrow$  exponential location parameter family), there is no reason to suspect that the nominal significance level could be exceeded to a practically relevant extent.
- (ii) Both the effective size and the power of the signed rank test for equivalence seems rather robust even against gross changes in the form of the underlying distribution.
- (iii) If the distribution of the intra-subject differences  $D_i$  is Gaussian, the loss in efficiency of the signed rank as compared to the  $t$ -statistic seems to be considerably more marked in the equivalence case than in testing problems with conventional form of the hypotheses (cf. Randles and Wolfe, 1979, Tab. 4.1.7).

*Remark.* Except for the case that the underlying distribution is exponential, the corresponding entries in column 4 and 5 of the above Table 5.18 are strikingly similar suggesting that each such pair actually gives two repeated estimates of the same quantity. This impression is far from misleading since it is not hard to derive analytically the following general result: Whenever the equivalence limits  $q'_+$ ,  $q''_+$  are chosen symmetrically about 1/2 and the location family under consideration is generated by a baseline distribution function symmetric about 0, the power function  $q_+ \mapsto \beta(F_\circ; q_+)$  of the signed rank test for equivalence is symmetric for its own part.

#### *Signed rank test for noninferiority*

In the noninferiority testing problem associated with (5.43), the hypotheses read

$$H_1 : q_+ \leq 1/2 - \varepsilon \text{ versus } K_1 : q_+ > 1/2 - \varepsilon. \quad (5.51)$$

Due to the facts stated before in the derivation of the signed rank test for two-sided equivalence, the test statistic  $U_+$  also satisfies the conditions for carrying out a construction of the kind described in § 2.3. Setting  $T_N = U_+$ ,  $\hat{\tau}_N = \hat{\sigma}^2[U_+]$ , and  $\theta_\circ = 1/2$  yields

$$\{(U_+ - 1/2 + \varepsilon)/\hat{\sigma}[U_+] > u_{1-\alpha}\} \quad (5.52)$$

as the rejection region of an asymptotically valid test for (5.51). As regards the choice of the equivalence margin  $\varepsilon$ , we propose to adopt the values ap-

pearing in Table 5.17 also for the noninferiority case. In particular, the simulation results shown below in Table 5.19 likewise refer to the specifications

Table 5.19 *Simulated rejection probability of the signed rank test for noninferiority at level  $\alpha = .05$  at the boundary of the null hypothesis (column 4) and  $q_+ = 1/2$  (rightmost column) for the distributional shapes and equivalence margins specified in Table 5.17. [Italicized values: power of the paired t-test for noninferiority against the alternative  $q_+ = 1/2$ .]*

Location Family	$1/2 - \varepsilon$	$n$	$\beta(F_0; 1/2 - \varepsilon)$	$\beta(F_0; 1/2)$
Gaussian	.3618	20	.06171	.32401 <i>.29771</i>
	"	30	.05459	.39560 <i>.39110</i>
	"	50	.04960	.53011 <i>.54886</i>
	.2398	20	.04298	.67383 <i>.71940</i>
	"	30	.03971	.81432 <i>.85689</i>
	"	50	.03991	.95033 <i>.96634</i>
	.3618	20	.06195	.32191
	"	30	.05484	.39325
	"	50	.05126	.53232
	.2398	20	.04515	.67292
Cauchy	"	30	.04050	.81476
	"	50	.03966	.95052
	.3618	20	.05978	.32320
	"	30	.05271	.39639
	"	50	.04893	.53045
	.2398	20	.04047	.67455
Uniform	"	30	.03924	.81512
	"	50	.03756	.94969
	.3618	20	.06413	.32651
	"	30	.05653	.39117
	"	50	.05242	.51927
	.2398	20	.04461	.65234
Exponential	"	30	.04207	.78600
	"	50	.03921	.92933
	.3618	20	.05962	.32186
	"	30	.05471	.39417
	"	50	.05013	.53298
	.2398	20	.04390	.67318
Laplace	"	30	.04105	.81568
	"	50	.03812	.94834
	.3618	20	.06068	.32253
	"	30	.05468	.39497
	"	50	.04885	.53138
	.2398	20	.04342	.67443
Logistic	"	30	.04092	.81620
	"	50	.03872	.94965

$1/2 - \varepsilon = .3618$  and  $1/2 - \varepsilon = .2398$  which under the location-shift models studied here transform to lower bounds for  $\vartheta$  in the way made precise in Table 5.17.

As to the conclusions to be drawn from the data shown in the above table, the only difference compared with the test for two-sided equivalence worth mentioning refers to a tendency toward a less strict control of the target significance level due to limitations in accuracy of the large-sample approximation involved. In all parametric submodels investigated, with samples of size  $n = 20$  and noninferiority margin  $\varepsilon = .1382$ , the size of the critical region of the signed-rank test for noninferiority is about 6 rather than 5 percent. On the one hand, this deviation from the target level seems still tolerable for most practical purposes. On the other, under the Gaussian model, it increases the power against the alternative  $q_+ = 1/2 \Leftrightarrow \delta = 0$  slightly above that of the paired  $t$ -test for noninferiority although the latter is UMP among all tests being invariant against transformations under which the test based on the signed rank statistic remains invariant as well. Through reducing the size of the latter to its target value, the sign of this difference in power would be reversed, leading to similar patterns as were obtained in the two-sided case.

---

## 5.5 A generalization of the signed rank test for equivalence for noncontinuous data

In the previous section, the construction of an equivalence version of the signed rank test was carried out under the basic assumption that the distribution of the intraindividual differences  $D_1, \dots, D_n$  is of the continuous type. For a considerable proportion of possible applications of the procedure to concrete data sets, this restriction is prohibitive since in practice the occurrence of ties between the  $D_i$  is quite common. Lack of continuity of the distribution of the  $D_i$  implies in particular that the probability

$$q_0 = P[D_i + D_j = 0] \quad (5.53)$$

of observing a pair of intra-subject differences of equal absolute value but opposite sign, does not necessarily vanish but might take on any value in the unit interval. Obviously, for  $q_0 > 0$ , it no longer makes sense to compare the true value of the functional  $q_+$  to  $1/2$  and consider  $|q_+ - 1/2|$  as a reasonable measure of distance between the actual distribution of the  $D_i$  and a distribution exhibiting the same form but being centered about 0. Hence, what we primarily need for purposes of adapting the signed rank test for equivalence to settings with noncontinuous data, is a meaningful generalization of the basic distance measure  $|q_+ - 1/2|$ .

In the continuous case, we obviously have  $q_+ = 1 - q_-$  with

$$q_- = P[D_i + D_j < 0] \quad (5.54)$$

so that  $|q_+ - 1/2| = 0$  is satisfied if and only if there holds  $q_+ = q_-$ . Irrespective of the type of the distributions involved, marked discrepancies between  $q_+$  and  $q_-$  are always at variance with equivalence of the treatments behind the primary observations  $X_i$  and  $Y_i$  from which the  $i$ th intra-subject difference is computed. In fact, even if the joint distribution of  $(X_i, Y_i)$  is discrete, it must be exchangeable whenever there exists no treatment difference at all, and exchangeability of  $(X_i, Y_i)$  clearly implies  $D_i = X_i - Y_i \stackrel{d}{=} Y_i - X_i = -D_i$  and hence  $q_+ = P[(-D_i) + (-D_j) > 0] = q_-$ . Now, in view of the basic restriction  $q_+ + q_0 + q_- = 1$ ,  $q_+$  and  $q_-$  coincide if and only if there holds the equation  $2q_+ + q_0 = 1$ , or equivalently,  $q_+/(1 - q_0) = 1/2$ . The latter fact suggests to replace  $q_+$  by its conditional counterpart  $q_+/(1 - q_0)$  throughout, where conditioning is on absence of any pair  $(i, j)$  of sampling units such that the Walsh average  $(D_i + D_j)/2$  of the associated intra-subject differences vanishes. This leads to use  $|q_+/(1 - q_0) - 1/2|$  as a basic measure of distance suitable for any type of distribution of the  $D_i$ . Defining, as usual, equivalence by the requirement that the value taken on by the selected measure of distance between the distributions under comparison be sufficiently small, the nonparametric testing problem to be dealt with in this section reads

$$H : \frac{q_+}{1 - q_0} \leq 1/2 - \varepsilon_1 \text{ or } \frac{q_+}{1 - q_0} \geq 1/2 + \varepsilon_2$$

versus     $K : 1/2 - \varepsilon_1 < \frac{q_+}{1 - q_0} < 1/2 + \varepsilon_2 . \quad (5.55)$

As before, the construction of an asymptotically valid testing procedure will be based on a natural estimator of the target functional in terms of which the hypotheses have been formulated. This estimator is given by

$$U_+^* = U_+/(1 - U_0) \quad (5.56)$$

where  $U_+$  is as in (5.42) and  $U_0$  denotes the  $U$ -statistic estimator of  $q_0$  to be computed analogously, namely as

$$U_0 = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^n I_{\{0\}}(D_i + D_j) \quad (5.57)$$

with

$$I_{\{0\}}(z) = \begin{cases} 1 & \text{for } z = 0 \\ 0 & \text{for } z \in \mathbb{R} \setminus \{0\} \end{cases} . \quad (5.58)$$

In order to determine the large-sample distribution of  $U_+^*$  and in particular its asymptotic variance, we first consider the joint large-sample distribution

of  $(U_+, U_0)$ . From the general asymptotic distribution theory for  $U$ -statistics (see, e.g., Randles and Wolfe, 1979, Sec. 3.6), it follows that  $\sqrt{n}(U_+ - q_+, U_0 - q_0)$  is asymptotically bivariate normal with expectation  $\mathbf{0}$ . The covariance matrix  $\Sigma_n = \begin{pmatrix} \sigma_{+n}^2 & \sigma_{0+;n} \\ \sigma_{+0;n} & \sigma_{0n}^2 \end{pmatrix}$ , say, of this distribution can be computed exactly, and with a view to using the testing procedure under derivation in finite samples, we prefer to keep the terms of order  $O(1/n)$  in the following formulae:

$$\sigma_{+n}^2 = \frac{4(n-2)}{n-1} \left( q_{1(2,3)}^+ - q_+^2 \right) + \frac{2}{n-1} q_+ (1-q_+), \quad (5.59a)$$

$$\sigma_{0n}^2 = \frac{4(n-2)}{n-1} \left( q_{1(2,3)}^0 - q_0^2 \right) + \frac{2}{n-1} q_0 (1-q_0), \quad (5.59b)$$

$$\sigma_{+0;n} = \frac{4(n-2)}{n-1} \left( q_{1(2,3)}^{+0} - q_+ q_0 \right) + \frac{2}{n-1} q_+ q_0. \quad (5.59c)$$

In these expressions, the symbol  $q_{1(2,3)}^+$  has the same meaning as before [recall (5.45)] whereas  $q_{1(2,3)}^0$  and  $q_{1(2,3)}^{+0}$  have to be defined by

$$q_{1(2,3)}^0 = P[D_i + D_j = 0, D_i + D_k = 0] \quad (5.60)$$

and

$$q_{1(2,3)}^{+0} = P[D_i + D_j > 0, D_i + D_k = 0], \quad (5.61)$$

respectively. Now, we can proceed by applying the so-called  $\delta$ -method (cf. Bishop et al., 1975, § 14.6) which allows us to infer that  $\sqrt{n}(U_+^* - q_+/(1-q_0))$  is asymptotically univariate normal with mean 0 and variance  $\sigma_{*n}^2 = \nabla g(q_+, q_0)' \Sigma_n (\nabla g(q_+, q_0))'$  where  $\nabla g(q_+, q_0)$  stands for the gradient (row) vector of the transformation  $(q_+, q_0) \mapsto q_+/(1-q_0)$ . Writing down  $\nabla g(q_+, q_0)$  explicitly and expanding the quadratic form yields, after some straightforward algebraic simplifications, the expression:

$$\sigma_{*n}^2 = \frac{\sigma_{+n}^2}{(1-q_0)^2} + \frac{q_+^2 \sigma_{0n}^2}{(1-q_0)^4} + \frac{2q_+ \sigma_{+0;n}}{(1-q_0)^3}. \quad (5.62)$$

The natural way of estimating the asymptotic variance  $\sigma_{*n}^2$  of  $\sqrt{n}U_+^*$  consists in plugging in  $U$ -statistic estimators of all functionals of the distribution of the  $D_i$  appearing on the right-hand side of equations (5.59a–c). Since the  $U$ -statistics for  $q_+$ ,  $q_{1(2,3)}^+$  and  $q_0$  have already been defined [recall (5.42), (5.46) and (5.57), respectively], all what remains to be done is to provide explicit formulae for the analogous estimators of  $q_{1(2,3)}^0$  and  $q_{1(2,3)}^{+0}$ :

$$\begin{aligned} \hat{q}_{1(2,3)}^0 &= {n \choose 3}^{-1} \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{k=j+1}^n (1/3) \left[ I_{\{0\}}(D_i + D_j) I_{\{0\}}(D_i + D_k) + \right. \\ &\quad \left. I_{\{0\}}(D_i + D_j) I_{\{0\}}(D_j + D_k) + I_{\{0\}}(D_i + D_k) I_{\{0\}}(D_j + D_k) \right], \end{aligned} \quad (5.63)$$

$$\begin{aligned} q_{1(2,3)}^{+0} = {}^n \binom{n}{3}^{-1} \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{k=j+1}^n & (1/6) \left[ I_{(0,\infty)}(D_i + D_j) I_{\{0\}}(D_i + D_k) + \right. \\ & I_{(0,\infty)}(D_i + D_j) I_{\{0\}}(D_j + D_k) + I_{(0,\infty)}(D_i + D_k) I_{\{0\}}(D_j + D_k) \\ & + I_{(0,\infty)}(D_i + D_k) I_{\{0\}}(D_i + D_j) + I_{(0,\infty)}(D_j + D_k) I_{\{0\}}(D_i + D_j) \\ & \left. + I_{(0,\infty)}(D_j + D_k) I_{\{0\}}(D_i + D_k) \right]. \quad (5.64) \end{aligned}$$

Additional results from the asymptotic theory of  $U$ -statistics (see, e.g., Lee, 1990, § 3.4.2) ensure the (strong) consistency of all these estimators for their expectations. This implies that the plug-in estimator  $\hat{\sigma}_{+n}^2$  obtained from (5.59a) by replacing  $q_+$  and  $q_{1(2,3)}^+$  with  $U_+$  and  $\hat{q}_{1(2,3)}^+$ , respectively, is consistent for  $\sigma_{+n}^2$ , and so on for  $\hat{\sigma}_{0n}^2$  and  $\hat{\sigma}_{+0;n}$ . Finally, consistency of  $U_+$ ,  $U_0$ ,  $\hat{\sigma}_{+n}^2$ ,  $\hat{\sigma}_{0n}^2$  and  $\hat{\sigma}_{+0;n}$  ensures that the asymptotic variance  $\sigma_{*n}^2$  of  $\sqrt{n}U_+^* = \sqrt{n}U_+/(1 - U_0)$  can be consistently estimated by

$$\hat{\sigma}_{*n}^2 = \frac{\hat{\sigma}_{+n}^2}{(1 - U_0)^2} + \frac{U_+^2 \hat{\sigma}_{0n}^2}{(1 - U_0)^4} + \frac{2U_+ \hat{\sigma}_{+0;n}}{(1 - U_0)^3}. \quad (5.65)$$

Now, in order to complete the derivation of a signed rank test for equivalence being asymptotically valid for arbitrary noncontinuous distributions of the  $D_i$ , it suffices to invoke for another time the general result stated in § 3.4, with the following specifications:  $N = n$ ,  $T_N = U_+/(1 - U_0)$ ,  $\theta = q_+/(1 - q_0)$ ,  $\sigma = \lim_{n \rightarrow \infty} \sigma_{*n}$ ,  $k = 1$ , and  $\tau_N = \sigma_{*n}/\sqrt{n}$ . This verifies that the decision rule

$$\text{Reject } H : \frac{q_+}{1 - q_0} \leq 1/2 - \varepsilon_1 \text{ or } \frac{q_+}{1 - q_0} \geq 1/2 + \varepsilon_2 \quad \text{iff} \\ \sqrt{n} \left| U_+/(1 - U_0) - (1 - \varepsilon_1 + \varepsilon_2)/2 \right| / \hat{\sigma}_{*n} < C_{\pm R}^*(\alpha; \varepsilon_1, \varepsilon_2), \quad (5.66)$$

with

$$C_{\pm R}^*(\alpha; \varepsilon_1, \varepsilon_2) = \left\{ 100\alpha\text{-percentage point of the } \chi^2 \text{ - distribution with } df = 1 \text{ and } \lambda_{nc}^2 = n(\varepsilon_1 + \varepsilon_2)^2 / 4\hat{\sigma}_{*n}^2 \right\}^{1/2} \quad (5.67)$$

defines an asymptotically valid test for the nonparametric equivalence problem (5.55), provided the limit  $\sigma_*^2$ , say, of (5.62) is a positive number.

For the implementation of the generalized signed rank test for equivalence allowing for arbitrary patterns of ties in the set  $\{D_1, \dots, D_n\}$  of observed intra-subject differences, another special program is available for download from **WKTSHEQ2 Source Code Package**. The program name is **srktie\_d**, and again, it exists both in a SAS and a R version.

*Simplifying computations by reducing raw data to counts*

Often computation of the  $U$ -statistics required for carrying out the generalized signed rank test for equivalence can largely be simplified by determining in a preliminary step the configuration of ties which occurred in the observed vector  $(D_1, \dots, D_n)$  of intra-subject differences. This is the case whenever the  $D_i$  take on their values in a finite lattice of known span  $w > 0$ , say. Without loss of generality, this lattice can be assumed symmetric about 0, i.e., as a set of the form  $\{ -rw, -(r-1)w, \dots, -w, 0, w, \dots, (r-1)w, rw \}$  where  $r \in \mathbb{N}$  denotes the number of points lying on both sides of the origin. Of course, a particularly frequent specific case of that kind occurs when the  $D_i$  are integer-valued random variables with finite support. Given a lattice containing the set of possible values of each  $D_i$  in the sample of intra-subject differences, let us define

$$M_k = \# \{ i \mid D_i = kw \}, \quad k = -r, \dots, r \quad (5.68)$$

so that for any  $k \in \{-r, \dots, r\}$ ,  $M_k$  counts the number of sample points tied at the respective point of the lattice. (Note that some of the frequencies  $M_k$  may and typically will be equal to zero.)

In terms of the frequencies  $M_k$ , the formulae for the  $U$ -statistic estimators to be considered in connection with the generalized signed rank test for equivalence can be rewritten as follows:

$$U_+ = \frac{1}{n(n-1)} \left[ 2 \sum_{k=1}^r \sum_{l=-k+1}^{k-1} M_k M_l + \sum_{k=1}^r M_k^2 - \sum_{k=1}^r M_k \right], \quad (5.42^*)$$

$$\begin{aligned} q_{1(2,3)}^+ &= \frac{1}{n(n-1)(n-2)} \left[ \sum_{k=-r+1}^r M_k \left( \sum_{l=-k+1}^r M_l \right)^2 - \sum_{k=1}^r M_k^2 \right. \\ &\quad \left. - 2 \sum_{k=1}^r \sum_{l=-k+1}^{k-1} M_k M_l + 2 \sum_{k=1}^r M_k - 2 \sum_{k=1}^r \sum_{l=-k+1}^r M_k M_l \right], \quad (5.46^*) \end{aligned}$$

$$U_0 = \frac{1}{n(n-1)} \left[ 2 \sum_{k=1}^r M_k M_{-k} + M_0(M_0 - 1) \right], \quad (5.57^*)$$

$$\begin{aligned} q_{1(2,3)}^0 &= \frac{1}{n(n-1)(n-2)} \left[ \sum_{k=-r}^r M_k M_{-k}^2 - 2 \sum_{k=1}^r M_k M_{-k} \right. \\ &\quad \left. - 3M_0^2 + 2M_0 \right], \quad (5.63^*) \end{aligned}$$

$$\begin{aligned} \hat{q}_{1(2,3)}^{+0} &= \frac{1}{n(n-1)(n-2)} \left[ \sum_{k=-r+1}^r \left( M_k M_{-k} \sum_{l=-k+1}^r M_l \right) \right. \\ &\quad \left. - \sum_{k=1}^r M_k (M_{-k} + M_0) \right]. \quad (5.64^*) \end{aligned}$$

These identities (formally proved in Firle, 1998, pp. 86–88) underlie the program supplied under the name `srktie_m` [→ Appendix B] which allows its user to compute both the test statistic and its critical upper bound at any desired significance level  $\alpha$ . Of course, in every setting such that the  $D_i$  are lattice variables, the function `srktie_d` can be used instead, and the results will necessarily coincide with those obtained by means of `srktie_m`. The only difference is that the algorithm underlying the M-version of the program enables to dispense with triple loops and typically needs much shorter execution time.

### *Example 5.5*

In a comparative trial of the efficacy of two mydriatic agents (compounds administered for the purpose of dilating the eye pupils)  $A$  and  $B$ , each of  $n = 24$  patients got dropped substance  $A$  to the one eye and  $B$  to the other. The outcome was measured as the increase of the pupil diameter [mm] attained after 30 minutes since administration of the tinctures. The individual results obtained in this way are shown in Table 5.20. Due to limited accuracy of measurements, all  $X_i$  and  $Y_i$  observed in this trial are multiples of  $w = 0.1$ , and so are the associated intra-subject differences  $D_i = X_i - Y_i$ . In order to test for equivalence in efficacy of both mydriatic drugs, decision rule (5.66) is clearly appropriate. Using the same equivalence limits as have been proposed in Example 5.4 for  $q_+$ , i.e., setting  $\varepsilon_1 = \varepsilon_2 = .2602$  and  $\alpha = .05$ , the procedure `srktie_m` yields with the entries in column 4 of the table as raw data the following results:  $U_+^* = .57769$ ,  $\hat{\sigma}_{*n} = .59819$ ,  $C_{\pm R}^*(\alpha; \varepsilon_1, \varepsilon_2) = .52345$ .

Table 5.20 *Results of a comparative trial of two mydriatic agents, with  $X_i$  and  $Y_i$  as the increase of the pupil diameter [mm] attained in the  $i$ th patient 30 minutes after administration of drug A and B, respectively.*

$i$	$X_i$	$Y_i$	$D_i$	$i$	$X_i$	$Y_i$	$D_i$
1	2.4	1.6	0.8	13	3.8	4.1	-0.3
2	2.6	2.4	0.2	14	4.2	4.2	0.0
3	3.3	3.3	0.0	15	3.5	3.4	0.1
4	3.9	4.0	-0.1	16	3.6	3.3	0.3
5	3.6	3.9	-0.3	17	4.0	4.3	-0.3
6	2.7	2.4	0.3	18	3.6	3.5	0.1
7	4.4	4.5	-0.1	19	3.8	4.0	-0.2
8	3.7	3.3	0.4	20	3.3	3.8	-0.5
9	3.6	3.0	0.6	21	2.7	2.5	0.2
10	3.4	3.2	0.2	22	4.0	4.1	-0.1
11	2.6	2.6	0.0	23	3.7	3.5	0.2
12	3.9	4.1	-0.2	24	3.3	3.4	-0.1

But with  $n = 24$  and  $\varepsilon_1 = \varepsilon_2 = .2602$ , these values do not lead to a rejection of the null hypothesis  $H : q_+/(1 - q_0) \leq 1/2 - \varepsilon_1$  or  $q_+/(1 - q_0) \geq 1/2 + \varepsilon_2$  since we have  $\sqrt{24}|.57769 - 1/2|/.59819 = .63626 > C_{\pm R}^*(\alpha; \varepsilon_1, \varepsilon_2)$ .

Table 5.21 is to give an impression of the multiformity of discrete distributions concentrated on the specific lattice  $\{-2, -1, 0, 1, 2\}$  leading to the same selected value of the target functional  $q_+/(1 - q_0)$ . From the entries in lines (b), (e) and (f), it becomes in particular obvious that symmetry of the distribution of the  $D_i$  is sufficient but by no means necessary for minimizing the distance measure  $|q_+/(1 - q_0) - 1/2|$  underlying hypotheses formulation (5.55).

Table 5.21 *Examples of distributions on a five-point lattice belonging to the left [(a),(d)], the right [(c),(g)] boundary and the center [(b), (e), (f)] of the equivalence hypothesis K of (5.55) with  $\varepsilon_1 = \varepsilon_2 = .2602$ .*

	$P[D_i = -2]$	$P[D_i = -1]$	$P[D_i = 0]$	$P[D_i = 1]$	$P[D_i = 2]$
(a)	.313300	.249480	.229700	.086820	.120700
(b)	.091300	.346251	.229700	.116049	.216700
(c)	.120700	.086820	.229700	.249480	.313300
(d)	.300000	.366789	.000000	.233211	.100000
(e)	.200000	.400000	.000000	.100000	.300000
(f)	.369855	.130145	.000000	.130145	.369855
(g)	.299300	.052384	.000000	.048016	.600300

The rejection probabilities of the generalized signed rank test for equivalence under all specific distributions covered by Table 5.21 have been determined by means of Monte Carlo simulation. The results of this simulation study which are shown in Table 5.22 allow the conclusion that correcting the equivalence version of the signed rank test for ties in the way established in this section, is another instance of an asymptotic construction along the lines of § 3.4 yielding an equivalence testing procedure which tends to mild conservatism in finite samples. Furthermore, by comparison with Table 5.18, it can be seen that the maximum power attainable with a given sample size  $n$ , is markedly lower in the discrete than in the continuous case, as to be expected according to intuition.

Table 5.22 Simulated rejection probabilities<sup>†</sup> of the tie-corrected version of the signed rank test for equivalence under the distributions shown in Table 5.21, for sample sizes varying over {20, 30, 50, 100}. [The values obtained under (a), (d) and (c), (g) allow to estimate the actual type-I error risk in finite samples; the entries into lines (b), (e) and (f) are values of the power against  $q_+/(1 - q_0) = 1/2$ .]

	$n = 20$	$n = 30$	$n = 50$	$n = 100$
(a)	.03910	.03657	.03606	.03619
(d)	.03695	.03738	.03422	.03622
(c)	.03834	.03485	.03263	.03537
(g)	.03348	.03790	.03445	.03564
(b)	.19518	.36714	.70671	.97368
(e)	.17185	.33191	.67104	.96635
(f)	.12228	.24260	.53774	.92583

<sup>†</sup> 100,000 replications of each Monte Carlo experiment

### Discussion

In order to gain better insight into the meaning of the distance measure  $|q_+/(1 - q_0) - 1/2|$  we chose as a basis for formulating a nonparametric equivalence hypothesis for settings with noncontinuous paired observations, it is helpful to analyze the extreme case that the  $D_i$  can take on only three different values, namely  $-1$ ,  $0$  and  $+1$ . Denoting the probabilities of these values by  $p_-$ ,  $p_0$  and  $p_+$ , respectively, it is easily verified that we can write  $q_+ = 2p_0p_+ + p_+^2$ ,  $q_0 = 2(1 - p_0 - p_+)p_+ + p_0^2$ . Hence, in the trinomial case, the alternative hypothesis to be established by means of the tie-corrected signed rank test for equivalence corresponds to the following subset of the parameter space:

$$\widetilde{K}_{\varepsilon_1, \varepsilon_2}^{\pm R} = \left\{ (p_0, p_+) \in (0, 1)^2 \mid p_0 + p_+ \leq 1, 1/2 - \varepsilon_1 < \frac{2p_0p_+ + p_+^2}{1 - 2(1 - p_0 - p_+)p_+ - p_0^2} < 1/2 + \varepsilon_2 \right\}. \quad (5.69)$$

In contrast, in the sign test for equivalence which, after the appropriate sufficiency reduction, refers to exactly the same family of distributions, the alternative hypothesis [cf. p. 72, (5.4)] is easily seen to admit the representation

$$\tilde{K}_{\varepsilon'_1, \varepsilon'_2}^S = \left\{ (p_0, p_+) \in (0, 1)^2 \mid p_0 + p_+ \leq 1, 1/2 - \varepsilon'_1 < \frac{p_+}{1 - p_0} < 1/2 + \varepsilon'_2 \right\} \quad (5.70)$$

where the  $\varepsilon'_\nu$ , like the  $\varepsilon_\nu$  ( $\nu = 1, 2$ ) of (5.69), can be arbitrarily chosen in the interval  $(0, 1/2)$ .

The graphs presented in Figure 5.1 and 5.2 visualize the differences in geometrical shape of the subspaces  $\tilde{K}_{\varepsilon_1, \varepsilon_2}^{\pm R}$  and  $\tilde{K}_{\varepsilon'_1, \varepsilon'_2}^S$ : Whereas the boundaries of the latter are simply straight lines, those of  $\tilde{K}_{\varepsilon_1, \varepsilon_2}^{\pm R}$  are slightly curved. If, as done in the first couple of graphs, the margins  $\varepsilon_\nu$  and  $\varepsilon'_\nu$  are set equal to each other (both for  $\nu = 1$  and  $\nu = 2$ ), the alternative hypothesis of the sign test for equivalence is much larger than that of the generalized signed rank test. On the other hand, if the  $\varepsilon_\nu$  and  $\varepsilon'_\nu$  are chosen in a way ensuring that the vertical section at  $p_0 = 0$  of both parameter subspaces coincide, then  $\tilde{K}_{\varepsilon_1, \varepsilon_2}^{\pm R}$  covers  $\tilde{K}_{\varepsilon'_1, \varepsilon'_2}^S$ , as illustrated by Figure 5.2. All in all, it is clear that we have no possibility to make the two versions of an alternative hypothesis about  $(p_0, p_+)$  perfectly coincident by a suitable adjustment of the respective tolerances  $\varepsilon_\nu$  and  $\varepsilon'_\nu$ . Admittedly, it would be desirable to base a nonparametric equivalence test for paired noncontinuous observations on a functional which reduces in the case of trinomially distributed  $D_i$  to the parametric function underlying the sign test for equivalence. However, there seems to exist no possibility to implement this idea without discarding all other information provided by the observations except that contained in the counting statistic  $(N_0, N_+)$  to which the data are reduced in the sign test.

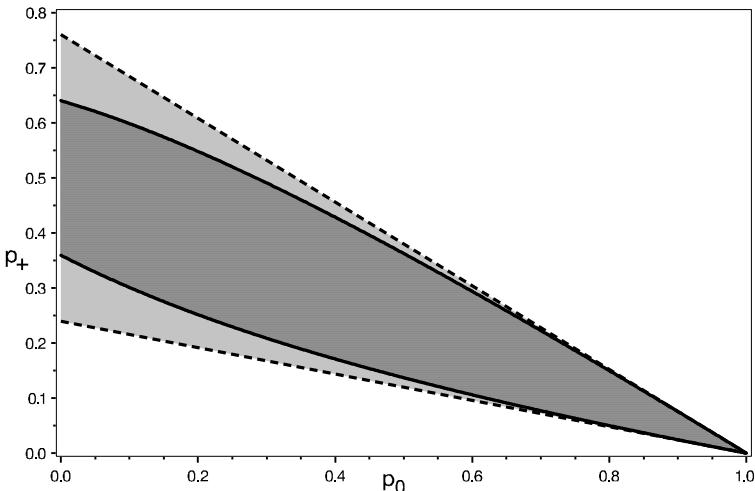


Figure 5.1 Boundaries of the equivalence hypotheses  $\tilde{K}_{\varepsilon_1, \varepsilon_2}^{\pm R}$  (—) and  $\tilde{K}_{\varepsilon'_1, \varepsilon'_2}^S$  (----) for  $\varepsilon_1 = \varepsilon'_1 = \varepsilon'_2 = \varepsilon_2 = .2602$ .

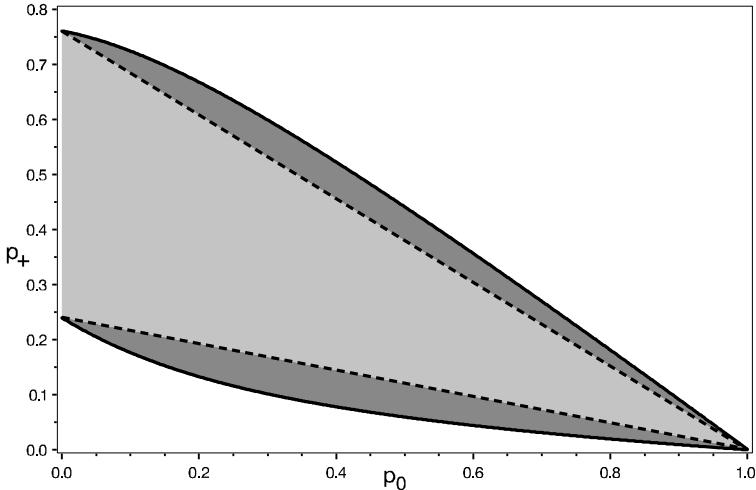


Figure 5.2 *Boundaries of the equivalence hypotheses*  $\tilde{K}_{\varepsilon_1, \varepsilon_2}^{\pm R}$  (—) and  $\tilde{K}_{\varepsilon'_1, \varepsilon'_2}^S$  (----) for  $\varepsilon_1 = \varepsilon_2 = .2602$ ,  $\varepsilon'_1 = \varepsilon'_2 = .4095$ .

#### *Tied-data version of the signed rank test for noninferiority*

A noninferiority analogue of the generalized signed rank test for equivalence derived in this section is obtained by replacing the critical region to be used in the two-sided case with

$$\left\{ (U_+/(1 - U_0) - 1/2 + \varepsilon) / \hat{\sigma}_{*n} > u_{1-\alpha} \right\} \quad (5.71)$$

where all symbols appearing in the expression for the test statistic have exactly the same meaning as before. In particular, the simplified formula for the  $U$ -statistics involved can be applied in the noninferiority case as well leading to considerable savings of computation time whenever the observed individual differences take values in a sufficiently coarse lattice of points. Of course, the basic property of the test given by (5.71) is that, asymptotically, its significance level does not exceed  $\alpha$  provided the testing problem is rewritten

$$H_1 : \frac{q_+}{1 - q_0} \leq 1/2 - \varepsilon \text{ versus } K_1 : \frac{q_+}{1 - q_0} > 1/2 - \varepsilon. \quad (5.72)$$

The extent to which the power of the test will change when the statement made about the target functional under the alternative hypothesis to be established is much less precise, can be expected to be of the same order of magnitude as in the corresponding tests for continuous data. In order to get a more precise idea of the finite-sample behavior of the test based on (5.71),

we studied by simulation its rejection probabilities under all discrete models considered in the two-sided case except those for which the true value of the target functional  $q_+/(1 - q_0)$  falls on the right-hand boundary of the equivalence hypothesis specified in (5.55). Regarding the type-I error risk, the differences between both versions of the test are negligible, so that in the noninferiority case, the size is likewise less than 4% for all sample sizes and both discrete models studied. Furthermore, except for  $n = 100$ , leaving the theoretical equivalence range unbounded to the right (except, of course, for the bound limiting the parameter space as a whole) leads to a marked increase in power indeed. For  $n = 20$  and  $n = 30$ , this difference exceeds 30% for all specific alternatives studied.

Table 5.23 *Simulated rejection probabilities<sup>†</sup> of the tie-corrected version of the signed rank test for noninferiority under the null hypothesis [(a), (d)] and three alternatives with  $q_+/(1 - q_0) = 1/2$  [(b), (e) and (f)] in samples of size  $n \in \{20, 30, 50, 100\}$ . [For a fully precise description of the distributions from which the data were generated see Table 5.21.]*

	$n = 20$	$n = 30$	$n = 50$	$n = 100$
(a)	.04000	.03702	.03571	.03694
(d)	.03618	.03461	.03527	.03638
(b)	.53048	.66813	.84671	.98430
(e)	.50557	.64704	.83111	.98098
(f)	.43928	.56918	.76780	.96391

<sup>†</sup> 100,000 replications of each Monte Carlo experiment

# 6

---

## *Equivalence tests for two unrelated samples*

---

### 6.1 Two-sample $t$ -test for equivalence

If the trial run in order to compare the two treatments  $A$  and  $B$  follows an ordinary two-arm design, the data set to be analyzed consists of  $m+n$  mutually independent observations  $X_1, \dots, X_m, Y_1, \dots, Y_n$ . For definiteness, we keep assigning the  $X$ 's to treatment group  $A$ , whereas the  $Y$ 's are assumed to belong to the group of subjects or patients given treatment  $B$ . The statistical model which is assumed to hold for these variables is the same as in the ordinary two-sample  $t$ -test for conventional one- and two-sided hypotheses. Accordingly, we suppose throughout this section that both the  $X_i$  and the  $Y_j$  follow a normal distribution with some common variance and possibly different expected values where all three parameters are unknown constants allowed to vary unrestrictedly over the respective part of the parameter space. Formally speaking, this means that the observations are assumed to satisfy the basic parametric model

$$X_i \sim \mathcal{N}(\xi, \sigma^2) \quad \forall i = 1, \dots, m, \quad Y_j \sim \mathcal{N}(\eta, \sigma^2) \quad \forall j = 1, \dots, n, \quad (6.1)$$

with  $\xi, \eta \in \mathbb{R}$ ,  $\sigma^2 \in \mathbb{R}_+$ .

In § 1.6, we have argued that the most natural measure of dissimilarity of two homoskedastic Gaussian distributions is the standardized difference of their means. Accordingly, under the present parametric model, we define equivalence of treatments  $A$  and  $B$  through the condition that the true value of this measure falls into a sufficiently narrow interval  $(-\varepsilon_1, \varepsilon_2)$  around zero. In other words, we formulate the testing problem as

$$\begin{aligned} H : & (\xi - \eta)/\sigma \leq -\varepsilon_1 \text{ or } (\xi - \eta)/\sigma \geq \varepsilon_2 \\ & \text{versus } K : -\varepsilon_1 < (\xi - \eta)/\sigma < \varepsilon_2 \quad (\varepsilon_1, \varepsilon_2 > 0), \end{aligned} \quad (6.2)$$

in direct analogy to the setting of the paired  $t$ -test for equivalence [recall (5.27), (5.28)].

The construction of an optimal solution to this problem exploits essentially the same ideas as underlie the derivation of the paired  $t$ -test for equivalence. In the unpaired case, both hypotheses remain invariant under any transformation of the form  $(x_1, \dots, x_m, y_1, \dots, y_n) \mapsto (a+bx_1, \dots, a+bx_m, a+by_1, \dots, a+by_n)$

with  $(a, b) \in \mathbb{R} \times \mathbb{R}_+$ , applying to each coordinate of a point in the joint sample space  $\mathbb{R}^N$  (with  $N = m + n$ ) the same scale change and the same translation. Denoting the group of all transformations of this form by  $\mathcal{G}$ , it is easy to show that the construction of a test which is uniformly most powerful among all level- $\alpha$  tests for (6.2) remaining invariant under  $\mathcal{G}$  has been accomplished as soon as a test has been found which is UMP at the same level for the reduced problem

$$\tilde{H} : \tilde{\theta} \leq -\tilde{\varepsilon}_1 \text{ or } \tilde{\theta} \geq \tilde{\varepsilon}_2 \quad \text{versus} \quad \tilde{K} : -\tilde{\varepsilon}_1 < \tilde{\theta} < \tilde{\varepsilon}_2 . \quad (6.2)$$

In the two-sample case,  $\tilde{\theta}$  has to be interpreted as the parameter of a non-central  $t$ -distribution with  $m + n - 2 = N - 2$  degrees of freedom, and  $(-\tilde{\varepsilon}_1, \tilde{\varepsilon}_2)$  denotes the equivalence interval of (6.2) rescaled by multiplication with  $\sqrt{mn/N}$ .

Except for differences in notation for the number of degrees of freedom and the limits of the equivalence range, (6.2) is the same as the testing problem (5.28) which has been considered in § 5.3 in connection with the paired  $t$ -test for equivalence. Hence, it follows [cf. (5.32), (5.33)] that in order to obtain an UMP level- $\alpha$  test for (6.2), we have to use a critical region of the form

$$\left\{ \tilde{C}_{\alpha; m, n}^1(-\tilde{\varepsilon}_1, \tilde{\varepsilon}_2) < T < \tilde{C}_{\alpha; m, n}^2(-\tilde{\varepsilon}_1, \tilde{\varepsilon}_2) \right\} , \quad (6.3)$$

where the critical constants  $\tilde{C}_{\alpha; m, n}^k(-\tilde{\varepsilon}_1, \tilde{\varepsilon}_2)$ ,  $k = 1, 2$ , have to be determined by solving the equation system

$$\begin{aligned} G_{-\tilde{\varepsilon}_1}^*(C_2) - G_{-\tilde{\varepsilon}_1}^*(C_1) &= \alpha = \\ G_{\tilde{\varepsilon}_2}^*(C_2) - G_{\tilde{\varepsilon}_2}^*(C_1) , \quad -\infty < C_1 < C_2 < \infty . \end{aligned} \quad (6.4)$$

In (6.4), the superscript \* is added to the symbol  $G_{\tilde{\varepsilon}}(\cdot)$  which has been used in § 5.3 for noncentral  $t$ -distribution functions, in order to make conspicuous the change in the way the number of degrees of freedom has to be determined. In the two-sample case, the latter must be set equal to  $N - 2$  instead of  $n - 1$ . The desired UMPI (uniformly most powerful invariant) test for the primary problem (6.2) is obtained by applying (6.3) with  $T$  as the ordinary two-sample  $t$ -statistic to be computed from the individual observations by means of the well known formula

$$T = \sqrt{mn(N-2)/N} (\bar{X} - \bar{Y}) / \left\{ \sum_{i=1}^m (X_i - \bar{X})^2 + \sum_{j=1}^n (Y_j - \bar{Y})^2 \right\}^{1/2} . \quad (6.5)$$

The simplified computational scheme for determining the critical bounds to the  $t$ -statistic derived in § 5.3 for symmetrically chosen equivalence limits to the standardized population mean, has likewise its direct counterpart in the two-sample case. More precisely speaking, it follows by Lemma A.1.6

that whenever both tolerances  $\varepsilon_1$  and  $\varepsilon_2$  are set equal to some common value  $\varepsilon > 0$ , say, the critical region (6.3) admits the representation

$$\left\{ |T| < \tilde{C}_{\alpha; m,n}(\varepsilon) \right\}, \quad (6.6)$$

where  $\tilde{C}_{\alpha; m,n}(\varepsilon)$  is explicitly given as the square root of the lower  $100\alpha$  percentage point of an  $F$ -distribution with  $1, N - 2$  degrees of freedom and noncentrality parameter  $\lambda_{nc}^2 = mn\varepsilon^2/N$ . In view of this, the practical implementation of the two-sample  $t$ -test for equivalence in a symmetric sense is extremely easy in any programming environment providing an intrinsic function for the computation of quantiles of noncentral  $F$ -distributions (like `finv` in SAS and `qf` in R). Moreover, making use of Table 6.1, no extra computational tools are needed at all to carry out the test in cases of balanced designs with common sample sizes being multiples of 5 between 10 and 75, pro-

Table 6.1 *Critical constant  $\tilde{C}_{.05; n,n}(\varepsilon)$  of the two-sample  $t$ -test for equivalence at level  $\alpha = 5\%$  with common size  $n$  of both samples in the symmetric case  $(\varepsilon_1, \varepsilon_2) = (-\varepsilon, \varepsilon)$ , for  $\varepsilon = .25(.25)1.00$  and  $n = 10(5)75$ .*

$n$	$\varepsilon =$			
	.25	.50	.75	1.00
10	0.07434	0.11864	0.25412	0.61365
15	0.07997	0.16083	0.46595	1.08697
20	0.08626	0.21755	0.73546	1.50303
25	0.09313	0.29167	1.00477	1.87065
30	0.10058	0.38434	1.25435	2.20346
35	0.10866	0.49320	1.48481	2.50990
40	0.11738	0.61250	1.69952	2.79542
45	0.12681	0.73572	1.90128	3.06383
50	0.13698	0.85816	2.09219	3.31790
55	0.14795	0.97741	2.27384	3.55971
60	0.15977	1.09257	2.44747	3.79090
65	0.17248	1.20355	2.61406	4.01275
70	0.18615	1.31055	2.77441	4.22631
75	0.20082	1.41389	2.92917	4.43246

vided one chooses  $\varepsilon_1 = \varepsilon_2 = \varepsilon \in \{.25, .50, .75, 1.00\}$ . The maxima of the power functions of the tests determined by these critical constants are shown in Table 6.2. Denoting, for  $\varepsilon_1 = \varepsilon_2 = \varepsilon$  and  $m = n$ , the rejection probability of the test with critical region (6.3) against the specific alternative  $\theta = (\xi - \eta)/\sigma = 0$  [ $\leftarrow$

$\xi = \eta]$  by  $\tilde{\beta}_{\alpha;n,n}(\varepsilon)$ , this maximal power is readily computed by means of the simple formula

$$\tilde{\beta}_{\alpha;n,n}(\varepsilon) = 2 G_0^*(\tilde{C}_{\alpha;n,n}(\varepsilon)) - 1, \quad (6.7)$$

with  $G_0^*(\cdot)$  as the central version of the  $t$ -distribution function with  $N - 2$  degrees of freedom.

Table 6.2 *Power attained at  $\theta = (\xi - \eta)/\sigma = 0$  when using the critical constants shown in Table 6.1 for the two-sample  $t$ -statistic.*

$n$	$\varepsilon =$			
	.25	.50	.75	1.00
10	.05844	.09312	.19772	.45288
15	.06317	.12662	.35514	.71368
20	.06829	.17106	.53343	.85890
25	.07381	.22820	.67995	.93250
30	.07977	.29786	.78525	.96845
35	.08620	.37654	.85779	.98554
40	.09314	.45801	.90679	.99348
45	.10062	.53614	.93946	.99710
50	.10867	.60710	.96099	.99873
55	.11734	.66945	.97505	.99945
60	.12666	.72319	.98414	.99976
65	.13667	.76901	.98998	.99990
70	.14740	.80781	.99370	.99996
75	.15888	.84051	.99606	.99998

Even if the restriction of a symmetric choice of the equivalence limits to the target parameter  $\theta = (\xi - \eta)/\sigma$  has to be abandoned, the practical implementation of the two-sample  $t$ -test for equivalence continues to be a quick and easy matter, provided one makes use of an appropriate computer program. A SAS macro or, alternatively, a R function which does all necessary computations determining in particular both critical bounds to the test statistic, is to be found in the **WKTSEQ2 Source Code Package** under the program name **tt2st**. The basic algorithm is the same as that underlying its analogue for the one-sample case [→ **tt1st**]. In addition to solving the pair of equations (6.4) for arbitrary choices of  $\varepsilon_1$  and  $\varepsilon_2 > 0$ , it outputs the power of the test against the alternative of coincident expected values.

*Example 6.1*

Table 6.3 shows the raw data obtained from a comparative trial of moxonodin [an alpha receptor blocking agent] and captopril [an inhibitor of the angiotensin converting enzyme] in the antihypertensive treatment of patients suffering from major depression. Since moxonodin was known to have not only

Table 6.3 *Reduction of diastolic blood pressure [mm Hg] observed in patients suffering from major depression after 4 weeks of daily treatment with 0.2–0.4 mg moxonodin (a) and 25–50 mg captopril (b), respectively.*

(a) moxonidin		(b) captopril	
<i>i</i>	$X_i$	<i>j</i>	$Y_j$
1	10.3	1	3.3
2	11.3	2	17.7
3	2.0	3	6.7
4	-6.1	4	11.1
5	6.2	5	-5.8
6	6.8	6	6.9
7	3.7	7	5.8
8	-3.3	8	3.0
9	-3.6	9	6.0
10	-3.5	10	3.5
11	13.7	11	18.7
12	12.6	12	9.6

$\bar{X} = 4.175, S_X = 7.050$	$\bar{Y} = 7.208, S_Y = 6.625$
--------------------------------	--------------------------------

antihypertensive but also neuroregulatory effects alleviating some of the key symptoms making up a major depression, the aim of the trial was to establish equivalence of this drug to the ACE inhibitor with respect to antihypertensive efficacy. Furthermore, only rather extreme increases in the reduction of blood pressure induced by captopril were considered undesirable, on the ground that the latter drug was known to produce comparatively weak effects. On the other hand, only comparatively small losses in antihypertensive effectiveness were considered compatible with equivalence. Making allowance for both of these aspects, a nonsymmetric choice of the equivalence range for the standardized difference of means was considered appropriate specifying  $(-\varepsilon_1, \varepsilon_2) = (-0.50, 1.00)$ .

With the data given in the above table, the test statistic (6.5) is computed to be  $T = \sqrt{12^2 \cdot 22/24} \cdot (4.175 - 7.208) / \sqrt{11 \cdot 7.050^2 + 11 \cdot 6.625^2} = -1.086$ .

On the other hand, with  $\alpha = .05$ ,  $m = n = 12$ ,  $\varepsilon_1 = 0.50$  and  $\varepsilon_2 = 1.00$ , **tt2st** yields (0.27977, 0.93088) as the critical interval for  $T$ . Since the observed value of  $T$  falls clearly outside this interval, the data do not allow to reject the null hypothesis of nonequivalence of moxonodin to captopril with respect to antihypertensive efficacy. Furthermore, the power against the alternative that the underlying Gaussian distributions coincide, is computed to be .21013. Thus, even under the most extreme alternative, the chance of establishing equivalence was rather low in the present study. In view of the small size of both samples, this is hardly surprising.

### *Two-sample t-test for noninferiority*

The modifications required in order to obtain an optimum testing procedure for the noninferiority problem associated with (6.2) are largely analogous to those described in § 5.3 for the case of paired observations. The first step consists of replacing the hypotheses of (6.2) with

$$H_1 : (\xi - \eta)/\sigma \leq -\varepsilon \quad \text{vs.} \quad K_1 : (\xi - \eta)/\sigma > -\varepsilon \quad (\varepsilon > 0). \quad (6.8)$$

Essentially the same arguments allowing one to establish the UMPI property of the  $t$ -test for two-sided equivalence show that the class of all invariant level- $\alpha$  tests for (6.8) contains a uniformly most powerful element, namely the test given by the rejection region

$$\left\{ T > t_{N-2;1-\alpha}(-\sqrt{mn/N}\varepsilon) \right\}, \quad (6.9)$$

where  $t_{N-2;1-\alpha}(-\sqrt{mn/N}\varepsilon)$  denotes the  $(1 - \alpha)$ -quantile of a  $t$ -distribution with  $df = N - 2$  and noncentrality parameter  $-\sqrt{mn/N}\varepsilon$ . As in the paired  $t$ -test for noninferiority, these critical values of the  $t$ -statistic are readily computed by calling the SAS intrinsic function **tinv** or the R function **qt**. A compilation of critical values for balanced designs with  $n = 10$  through 75 observations per group and equivalence margins  $\varepsilon \in \{k/10 \mid k = 1, \dots, 5\}$  is shown in Table 6.4. The power of the test against any alternative  $(\xi, \eta, \sigma^2)$  with  $(\xi - \eta)/\sigma = \theta > -\varepsilon$  is as easy to compute as the critical constant, namely by evaluating the survivor function  $1 - G^*_{\sqrt{mn/N}\theta}(t)$  of the noncentral  $t$ -distribution with  $df = N - 2$  and noncentrality parameter  $\sqrt{mn/N}\theta$  at  $t = t_{N-2;1-\alpha}(-\sqrt{mn/N}\varepsilon)$ . If the alternative of interest is a null alternative specifying  $\xi = \eta$ , power computation is particularly simple since we then can write  $\text{POW}_0 = G_0^*(-t_{N-2;1-\alpha}(-\sqrt{mn/N}\varepsilon))$  where  $G_0^*(\cdot)$  is as in (6.7). The results of computing  $\text{POW}_0$  for the same sample sizes and choices of the equivalence margin as in Table 6.4 are the entries in Table 6.5.

Table 6.4 Critical constant  $t_{2n-2,.95}(-\sqrt{n/2}\varepsilon)$  of the two-sample *t*-test for noninferiority at level  $\alpha = 5\%$  in the case of a balanced design with  $n = 10(5)75$  and equivalence margin  $\varepsilon = .10(.10).50$ .

$n$	$\varepsilon =$				
	.10	.20	.30	.40	.50
10	1.49038	1.24922	1.01058	0.77445	0.54081
15	1.41206	1.12533	0.84095	0.55890	0.27916
20	1.35708	1.03048	0.70613	0.38404	0.06418
25	1.31267	0.95034	0.59022	0.23231	-0.12342
30	1.27442	0.87948	0.48672	0.09614	-0.29229
35	1.24029	0.81518	0.39223	-0.02857	-0.44723
40	1.20915	0.75584	0.30468	-0.14435	-0.59127
45	1.18032	0.70044	0.22269	-0.25293	-0.72645
50	1.15334	0.64826	0.14531	-0.35552	-0.85427
55	1.12787	0.59879	0.07183	-0.45303	-0.97581
60	1.10369	0.55163	0.00169	-0.54615	-1.09192
65	1.08061	0.50649	-0.06552	-0.63544	-1.20329
70	1.05849	0.46311	-0.13016	-0.72134	-1.31045
75	1.03721	0.42131	-0.19249	-0.80420	-1.41385

Table 6.5 Power against the null alternative  $\theta = 0$  attained in the two-sample *t*-test for noninferiority for the sample sizes and equivalence-margin specifications underlying Table 6.4.

$n$	$\varepsilon =$				
	.10	.20	.30	.40	.50
10	.07672	.11379	.16280	.22436	.29763
15	.08448	.13500	.20375	.29034	.39109
20	.09138	.15465	.24221	.35154	.47458
25	.09777	.17335	.27891	.40864	.54886
30	.10380	.19139	.31415	.46187	.61445
35	.10957	.20891	.34806	.51135	.67194
40	.11513	.22601	.38071	.55720	.72197
45	.12053	.24275	.41215	.59954	.76526
50	.12579	.25917	.44238	.63852	.80248
55	.13094	.27528	.47144	.67428	.83433
60	.13599	.29112	.49933	.70700	.86145
65	.14095	.30669	.52607	.73686	.88446
70	.14584	.32201	.55168	.76404	.90389
75	.15066	.33707	.57619	.78871	.92025

A glance over Tables 6.2 and 6.5 shows that the differences in power between the two- and the one-sided version of the two-sample  $t$ -test for equivalence are roughly comparable to those found in § 5.3 between both versions of the paired  $t$ -test for equivalence: Given some common value of the equivalence margin, the much higher precision of the statement about the target parameter which can be made in the two-sided equivalence test in case of a significance decision, is bought at the price that the sample sizes required for ensuring the same level of power are considerably larger.

---

## 6.2 Mann-Whitney test for equivalence

The equivalence testing problem treated in the previous section can easily be embedded in a purely nonparametric framework. To see this, it suffices to recall the elementary fact that under the specific parametric assumption that both the  $X_i$  and the  $Y_j$  are normally distributed with common variance  $\sigma^2$  and arbitrary expectation  $\xi$  and  $\eta$ , respectively, the probability  $\pi_+$ , say, of obtaining an observation from the first distribution which exceeds some single observation drawn from the second, can be written  $\pi_+ = \Phi((\xi - \eta)/\sqrt{2}\sigma)$ . Since the standard normal cdf  $\Phi$  is well known to be both continuous and strictly increasing, this identity implies that the testing problem (6.2) is equivalent to

$$H : \pi_+ \leq 1/2 - \varepsilon'_1 \text{ or } \pi_+ \geq 1/2 + \varepsilon'_2 \quad \text{versus} \quad K : 1/2 - \varepsilon'_1 < \pi_+ < 1/2 + \varepsilon'_2 , \quad (6.10)$$

provided  $\pi_+$  and the  $\varepsilon'_\nu$  are defined by

$$\pi_+ = P[X_i > Y_j] ; \quad \varepsilon'_\nu = \Phi(\varepsilon_\nu/\sqrt{2}) - 1/2 , \quad \nu = 1, 2 . \quad (6.11)$$

Since the Mann-Whitney functional  $\pi_+$  is well defined for any pair  $(F, G)$  of *continuous* distribution functions on  $\mathbb{R}^1$  such that

$$X_i \sim F \quad \forall i = 1, \dots, m , \quad Y_j \sim G \quad \forall j = 1, \dots, n , \quad (6.12)$$

it is natural to treat (6.10) as a nonparametric formulation of the problem of testing for equivalence of two continuous distributions of arbitrary shape. On the following pages, we will show how an asymptotically distribution-free solution to that testing problem can be based on the large-sample distribution of the  $U$ -statistic estimator  $W_+$ , say, of  $\pi_+$ . As is well known (see, e.g., Lehmann, 1975, p. 335), this estimator is simply given as the Wilcoxon two-sample statistic in its Mann-Whitney form divided by  $mn$  so that we have

$$W_+ = (1/mn) \sum_{i=1}^m \sum_{j=1}^n I_{(0, \infty)}(X_i - Y_j) . \quad (6.13)$$

Not surprisingly, the role played here by the statistic  $W_+$  and the functional  $\pi_+$  will be analogous to that of the quantities  $U_+$  [ $\rightarrow$  p. 100, (5.42)] and  $q_+$  [ $\rightarrow$  (5.41)] in the derivation of a nonparametric equivalence test for paired observations.

Irrespective of the shape of the underlying distribution functions  $F$  and  $G$ , the exact variance of  $W_+$  is given as

$$\text{Var}[W_+] = (1/mn) \cdot \\ \left( \pi_+ - (m+n-1)\pi_+^2 + (m-1)\Pi_{XXY} + (n-1)\Pi_{XYY} \right), \quad (6.14)$$

where

$$\begin{aligned} \Pi_{XXY} &= P[X_{i_1} > Y_j, X_{i_2} > Y_j], \\ \Pi_{XYY} &= P[X_i > Y_{j_1}, X_i > Y_{j_2}] \end{aligned} \quad (6.15)$$

(see, e.g., Lehmann, 1975, (2.20)). Natural estimators of the functionals  $\Pi_{XXY}$  and  $\Pi_{XYY}$  are given by the  $U$ -statistics

$$\widehat{\Pi}_{XXY} = \frac{2}{m(m-1)n} \times \\ \sum_{i_1=1}^{m-1} \sum_{i_2=i_1+1}^m \sum_{j=1}^n I_{(0,\infty)}(X_{i_1} - Y_j) \cdot I_{(0,\infty)}(X_{i_2} - Y_j), \quad (6.16a)$$

$$\widehat{\Pi}_{XYY} = \frac{2}{mn(n-1)} \times \\ \sum_{i=1}^m \sum_{j_1=1}^{n-1} \sum_{j_2=j_1+1}^n I_{(0,\infty)}(X_i - Y_{j_1}) \cdot I_{(0,\infty)}(X_i - Y_{j_2}). \quad (6.16b)$$

As is true for any two-sample  $U$ -statistic estimator with square integrable kernel (cf. Randles and Wolfe, 1979, Corollary 3.4.9),  $W_+$ ,  $\widehat{\Pi}_{XXY}$  and  $\widehat{\Pi}_{XYY}$  are consistent for the respective estimands  $\pi_+$ ,  $\Pi_{XXY}$  and  $\Pi_{XYY}$ . Hence, a consistent estimator of  $\text{Var}[W_+]$  is given by

$$\hat{\sigma}^2[W_+] = \frac{1}{mn} \left( W_+ - (m+n-1)W_+^2 + (m-1)\widehat{\Pi}_{XXY} + (n-1)\widehat{\Pi}_{XYY} \right). \quad (6.17)$$

In view of the asymptotic normality of the Mann-Whitney statistic (see, e.g. Lehmann, 1975, p. 69), consistency of  $\hat{\sigma}^2[W_+]$  implies that the statistic  $(W_+ - \pi_+)/\hat{\sigma}[W_+]$  converges in law to a standard Gaussian variable whenever we have  $0 < \pi_+ < 1$ ,  $m, n \rightarrow \infty$  and  $m/n \rightarrow \lambda$  (with  $\lambda$  denoting some positive real number). Hence, we can apply once more the general theory outlined in

§ 3.4 leading to the conclusion that an asymptotically valid testing procedure for the equivalence problem (6.10) is obtained by using the decision rule:

Reject nonequivalence if and only if

$$\left| W_+ - 1/2 - (\varepsilon'_2 - \varepsilon'_1)/2 \right| / \hat{\sigma}[W_+] < C_{MW}(\alpha; \varepsilon'_1, \varepsilon'_2), \quad (6.18)$$

with

$$C_{MW}(\alpha; \varepsilon'_1, \varepsilon'_2) = \left\{ 100\alpha - \text{percentage point of the } \chi^2 \text{ - distribution with } df = 1 \text{ and } \lambda_{nc}^2 = (\varepsilon'_1 + \varepsilon'_2)^2 / 4\hat{\sigma}^2[W_+] \right\}^{1/2}. \quad (6.19)$$

At the URL associated with this book, one finds also a tool for the quick and easy implementation of the testing procedure (6.18), appropriately termed Mann-Whitney test for equivalence in the research paper where its basic properties have first been rigorously established (Wellek, 1996). The program which had originally been written as another SAS macro and translated later on in R, is named `mawi` and requires as input data the actual values of the significance level  $\alpha$ , the sample sizes  $m, n$  and the tolerances  $\varepsilon'_1, \varepsilon'_2$ , as well as the name of a file containing the raw data under analysis. In the output, the user finds in addition to the input data, the values computed for the Mann-Whitney statistic  $W_+$ , its estimated standard deviation  $\hat{\sigma}[W_+]$ , the critical bound  $C_{MW}^a(\alpha; \varepsilon'_1, \varepsilon'_2)$  and an indicator which is set equal to 1 if and only if the data fall into the rejection region corresponding to (6.18).

### *Example 6.1 (continued)*

In order to illustrate the use of the testing procedure (6.18), we reanalyze the blood pressure reduction data of Table 6.3 by means of the Mann-Whitney statistic. According to (6.11), the equivalence hypothesis  $-.50 < (\xi - \eta)/\sigma < 1.00$  on two Gaussian distributions with means  $\xi, \eta$  and common variance  $\sigma^2$  can be rewritten as  $.3618 < \pi_+ < .7602$ . The corresponding specification of the  $\varepsilon'_\nu$  in (6.10), (6.18), (6.19) is  $\varepsilon'_1 = .1382, \varepsilon'_2 = .2602$ . With these values and the raw data displayed in Table 6.3, the program `mawi` yields  $W_+ = .41667, \hat{\sigma}[W_+] = .11133, C_{MW}(.05; \varepsilon'_1, \varepsilon'_2) = .30078$ . Since we have  $|.41667 - 1/2 - (.2602 - .1382)/2|/.11133 = 1.2964$ , the observed value of the test statistic clearly exceeds its critical upper bound. Thus it follows that the Mann-Whitney test for equivalence leads to the same decision as the parametric test derived in the previous subsection, namely acceptance of the null hypothesis of nonequivalence of both drugs with respect to their antihypertensive efficacy.

In order to investigate the level and power of the Mann-Whitney test for equivalence in real applications with samples of finite or even small size, one has to rely on Monte Carlo simulation once more. The fact that the ordinary Mann-Whitney test is well known (see Hájek et al., 1999, pp. 96–7) to be optimal against logistic shift alternatives suggests to restrict such an investigation to constellations where the cdf and the density of the  $Y_j$  has one of the standard forms listed in Table 5.16 and the distribution of the  $X_i$  is the result of shifting that of the  $Y_j$  by a real constant  $\vartheta$ . Accordingly, in the majority of simulation experiments run to study the finite sample behavior of the testing procedure (6.18), random numbers of the form  $X_i = X_i^\circ + \vartheta$ ,  $Y_j = Y_j^\circ$  were generated with  $(X_1^\circ, \dots, X_m^\circ, Y_1^\circ, \dots, Y_n^\circ)$  as a sample of size  $m + n$  from a standardized Gaussian, Cauchy, uniform, exponential, Laplace and logistic distribution, respectively. In addition, one constellation was investigated where both distributions under comparison differ in form. The rationale behind the procedure of generating pairs of samples from distributions of this latter kind was the same as for the location models: The distribution from which the  $Y_j$  were taken, was fixed (at  $\mathcal{N}(0, 1)$ , i.e., the standard Gaussian), and the center  $\vartheta$  of the distribution of the  $X_i$  (belonging to the Gaussian family as well but with threefold standard deviation) was successively shifted from zero to the point at which the functional  $\pi_+$  [recall (6.11)] reaches  $1/2 - \varepsilon'_1$ ,  $1/2 + \varepsilon'_2$ , and  $1/2$ , respectively.

For  $X_i = X_i^\circ + \vartheta$ ,  $Y_j = Y_j^\circ$  with  $X_i^\circ$  and  $Y_j^\circ$  following the (continuous) cdf  $F_\circ$  and  $G_\circ$ , respectively, the functional  $\pi_+$  with respect to which equivalence of the underlying distributions is asserted under the alternative hypothesis of (6.10), admits the representation

$$\pi_+ = \int_{-\infty}^{\infty} [1 - F_\circ(y - \vartheta)] dG_\circ(y) . \quad (6.20)$$

It is easy to verify that in any setting with  $F_\circ = G_\circ$  and  $F_\circ(-x) = 1 - F_\circ(x) \forall x \in \mathbb{R}$ , i.e., in any ordinary shift model with a symmetric common baseline cdf, the integral on the right-hand side of (6.20) coincides with the expression on the left of equation (5.50), provided,  $\vartheta$  is replaced with  $\vartheta/2$  in the latter. Hence except for the exponential distribution, for all baseline cdf's considered in § 5.4, the limits  $-\varepsilon_1, \varepsilon_2$ , say, of the equivalence range for the location parameter  $\vartheta$  which corresponds to some given equivalence range  $(1/2 - \varepsilon'_1, 1/2 + \varepsilon'_2)$  for  $\pi_+$  under the two-sample shift model generated by  $F_\circ$ , can be determined by means of the formulae shown in the third column of Table 5.16. In the two-sample shift model obtained by setting  $F_\circ(x) = \Phi(x/3)$ ,  $G_\circ(y) = \Phi(y)$  such that the distributions under comparison differ in dispersion for any value of the location parameter,  $\pi_+$  admits a simple representation through the normal distribution function as well. It reads  $\pi_+ = \Phi(\vartheta/\sqrt{10})$ , in close analogy to the ordinary Gaussian shift model given by  $F_\circ = G_\circ = \Phi$ . Making use of these relationships one obtains the values displayed in Table 6.6 which is a direct two-sample counterpart to Table 5.17.

Specifying the equivalence limits to  $\pi_+$  by choosing  $\varepsilon'_1 = \varepsilon'_2 = .20$  can be motivated as explained in § 1.7 [cf. Table 1.1, (ii)]. The other specification of  $\varepsilon'_1, \varepsilon'_2$  referred to in Table 6.6 corresponds to the equivalence range used in Example 6.1 [to be resumed below] for the standardized difference of the means in the Gaussian shift model.

Table 6.6 *Equivalence ranges for the location parameter  $\vartheta$  corresponding to  $.30 < \pi_+ < .70$  and  $.3618 < \pi_+ < .7602$ , respectively, in seven different models for the two-sample problem with  $X_i \sim F_o(\cdot - \vartheta)$ ,  $Y_j \sim G_o(\cdot)$ .*

Family of Distributions	Equivalence Range $(-\varepsilon_1, \varepsilon_2)$ for $\vartheta$ corresponding to $1/2 - \varepsilon'_1 < \pi_+ < 1/2 + \varepsilon'_2$ , for $(1/2 - \varepsilon'_1, 1/2 + \varepsilon'_2) =$	
	$(0.3000, 0.7000)$	$(0.3618, 0.7602)$
Gaussian	$(-0.7416, 0.7416)$	$(-0.5000, 1.0000)$
Cauchy	$(-1.4531, 1.4531)$	$(-0.9274, 2.1324)$
Uniform	$(-0.2254, 0.2254)$	$(-0.1494, 0.3075)$
Exponential	$(-0.5108, 0.5108)$	$(-0.3235, 0.7348)$
Laplace	$(-0.8731, 0.8731)$	$(-0.5770, 1.2070)$
Logistic	$(-1.2636, 1.2636)$	$(-0.8491, 1.7127)$
$\mathcal{N}(\vartheta, 9)/\mathcal{N}(0, 1)$	$(-1.6583, 1.6583)$	$(-1.1180, 2.2356)$

In the simulation study of the Mann-Whitney test for equivalence, a balanced design was assumed throughout, and the common sample size  $n$  was chosen from the set  $\{12, 24, 36\}$ . For each of these sample sizes and all  $7 \times 2$  constellations covered by Table 6.6, the rejection probability of the test at the usual nominal significance level of 5% under  $\vartheta = -\varepsilon_1$ ,  $\vartheta = \varepsilon_2$  and  $\vartheta = 0 (\Leftrightarrow \pi_+ = 1/2)$  was determined on the basis of 100,000 replications of the respective Monte Carlo experiment. As can be seen from the results of this study summarized in Table 6.7, the nominal level is strictly maintained in all settings with a symmetric choice of the equivalence limits, even for sample sizes as small as 12. The only cases where the test shows some anticonservative tendency are those where very small samples are used for establishing an equivalence hypothesis exhibiting marked asymmetry, and even then the extent of anticonservatism seems not really serious from a practical point of view. Under the ordinary shift model with a standard Gaussian cdf as baseline, the loss in power as compared to the optimal parametric procedure (i.e., the two-sample  $t$ -test for equivalence derived in the preceding section) is roughly of the same order of magnitude as in the analogous paired-sample setting [cf. p. 105].

Table 6.7 Simulated rejection probabilities of the Mann-Whitney test for equivalence at both boundaries of the hypotheses (columns 5, 6) and  $\pi_+ = 1/2$  (rightmost column) with samples of common size  $n = 12, 24, 36$  from distributions belonging to the families listed in Table 6.6 and equivalence range  $(1/2 - \varepsilon'_1, 1/2 + \varepsilon'_2) = (.3000, .7000), (.3618, .7602)$ . [The italicized values give the power of the two-sample t-test for equivalence at level  $\alpha = .05$  against  $\pi_+ = 1/2$ .]

Family of Distributions	$(1/2 - \varepsilon'_1, 1/2 + \varepsilon'_2)$	Rejection Probability					
		$m$	$n$	$1/2 - \varepsilon'_1$	$1/2 + \varepsilon'_2$	at $\pi_+ =$ $1/2$	
Gaussian	(.3000, .7000)	12	12	.05015	.04941	.22020 <i>.24605</i>	
	"	24	24	.04510	.04565	.57929 <i>.63971</i>	
	"	36	36	.04427	.04476	.82186 <i>.86027</i>	
	(.3618, .7602)	12	12	.05514	.04327	.19727 <i>.21013</i>	
	"	24	24	.05273	.04026	.46811 <i>.49669</i>	
	"	36	36	.04828	.03900	.65010 <i>.67635</i>	
	Cauchy	(.3000, .7000)	12	12	.04925	.04780	.20180
	"	24	24	.04644	.04536	.56918	
	"	36	36	.04466	.04442	.81617	
	(.3618, .7602)	12	12	.05107	.04336	.17994	
Uniform	"	24	24	.05143	.03939	.45973	
	"	36	36	.04967	.03962	.64708	
	(.3000, .7000)	12	12	.04503	.04397	.19944	
	"	24	24	.04312	.04282	.56956	
	"	36	36	.04267	.04291	.81578	
	(.3618, .7602)	12	12	.04872	.03800	.17954	
Exponential	"	24	24	.04892	.03709	.45624	
	"	36	36	.04928	.03824	.64447	
	(.3000, .7000)	12	12	.04631	.04557	.20216	
	"	24	24	.04471	.04404	.56989	
	"	36	36	.04285	.04322	.81782	
	(.3618, .7602)	12	12	.04949	.04054	.18135	
	"	24	24	.05028	.03879	.45580	
	"	36	36	.04782	.03860	.64542	

Table 6.7 (*continued*)

Family of Distribution	$(1/2 - \varepsilon'_1, 1/2 + \varepsilon'_2)$	Rejection Probability				
		$m$	$n$	at $\pi_+ =$	$1/2 - \varepsilon'_1$	$1/2 + \varepsilon'_2$
Laplace	(.3000, .7000)	12	12	.04743	.04745	.20001
	"	24	24	.04582	.04425	.56931
	"	36	36	.04255	.04363	.81818
	(.3618, .7602)	12	12	.05138	.04177	.17918
	"	24	24	.05031	.03961	.46001
	"	36	36	.04833	.03904	.64797
Logistic	(.3000, .7000)	12	12	.04518	.04577	.20179
	"	24	24	.04371	.04507	.57006
	"	36	36	.04324	.04445	.81664
	(.3618, .7602)	12	12	.05102	.03979	.17807
	"	24	24	.05045	.03922	.45537
	"	36	36	.04650	.03973	.64290
$\mathcal{N}(\vartheta, 9)/\mathcal{N}(0, 1)$	(.3000, .7000)	12	12	.04862	.05058	.17883
	"	24	24	.04506	.04605	.47684
	"	36	36	.04403	.04431	.73950
	(.3618, .7602)	12	12	.05429	.04447	.16224
	"	24	24	.05090	.04088	.39080
	"	36	36	.04914	.03921	.58209

All in all, the simulation results shown in Table 6.7 allow the following conclusions:

- (i) As long as the equivalence range is chosen as an interval symmetric about 1/2 (which will be the case in the majority of practical applications), sample sizes of one dozen per group are sufficient to ensure that the Mann-Whitney test for equivalence is strictly valid with respect to the significance level. This is even true for pairs of underlying distribution function which satisfy the null hypotheses but exhibit gross departure from homoskedasticity. The anticonservative tendency to be observed in very small samples under some parametric submodels for markedly asymmetric choices of the equivalence limits is only slight and can be ignored for most practical purposes.
- (ii) The power attained by the Mann-Whitney test for equivalence under various shift models shows little sensitivity even to gross changes in form of the baseline distribution. Substantial differences in power can only be seen if any of the traditional location-shift models is contrasted with the model involving heteroskedasticity in addition to shift in location. In fact, heteroskedasticity (as well as other differences in form of the

two distributions to be compared) has a marked negative impact on the efficiency of the test.

- (iii) Conclusion (iii) stated on p. 106 with regard to the efficiency of the signed rank test for equivalence under the Gaussian submodel relative to the (paired)  $t$ -test applies mutatis mutandis to the Mann-Whitney test for equivalence as well.

By direct analytical arguments, some interesting and intuitively plausible properties of the (exact) power function of the Mann-Whitney test for equivalence can be established which hold under any model obtained by replacing the standard Gaussian cdf with an arbitrary continuous  $F_\circ$  in the setting of the ordinary two-sample  $t$ -test. To be more specific, let us denote by  $\beta_{F_\circ}^{m,n}(\vartheta, \sigma)$  the rejection probability of the test when applied with data such that  $X_i = \sigma X_i^\circ + \vartheta$ ,  $X_i^\circ \sim F_\circ \forall i = 1, \dots, m$ ,  $Y_j = \sigma Y_j^\circ$ ,  $Y_j^\circ \sim F_\circ \forall j = 1, \dots, n$ . Then, the following statements hold true:

- (a)  $F_\circ$  symmetric about 0 or  $m = n$

$$\Rightarrow \beta_{F_\circ}^{m,n}(-\vartheta, \sigma) = \beta_{F_\circ}^{m,n}(\vartheta, \sigma) \quad \forall (\vartheta, \sigma) \in \mathbb{R} \times \mathbb{R}_+ ;$$

- (b)  $\beta_{F_\circ}^{m,n}(\vartheta, \sigma) = \beta_{F_\circ}^{m,n}(\vartheta/\sigma, 1) \quad \forall (\vartheta, \sigma) \in \mathbb{R} \times \mathbb{R}_+ .$

The second of these results implies that in any (not only the Gaussian!) location-scale model for the two-sample setting which assumes homogeneity of the underlying distributions with respect to the scale parameter  $\sigma$ , the equivalence hypothesis which the test presented in this section is tailored for, is a statement about  $\vartheta/\sigma$  rather than  $\vartheta$  per se. This is one of several basic properties distinguishing the testing procedure under consideration from the interval inclusion procedure with Mann-Whitney based confidence limits to  $\vartheta$  as proposed by Hauschke et al. (1990). In fact, the interval estimation method forming the basis of the latter presupposes a simple location-shift model as becomes obvious from its construction (cf. Lehmann, 1963). Hence, it cannot be taken for granted that the corresponding interval inclusion test keeps being valid if the underlying nonparametric model covers also pairs of distributions differing in dispersion or other shape parameters in addition to location. Thus, apart from the fact that both approaches relate to totally different formulations of hypotheses, it is clear that the test obtained on the preceding pages is (asymptotically) distribution-free in a much stronger sense than the interval inclusion procedure based on Moses\*-Lehmann confidence limits.

---

\*The technique of computing nonparametric confidence limits for the shift in location between two continuous distributions of the same shape and dispersion by means of the Mann-Whitney null distribution has been originally introduced in a textbook chapter authored by Moses (1953, pp. 443-5).

*Mann-Whitney test for noninferiority*

If we want to test for noninferiority in the nonparametric two-sample setting (6.12) with underlying distribution functions  $F$  and  $G$  of the continuous type, this means that we have to consider the hypotheses

$$H_1 : \pi_+ \leq 1/2 - \varepsilon' \quad \text{versus} \quad K_1 : \pi_+ > 1/2 - \varepsilon'. \quad (6.21)$$

Since  $(W_+ - \pi_+)/\hat{\sigma}[W_+]$  converges in law to a standard normal variable as  $m, n \rightarrow \infty$  (under the very weak conditions made precise above in the derivation of the Mann-Whitney test for two-sided equivalence), the general result stated in §2.3 implies that an asymptotically valid test for (6.21) is given by the rejection region

$$\{(W_+ - 1/2 + \varepsilon')/\hat{\sigma}[W_+] > u_{1-\alpha}\}. \quad (6.22)$$

Its behavior in finite samples can be assessed from the simulation results presented in Table 6.8 which, except for omitting the right-hand limit of the theoretical equivalence range to the target functional, is organized analogously to Table 6.7. As in the Mann-Whitney test for two-sided equivalence, exceedances of the target significance level which might be of some concern for practical purposes, can be observed, if at all, only for the smallest of the sample sizes investigated, i.e.,  $m = n = 12$ . Under the homoskedastic Gaussian shift model, the loss in power as compared to the optimal parametric test for (6.21), is roughly the same as in the case of two-sided equivalence. With regard to (in-)sensitivity against changes in the form of the distribution functions compared, the results obtained for the noninferiority case are likewise well comparable to those shown in the previous table: The most marked differences across the models arise through allowing for heteroskedasticity which, for a variance ratio of 9, reduces the power against null alternatives by up to about 6%.

Table 6.8 *Simulated rejection probabilities of the Mann-Whitney test for noninferiority at level  $\alpha = .05$  at the boundary of the hypotheses (column 5) and  $\pi_+ = 1/2$  (rightmost column) with samples of common size  $n = 12, 24, 36$  from distributions belonging to the families listed in Table 6.6 and the same specifications of the equivalence margin  $\varepsilon'$ . [The italicized values give the power of the two-sample t-test for equivalence at level  $\alpha = .05$  against  $\pi_+ = 1/2$ .]*

Family of Distributions	Equivalence margin $\varepsilon'$	Rejection Probability				
		<i>m</i>	<i>n</i>	<i>at</i> $1/2 - \varepsilon'$	$\pi_+ =$	$\pi_+ = 1/2$
Gaussian	.2000	12	12	.04902	<i>.54167</i>	.56858
	"	24	24	.04516	<i>.78273</i>	.81853
	"	36	36	.04344	<i>.90821</i>	.92911
	.1382	12	12	.05668	<i>.34380</i>	.33276
	"	24	24	.04836	<i>.51296</i>	.53508
	"	36	36	.05036	<i>.65860</i>	.68286
	.2000	12	12	.05146	<i>.53895</i>	.58933
	"	24	24	.04846	<i>.78933</i>	.91015
	.1382	12	12	.05744	<i>.34597</i>	.52233
	"	24	24	.05315	<i>.65860</i>	.68286
Uniform	.2000	12	12	.04748	<i>.54198</i>	.58933
	"	24	24	.04348	<i>.78192</i>	.90783
	"	36	36	.04215	<i>.34687</i>	.51843
	.1382	12	12	.05607	<i>.65595</i>	.68286
	"	24	24	.05014	<i>.54265</i>	.58933
	"	36	36	.04708	<i>.78407</i>	.91012
Exponential	.2000	12	12	.04908	<i>.34678</i>	.34678
	"	24	24	.04447	<i>.51946</i>	.51946
	"	36	36	.04260	<i>.65444</i>	.65444
	.1382	12	12	.05722	<i>.51946</i>	.51946
	"	24	24	.04937	<i>.65444</i>	.65444

Table 6.8 (*continued*)

Family of Distributions	Equivalence margin $\epsilon'$	Rejection Probability				
		$m$	$n$	at $1/2 - \epsilon'$	$\pi_+ =$	$1/2$
Laplace	.2000	12	12	.05055	.54191	
	"	24	24	.04547	.78196	
	"	36	36	.04332	.90967	
	.1382	12	12	.05791	.34464	
	"	24	24	.05098	.51735	
	"	36	36	.04882	.65821	
Logistic	.2000	12	12	.04824	.53956	
	"	24	24	.04443	.78391	
	"	36	36	.04322	.90729	
	.1382	12	12	.05694	.34545	
	"	24	24	.04999	.51625	
	"	36	36	.04766	.65516	
$\mathcal{N}(\vartheta, 9)/\mathcal{N}(0, 1)$	.2000	12	12	.04966	.49628	
	"	24	24	.04424	.72957	
	"	36	36	.04410	.86861	
	.1382	12	12	.05744	.32071	
	"	24	24	.05119	.46791	
	"	36	36	.04839	.59737	

### 6.3 Two-sample equivalence tests based on linear rank statistics

In this section, general results derived in Janssen (1999, 2000) will be exploited to establish two-sample equivalence tests which are based on Wilcoxon's rank-sum and the Savage statistic, respectively. From a practical point of view, these tests are to supplement the repertoire of non- and semiparametric equivalence testing procedures which are made available in other parts of this chapter for commonly occurring two-sample problems with continuous data. The first of these tests to which the procedures considered in the present section provide a useful alternative, is the Mann-Whitney test derived in § 6.2. A semiparametric equivalence test which, like the test based on the Savage statistic, is tailored for the analysis of survival data and allows for right censoring in addition, will be presented under the heading "log-rank test for equivalence." Both of these tests are large-sample procedures so that, by construction, maintenance of the target significance level is guaranteed only asymptotically. Exact tests for interval hypotheses about distribution func-

tions satisfying the proportional hazards assumption have been considered by Munk (1996).

The general strategy for constructing linear rank tests for equivalence hypotheses is as follows: At a first stage, a specific semiparametric submodel is assumed and the critical constants for the test statistic computed from the exact distribution which the selected linear rank statistic follows under that submodel. At a second stage, the rejection probabilities of the test are studied by simulation under distributions outside the semiparametric submodel for which the construction was originally carried out, with a view to establishing that the exact computations done under the submodel provide satisfactory approximations to the operation characteristic of the test under the full nonparametric model for the two-sample setting with continuous observations.

The semiparametric model which we start from is given by the relationship

$$1 - F(t) = (1 - G(t))^\theta \quad (6.23)$$

which has to hold for all  $t \in \mathbb{R}$  and some  $\theta > 0$ . In classical nonparametric theory, each cdf  $F$  satisfying (6.23) is said to constitute a Lehmann alternative to  $G$  (see, e.g. Randles and Wolfe, 1979, § 4.3). In a survival context where both cdf's have support  $[0, \infty)$  and are assumed to be differentiable, the class of all pairs  $(F, G)$  satisfying (6.23) is called a proportional hazards or relative risk model (cf. Kalbfleisch and Prentice, 2002, § 2.3.2).

In developing the details of the approach, some additional pieces of notation will be used. The probability distributions on the real line defined by the cdf's  $F$  and  $G$  under comparison, are denoted by  $\mathcal{P}$  and  $\mathcal{Q}$ , respectively. The proversion probability  $\pi_+$  defined in (6.11) will be explicitly written as a functional of  $(\mathcal{P}, \mathcal{Q})$  setting

$$\pi_+ = \kappa_{MWW}(\mathcal{P}, \mathcal{Q}) = \int \mathcal{Q}((-\infty, t]) d\mathcal{P}(t). \quad (6.24)$$

The second functional which will be used in the sequel as a measure of dissimilarity between the distributions from which the samples  $(X_1, \dots, X_m)$  and  $(Y_1, \dots, Y_n)$  are taken, is usually named after Savage (1956). It is particularly useful in the analysis of samples consisting of positive values as obtained from survival distributions. In order to give a formal definition, let  $\Lambda_{\mathcal{P}}(t) = -\log(1 - F(t))$  denote the cumulative hazard function corresponding to  $F$ . Recall that  $\Lambda_{\mathcal{P}}(t) = \int_0^t \lambda_{\mathcal{P}}(s) ds$  holds where  $\lambda_{\mathcal{P}}(s) = \frac{f(s)}{1 - F(s)}$  is the hazard rate at  $s$  whenever  $F$  has a density  $f$ . Denoting the cumulative hazard of the second distribution under analysis by  $\Lambda_{\mathcal{Q}}$  the Savage functional is defined by

$$\kappa_{SAV}(\mathcal{P}, \mathcal{Q}) = \int \Lambda_{\mathcal{Q}}(t) d\mathcal{P}(t). \quad (6.25)$$

An equivalent representation involving the two cumulative hazards only, reads

$$\kappa_{SAV}(\mathcal{P}, \mathcal{Q}) = \int \exp(-\Lambda_{\mathcal{P}}(t)) d\Lambda_{\mathcal{Q}}(t). \quad (6.26)$$

Interestingly,  $\kappa_{SAV}(\mathcal{P}, \mathcal{Q})$  has an interpretation as the mean relative risk of the two populations. Actually, in the proportional hazards model (6.23), there holds  $\Lambda_{\mathcal{P}} = \theta \Lambda_{\mathcal{Q}}$  implying that  $\kappa_{SAV}(\mathcal{P}, \mathcal{Q}) = 1/\theta$  is exactly the (time independent) relative risk between the populations. Without the assumption of proportional hazards  $\kappa_{SAV}(\mathcal{P}, \mathcal{Q})$  is given by an average. With a view to symmetry, it is more convenient to consider the logarithmic version of (6.25), namely

$$\kappa_{LMRR}(\mathcal{P}, \mathcal{Q}) \equiv \log \int \Lambda_{\mathcal{Q}}(t) d\mathcal{P}(t) \quad (6.27)$$

which is called the logarithmic mean relative risk functional.

It is easy to see that  $\kappa_{MWW}$  and  $\kappa_{LMRR}$  are invariant with respect to strictly increasing transformations of the observations  $X_i$  and  $Y_j$ . Under the semiparametric model (6.23), both functionals are simple functions of the scalar parameter  $\theta$  only since we then have:

$$\kappa_{MWW}(\mathcal{P}, \mathcal{Q}) = 1/(1 + \theta), \quad \kappa_{LMRR}(\mathcal{P}, \mathcal{Q}) = -\log \theta. \quad (6.28)$$

Further important properties can be found in the Appendix to Janssen and Wellek (2008) where also proofs for all more technical results underlying the development presented here are given.

Put in general terms, by a linear rank test for equivalence of  $F$  and  $G$  (or  $\mathcal{P}$  and  $\mathcal{Q}$ ) with respect to some functional  $\kappa(\cdot, \cdot)$ , we mean a testing procedure which rejects the null hypothesis of relevant differences if and only if it turns out that

$$c_1 < S_N \equiv m^{-1/2} \sum_{i=1}^m a_N(R_N(X_i)) < c_2. \quad (6.29)$$

The form of the test statistic  $S_N$  appearing in this critical inequality is the same as in a linear rank test of the null hypothesis  $\mathcal{P} = \mathcal{Q}$  of the two-sided testing problem traditionally considered in deriving nonparametric tests for two independent samples. This is to say that  $R_N(X_i)$  stands for the rank of the  $i$ th observation from  $\mathcal{P}$  with respect to the pooled sample  $(X_1, \dots, Y_n)$ , and  $(a_N(1), \dots, a_N(N))$  denotes a suitable score vector to be chosen in dependence of the functional in terms of which the hypotheses are formulated. Up to now, for nonparametric equivalence problems, no results about an optimal choice of the scores are available. However, as argued by Janssen and Wellek (2008), a promising way of choosing  $a_N(\cdot)$  is to adopt exactly the same specifications as lead to rank tests of the classical null hypothesis  $\mathcal{P} = \mathcal{Q}$  which are efficient against alternatives defined in terms of the selected functional. As shown by Janssen (2000), the efficient score vectors corresponding to  $\kappa_{MWW}$  and  $\kappa_{LMRR}$  have elements

$$a_N^{MWW}(r) = \frac{r}{N+1} - 1/2 \quad (6.30a)$$

and

$$a_N^{LMRR}(r) = \sum_{q=1}^r \frac{1}{N+1-q} - 1, \quad (6.30b)$$

respectively.

Given the scores, the major problem left to solve is to determine the critical constants  $c_1, c_2$  which, of course, have to depend both on the target significance level  $\alpha$  and the equivalence region specified by the alternative hypothesis. Setting  $\phi_N \equiv I_{(c_1, c_2)}(S_N)$ , a first order approximation to the power of the equivalence test given by (6.29) is obtained by applying a central limit theorem derived by Janssen (2000). The accuracy is satisfactory for practical purposes in neighborhoods of the null model assuming identical distributions in both populations sampled, without restricting consideration to a specific semiparametric submodel. In order to state the result, let  $V_N^2$  denote the variance of the linear rank statistic  $S_N$  under  $\mathcal{P} = \mathcal{Q}$ . Furthermore, let  $\vartheta$  stand for the value taken on by an arbitrary functional  $\kappa(\cdot, \cdot)$  in its centered version  $\kappa(\mathcal{P}, \mathcal{Q}) - \kappa(\mathcal{P}, \mathcal{P})$ . Then we can write

$$E_{(\mathcal{P}, \mathcal{Q})}(\phi_N) \approx \Phi\left(\frac{c_2 - \vartheta m^{1/2} n/N}{V_N}\right) - \Phi\left(\frac{c_1 - \vartheta m^{1/2} n/N}{V_N}\right). \quad (6.31)$$

When Mann-Whitney-Wilcoxon scores are used, the null variance of  $S_N$  can be computed exactly from the equation

$$V_N^2 = \frac{n}{12(N+1)} \equiv V_{N; MWW}^2. \quad (6.32a)$$

For the normalized sum of Savage scores, the exact formula for the null variance is relatively complicated (see, e.g. Hájek et al., 1999, Eq. (4.2.1.18)) and will be replaced with the approximation

$$V_N^2 \approx \frac{n}{N} \int_0^1 \left(1 + \log(1-u)\right)^2 du = \frac{n}{N} \equiv V_{N; LMRR}^2. \quad (6.32b)$$

The validity of (6.31) for any pair  $(\mathcal{P}, \mathcal{Q})$  of underlying distributions has two major implications. Firstly, if the equivalence range for the centered functional  $\vartheta$  is chosen to be  $(-\varepsilon_1, \varepsilon_2)$ , it suggests to determine the critical constants  $c_1, c_2$  by solving the equations  $\Phi((c_2 + \varepsilon_1 m^{1/2} n/N)/V_N) - \Phi((c_1 + \varepsilon_1 m^{1/2} n/N)/V_N) = \alpha = \Phi((c_2 - \varepsilon_2 m^{1/2} n/N)/V_N) - \Phi((c_1 - \varepsilon_2 m^{1/2} n/N)/V_N)$  and to subsequently use the result as initial values for an iteration scheme based on the exact distribution of the test statistic. A second implication is that any exact solution established for a suitably chosen submodel can be used as an approximate solution for the full nonparametric model.

The key fact which allows us to construct the tests in the semiparametric proportional hazards model (6.23) as exact procedures is that, under Lehmann alternatives, the distributions of linear rank statistics are known exactly (at least in principle) and do not depend on the “baseline” distribution function  $G$ . The task of deriving formulae for the exact distribution of  $S_N$  in the proportional hazards model is greatly facilitated through representing test

statistics of that form in terms of random group indicators. For each  $r = 1, \dots, N$ , let

$$Z_{Nr} = \begin{cases} 1 & \text{if rank } r \text{ is taken on by an observation belonging to Sample 1} \\ 0 & \text{otherwise} \end{cases}. \quad (6.33)$$

Then, the linear rank statistic given by the scores  $a_N(1), \dots, a_N(N)$  can be written

$$S_N = m^{-1/2} \sum_{r=1}^N a_N(r) Z_{Nr}. \quad (6.34)$$

Furthermore, under (6.23), the exact joint distribution of the group indicators  $Z_{Nr}$  is well known (see, e.g., Savage, 1956; Kalbfleisch and Prentice, 2002, p. 106) to be given by the probability mass function

$$\begin{aligned} P_{\theta, m, n}((Z_{N1}, \dots, Z_{NN}) = (z_1, \dots, z_N)) = \\ m!n! \prod_{r=1}^N \frac{z_r + (1 - z_r)/\theta}{\sum_{q=r}^N z_q + \sum_{q=r}^N (1 - z_q)/\theta}, \\ \forall (z_1, \dots, z_N) \in \{0, 1\}^N \text{ with } \sum_{r=1}^N z_r = m. \end{aligned} \quad (6.35)$$

By means of (6.35), an exact expression for the distribution function of (6.34) can be established. Observing that any linear rank statistic is constant whenever  $N = m$  or  $N = n$  holds, yields the boundary conditions  $S_N = 0$  for  $m = 0$  and  $S_N = \sum_{r=1}^N a_N(r)$  if  $m = N$  for the following recursion formula which is very useful for computational purposes and holds under proportionality of hazards for each  $m, n \geq 1$  and  $x \in \mathbb{R}$ :

$$\begin{aligned} P_{\theta, m, n} \left( \sum_{r=1}^N a_N(r) Z_{Nr} \leq x \right) = \\ \frac{m}{m + n/\theta} P_{\theta, m-1, n} \left( \sum_{r=1}^{N-1} a_N(r+1) Z_{N-1, r} \leq x - a_N(1) \right) \\ + \frac{n/\theta}{m + n/\theta} P_{\theta, m, n-1} \left( \sum_{r=1}^{N-1} a_N(r+1) Z_{N-1, r} \leq x \right). \end{aligned} \quad (6.36)$$

As special cases of (6.36), recursion formulae for the distribution functions of the Wilcoxon and Savage statistics are obtained. For the sake of simplicity, it is advantageous to work with the sum of the ranks themselves and the scores  $a_N^{LMRR}(r) + 1$ , respectively. Using the corresponding expressions for the original test statistics  $S_N$  requires just elementary steps of rescaling. Dealing

with  $S_{N; MWW}$  first, we set  $a_N(r) = r$  so that that we have  $\sum_{r=1}^N a_N(r) = N(N+1)/2$ . Then, (6.36) simplifies to

$$\begin{aligned} P_{\theta,m,n} \left( \sum_{i=1}^m R_N(X_i) \leq x \right) &= \\ \frac{m}{m+n/\theta} P_{\theta,m-1,n} \left( \sum_{i=1}^{m-1} R_{N-1}(X_i) \leq x - m \right) \\ + \frac{n/\theta}{m+n/\theta} P_{\theta,m,n-1} \left( \sum_{i=1}^m R_{N-1}(X_i) \leq x - m \right). \end{aligned} \quad (6.36a)$$

An analogue of (6.36a) is obtained by setting  $b_N(r) = a_N^{LMRR}(r) + 1$  with  $a_N^{LMRR}(r)$  being defined as in (6.30b). These shifted Savage scores sum up to  $N$ , and the recursion formula reads

$$\begin{aligned} P_{\theta,m,n} \left( \sum_{i=1}^m b_N(R_N(X_i)) \leq x \right) &= \\ \frac{m}{m+n/\theta} P_{\theta,m-1,n} \left( \sum_{i=1}^{m-1} b_{N-1}(R_{N-1}(X_i)) \leq x - \frac{m}{N} \right) \\ + \frac{n/\theta}{m+n/\theta} P_{\theta,m,n-1} \left( \sum_{i=1}^m b_{N-1}(R_{N-1}(X_i)) \leq x - \frac{m}{N} \right). \end{aligned} \quad (6.36b)$$

The recursion formula (6.36a) is equivalent to that of Shorack (1967) for the Mann-Whitney form of the Wilcoxon statistic. A recursion for the distribution of the Savage statistic was used by Munk (1996) (see also Davies, 1971).

For a closer study of the properties of equivalence tests with rejection regions of the form (6.29), it is convenient to denote the centered version of the functional  $\kappa(\cdot, \cdot)$  for which the test statistic is tailored, by an extra symbol setting

$$\vartheta \equiv \vartheta(\mathcal{P}, \mathcal{Q}) = \kappa(\mathcal{P}, \mathcal{Q}) - \kappa(\mathcal{P}, \mathcal{P}). \quad (6.37)$$

The relationship between  $\vartheta$  and the hazard ratio  $\theta$  depends on the selected functional under consideration. For the Mann-Whitney-Wilcoxon functional  $\kappa_{MWW}$  and the log Savage functional  $\kappa_{LMRR}$ , there holds

$$\theta = \frac{1 - 2\vartheta}{2\vartheta + 1}, \quad \vartheta \in (-1/2, 1/2), \quad (6.38a)$$

and

$$\theta = \exp(-\vartheta), \quad \vartheta \in \mathbb{R}, \quad (6.38b)$$

respectively.

In both of these cases, we write  $E_\vartheta(\varphi_N) = E_{(\mathcal{P}, \mathcal{Q})}(\varphi_N)$  for the rejection probability of the test under any pair  $(P, Q)$  of underlying distributions which

satisfy (6.23) with  $\theta = \exp(-\vartheta)$ . Power computations are considerably simplified making use of the following results for the tests  $\varphi_N = \varphi_N^{MWW}$  or  $\varphi_N = \varphi_N^{LMRR}$  given by (6.29) for fixed  $c_1 < c_2$ :

- a) The function  $\vartheta \mapsto E_\vartheta(\varphi_N)$  is continuous.
- b) If there holds  $E_0(\varphi_N) > \alpha$  and  $\min S_N < c_1 < c_2 < \max S_N$ , then there exist equivalence margins  $\Delta_1 < 0 < \Delta_2$  for  $\vartheta$  such that

$$E_{\Delta_1}(\varphi_N) = E_{\Delta_2}(\varphi_N) = \alpha. \quad (6.39)$$

- c) Denoting the distribution of any real-valued statistic  $T_N$  on the sample space of  $(X_1, \dots, Y_n)$  under  $\vartheta(P, Q) = \vartheta$  by  $\mathcal{L}(T_N | \vartheta)$ , we have for each  $\vartheta$  and all  $(m, n)$

$$\mathcal{L}(S_N^{MWW} | \vartheta) = \mathcal{L}(-S_N^{MWW} | -\vartheta). \quad (6.40a)$$

In the balanced case  $m = n$ , there holds

$$\mathcal{L}(S_N^{LMRR} | \vartheta) = \mathcal{L}(-S_N^{LMRR} | -\vartheta) \quad (6.40b)$$

in addition.

All exact numerical results which will be presented below for the Mann-Whitney-Wilcoxon case have been obtained by means of a MATLAB® function implementing the recursion scheme (6.36a). Even for sample sizes of 100 each, execution time of this program on a standard Intel PC is reasonably short so that the algorithm is well suited for routine applications. The computational effort entailed in exact calculation of the distribution of the Savage statistic under arbitrary Lehmann alternatives is much higher than for the Wilcoxon statistic. Another MATLAB® program we used for determining the critical constants of the exact version of the Savage test for equivalence as well as for computing its power, is based on formula (6.35) rather than the recursion relation (6.36b). The program can conveniently be used as long as the larger of both sample sizes does not exceed 12. For larger sample sizes, we recommend to rely on the approximation obtained by applying (6.31) and check on its accuracy by means of Monte Carlo simulations. In view of the independence of the exact distribution of the baseline distribution function  $G$ , such simulations can easily be carried out by generating pseudo random observations from suitably scaled exponential distributions.

With a view to power comparisons between the Wilcoxon and the Savage test, it is essential to note that in the nonparametric framework, different functionals (scales) require different specifications of the equivalence margins. Exploiting the relationships (6.38a,b) between the hazard ratio  $\theta$  and the respective version of the centered functional  $\vartheta$ , it is easy to establish a rule for adjusting to each other the equivalence regions to be used in the Wilcoxon and the Savage test. From (6.38a), we know that if  $\vartheta_{MWW} \equiv_{MWW} -1/2$  is

chosen to be the functional of primary interest and the proportional hazards assumption holds true, the equivalence hypothesis

$$-\Delta < \vartheta_{MWW} < \Delta \quad (6.41a)$$

is the same as

$$\left| \frac{1}{1+\theta} - \frac{1}{2} \right| < \Delta \quad (6.41b)$$

for any  $\Delta > 0$ . The latter statement is in turn equivalent to  $|\log(\theta)| < \log((.5 + \Delta)/(.5 - \Delta))$ , where, according to (6.38b),  $-\log(\theta)$  coincides with the centered Savage functional in its logarithmic version. Denoting the latter by  $\vartheta_{LMRR}$ , this implies that  $\vartheta_{MWW}$  falls in the equivalence interval  $(-\Delta, \Delta)$  if and only if the distributions under comparison belong to the equivalence region

$$|\vartheta_{LMRR}| < \log\left(\frac{0.5 + \Delta}{0.5 - \Delta}\right) \quad (6.41c)$$

defined in terms of the Savage functional.

Thus, if primary interest is in assessing equivalence in terms of the Mann-Whitney-Wilcoxon functional and the equivalence margin for  $\vartheta_{MWW}$  is chosen to be  $\Delta = 0.2$  [recall Table 1.1 (ii)], it follows that a semi-parametric competitor to the Wilcoxon rank-sum test for equivalence is given by the Savage test of  $|\vartheta_{LMRR}| \geq \log(7/3)$  versus  $|\vartheta_{LMRR}| < \log(7/3) = 0.847$ . Comparing both tests in terms of their power functions leads to the couple of graphs shown in Figure 6.1 relating to the usual specification  $\alpha = .05$  of the (nominal) significance level and samples of size  $m = n = 12$ . Not surprisingly (in view of the local optimality of the Savage test for the traditional one-sided testing problem about two distribution functions satisfying the proportional hazards model — see, e.g., Hájek et al., 1999, p. 106), the Wilcoxon test turns out uniformly inferior with respect to power. Furthermore, the power function of the Wilcoxon test fails to cross the level- $\alpha$  line exactly at the boundary points of the equivalence range which, within the limited range of resolution provided by the graph, cannot be seen for the Savage test. This reflects the obvious fact that, as compared to the Savage statistic, the distribution of the simple sum of ranks exhibits much coarser discreteness.

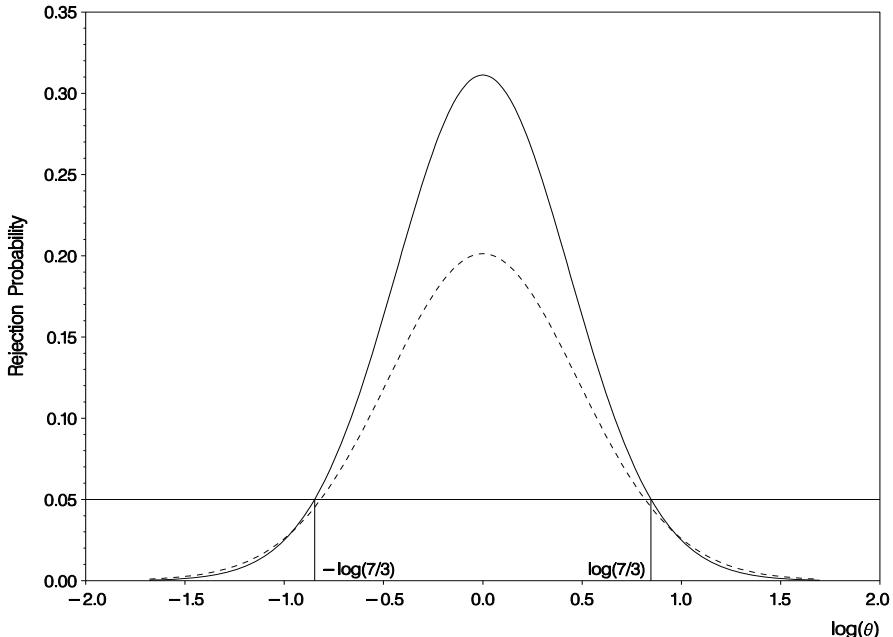


Figure 6.1 *Exact power function of the Wilcoxon [dashed] and the Savage test for equivalence [solid line] at level  $\alpha = .05$  for two samples of sizes  $m = n = 12$  from distributions satisfying a proportional hazards model.*

### Illustrations

In order to demonstrate the use of the linear rank tests discussed in this subsection, we first revert once more to the *data set shown in Table 6.3*. Replacing the individual observations with their ranks with respect to the combined sample yields:

*Moxonodin:*

8; 23; 14; 19; 2; 16; 11; 7; 12; 9; 24; 17

*Captopril:*

18; 20; 6; 1; 13; 15; 10; 5; 3; 4; 22; 21.

Summing up all ranks associated with the observations making up the first sample gives  $\tilde{S}_N^{MWW} \equiv \sum_{i=1}^m R_N(X_i) = 138$ . As follows from (6.39) and (6.40a), choosing a theoretical equivalence range to  $\vartheta_{MWW}$  of the form (6.41a) implies that the critical interval of the exact test for equivalence based on  $\tilde{S}_N^{MWW}$  is in turn symmetric about the expected value of  $\sum_{i=1}^m R_N(X_i)$  under the traditional null hypothesis  $P = Q$ , with limits  $m(N + 1)/2 \mp c$ . In view

of (6.31), (6.32a) and (4.26), the critical constant  $c$  is approximately given by

$$c \approx mn\Delta - \sqrt{mn(N+1)/12} u_{1-\alpha}. \quad (6.42)$$

For  $m = n = 12$ ,  $\Delta = .20$  and  $\alpha = .05$ , this yields  $c \approx .3078$ . Taking the smallest integer being at least as large as this approximate value of  $c$  as an initial value for iterative determination of the exact  $c$  yields after a few runs of the MATLAB® function mentioned before,  $c = 5$ . Since 138 falls outside the corresponding critical interval (145, 155), the exact Wilcoxon test for equivalence cannot reject the null hypothesis of relevant differences between both underlying distribution functions.

The discrepancy between the approximate and the exact value of  $c$  should not be overestimated. First of all, sample sizes are obviously too small for expecting high accuracy of results based on first order asymptotics. Furthermore, it must be kept in mind that even for  $m = n = 12$ , the range of  $\tilde{S}_N^{MWW}$  is fairly large (with bounds 78 and 222) so that an approximation error of about 5 does not matter too much on that scale.

Replacing unit with Savage scores, we obtain from the ranks obtained in the moxonodin arm of the trial  $\tilde{S}_N^{LMRR} \equiv \sum_{i=1}^m b_N(R_N(X_i)) = 10.3242$ . In order to determine the limits of the critical interval to be used in the exact test based on  $\tilde{S}_N^{LMRR}$ , we proceed in essentially the same way as in the exact Wilcoxon test for equivalence. Again, for a symmetric specification of the two equivalence margins, there is only a single critical constant  $c$  to be determined from the distribution of  $\sum_{i=1}^m b_N(R_N(X_i))$  at one of the boundary points of the equivalence hypothesis (6.41c) we are interested in. This time, the first order approximation based on (6.31) and (6.32) yields

$$c \approx \frac{mn}{N} \log \left( \frac{0.5 + \Delta}{0.5 - \Delta} \right) - \sqrt{\frac{mn}{N}} u_{1-\alpha}. \quad (6.43)$$

Plugging in the same values as before in the expression on the right-hand side of (6.43) yields  $c = 1.0547$  as initial estimate of the exact critical constant. Repeated runs of the computational procedure described above show that this initial value has to be diminished to  $c = 0.9766$  in order to make sure that the test is carried out in a way maintaining the 5% level exactly. Since the observed value of  $\sum_{i=1}^m b_N(R_N(X_i))$  is smaller than the right-hand limit of the corresponding critical interval  $(m - c, m + c) = (11.0234, 12.9766)$ , the Savage test fails to lead to a positive decision either.

As a second example, we analyze a subset of *Kalbfleisch and Prentice's (2002, pp. 378–9) Data Set I* referring to a clinical trial conducted by the Veterans Administration Lung Cancer Study Group of the U.S. In both arms of this trial, patients were in particular stratified with regard to pretreatment status. We suppose that a preliminary test for equivalence of patients with and without previous treatment shall be performed in order to make sure that non-stratified randomization and the assumption of a common treatment effect for the whole study population were appropriate. Discarding a few patients

who provided censored values and restricting the analysis to the standard chemotherapy arm, there were  $m = 20$  and  $n = 44$  patients with and without previous treatment, respectively. In order to avoid technical complications, a few ties contained in the data set were broken at random. The empirical survival functions obtained from these data are shown in Figure 6.2.

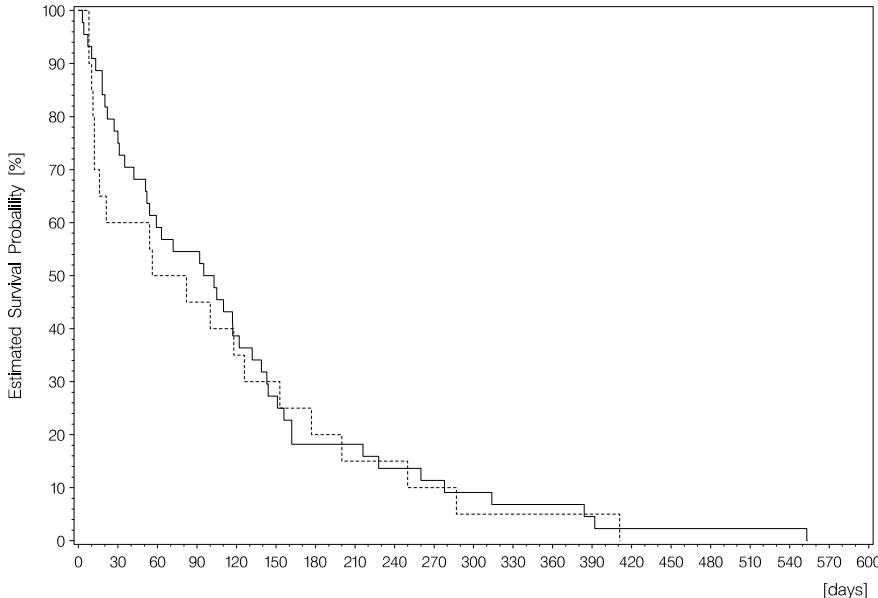


Figure 6.2 *Empirical survival functions observed in lung-cancer patients with [dashed] and without previous therapy [solid line].*

The set of ordered ranks of the survival times making up the first of these two samples [ $\leftrightarrow$  pretreatment yes] are:

$$\begin{array}{cccccccccc} 4; & 5; & 6; & 8; & 9; & 10; & 12; & 16; & 25; & 27; \\ 31; & 34; & 40; & 42; & 48; & 52; & 53; & 56; & 59; & 63. \end{array}$$

The corresponding values of the (noncentered) Wilcoxon and Savage statistics are computed to be

$$\tilde{S}_N^{MWW} = 600 \quad \text{and} \quad \tilde{S}_N^{LMRR} = 18.6004.$$

Using the same specification of the equivalence margin  $\Delta$  as before, Equation (6.42) gives  $c = 62.4275$  as an approximate value of the critical constant for the

Wilcoxon test for equivalence. The computational scheme for determining the exact critical interval for  $\tilde{S}_N^{MWW}$  described in the previous example does not have to rely on the assumption that a balanced study design is used. Hence, we can proceed as before using 63 as the initial value in an iterative search for the largest integer  $c$  such that the exact value of  $P_\theta [m(N+1)/2 - c < \tilde{S}_N^{MWW} < m(N+1)/2 + c]$  at  $\theta \in \{3/7, 7/3\}$  does not exceed  $\alpha = .05$ . In this way, we obtain  $c = 76$  corresponding to a critical interval of (576, 724) for  $\tilde{S}_N^{MWW}$ . Since the latter covers the observed value of the sum of ranks for the pretreatment group the decision is in favor of equivalence.

In order to compute the critical interval for the sum of Savage scores of the observations making up the first sample, we start with applying formula (6.43) which gives  $c = 5.5505$  as an approximate value of the distance of both endpoints from  $m = 20$ . However, in the unbalanced case the exact critical interval for  $\tilde{S}_N^{LMRR}$  is no longer symmetric about its value expected under the assumption of identical populations. Thus, we have to determine two different integers  $c_1$  and  $c_2$  such that the probability content of  $(m - c_1, m + c_2)$  comes as close as possible to  $\alpha$  both under  $\theta = 3/7$  and  $\theta = 7/3$ . A heuristic algorithm for determining  $(c_1, c_2)$  is as follows: In a first step, starting from the large-sample solution,  $c_2$  is determined as the largest positive number such that there holds  $P_\theta [m - c_2 < \tilde{S}_N^{LMRR} < m + c_2] \leq \alpha$  under  $\theta = 3/7$ . From a set of Monte Carlo simulation experiments of the kind described above [→ p. 142], it has been found that the initial approximation has to be replaced with  $c_2 = 4.98$  in order to meet this condition. In order to ensure that the significance level is kept also under  $\theta = 7/3$ , the left-hand critical limit has to be increased from  $m - c_2$  to  $m - c_1$  with  $c_1 = 4.86$ , implying that the critical interval for the exact Savage test is estimated to be  $(m - c_1, m + c_2) = (15.14, 24.98)$ . Again, the observed value of the test statistic is included so that the qualitative decision is the same as has to be taken in the Wilcoxon test for equivalence with the same data.

#### *Robustness/sensitivity of the exact tests against violations of the proportional hazards assumption*

Table 6.9 shows some simulation results on the level and power of the exact linear rank test for equivalence with Wilcoxon scores, under two common location-shift models. For ease of comparison, the corresponding entries in Table 6.7 obtained under the same specifications for the asymptotic test derived in § 6.2, are added as italicized values.

Table 6.9 *Simulated level and power of the exact rank-sum test for equivalence under Gaussian and Cauchy shift models for the same sizes and specification of the equivalence margins as in the corresponding part of Table 6.7. [The italicized values are taken from the latter, showing the rejection probabilities of the asymptotic Mann-Whitney-Wilcoxon test of Section 6.2 under the same specifications.]*

Family of Distributions	<i>m</i>	<i>n</i>	Rejection Probability at $\vartheta =$		
			−0.20	+0.20	0.00
Gaussian	12	12	.04568	.04580	.20089
			<i>.05015</i>	<i>.04941</i>	<i>.22020</i>
		24	.04851	.04785	.59286
	36		<i>.04510</i>	<i>.04565</i>	<i>.57929</i>
		36	.04894	.04958	.83373
			<i>.04427</i>	<i>.04476</i>	<i>.82186</i>
Cauchy	12	12	.04796	.04767	.20363
			<i>.04925</i>	<i>.04780</i>	<i>.20180</i>
		24	.05191	.05465	.59412
	36		<i>.04644</i>	<i>.04536</i>	<i>.56918</i>
		36	.05425	.05466	.83441
			<i>.04466</i>	<i>.04442</i>	<i>.81617</i>

In order to study the behavior of the Savage test for equivalence when the two distribution functions under comparison do not satisfy the semiparametric submodel (6.23), an appropriate choice of a model for generating the data by Monte Carlo simulation is the lognormal distribution. In principle, the log transformation is dispensible since both the functional in terms of which the hypotheses are defined, and the test statistics remain invariant under monotone transformations of any kind applied to each observation. But Savage scores are commonly used in the nonparametric analysis of survival data, and the lognormal distribution is among the most popular survival models with nonproportional hazards. Under nonproportional hazards, one has to refer to the general definition (6.27) of the log Savage functional. In the simulation experiments whose results are summarized in Table 6.10, both distributions of log survival times had unit variance, and the expected values were determined through solving the equation  $\int \Lambda_Q(t)dP(t) = \theta^*$  with  $\theta^* = 7/3$  [ $\leftrightarrow$  left column],  $3/7$  [ $\leftrightarrow$  middle column], and  $1$  [ $\leftrightarrow$  right column], respectively.

Table 6.10 *Level and power of the exact Savage test for equivalence under the lognormal model. Distribution underlying Sample 1 ( $\mathcal{P}$ ) :  $\mathcal{LN}(\mu, 1)$ ,  $\mu \in \{-0.8444, 0.0000, 1.0630\}$ ; baseline distribution ( $\mathcal{Q}$ ) :  $\mathcal{LN}(0, 1)$ . [The three values of  $\mu$  correspond to values  $7/3$ ,  $1$ , and  $3/7$  of the generalized Savage functional as defined in (6.26-7).]*

$m$	$n$	Rejection Probability		
		at $\mu =$		
		1.0630	-0.8444	0.0000
12	12	.0241	.0633	.3129
24	24	.0161	.0786	.7560
36	36	.0127	.0760	.9279

The simulation results shown in both tables admit the following conclusions:

- In the Gaussian shift model, the performance of the MWU rank-sum test with critical constants determined under the proportional hazards model is very satisfactory: the maximum rejection probability under the null hypothesis is slightly smaller than the nominal level throughout.
- Replacing the Gaussian with the Cauchy distribution leads to a slight anticonservatism of the MMW test for equivalence whose extent seems tolerable for most practical purposes.
- The Savage test turns out to be markedly more sensitive to violations of proportionality of hazards; its application in settings where restriction to this semiparametric model seems unrealistic must be discouraged.

#### *Noninferiority versions of the exact semiparametric tests*

If the alternative hypothesis one aims to establish is given by the class of all pairs  $(\mathcal{P}, \mathcal{Q})$  for which the centered functional  $\vartheta$  of interest satisfies  $-\Delta < \vartheta < \infty$ , the natural form of the critical function of a suitable test is

$$\varphi_N^* = I_{(c^*, \infty)}(S_N) \quad (6.44)$$

being the indicator function of the noninferiority analogue of a critical region of the form (6.29).

The power function of  $\varphi_N^*$  can be approximated (in the same sense as made precise above for the case of two-sided equivalence [recall the comments on

(6.31)] by means of the formula

$$E_{(\mathcal{P}, \mathcal{Q})}(\phi_N^*) \approx 1 - \Phi\left(\frac{c^* - \vartheta m^{1/2} n/N}{V_N}\right). \quad (6.45)$$

For the practical implementation of the noninferiority versions of the linear rank tests considered above, (6.45) plays the same role as (6.31) in testing for equivalence in the strict sense. I.e., it can be used for finding a good initial estimate of the exact critical lower bound to the linear rank statistic under consideration. In the noninferiority case, the exact critical constant has to be determined as the  $(1 - \alpha)$ -quantile of the distribution of  $S_N$  under  $\vartheta = -\Delta$ . As in the case of two-sided equivalence, (6.45) is in particular a convenient starting point for computing the exact critical lower bound to the linear rank statistic for a test for noninferiority at prespecified level  $\alpha$  in the proportional hazards model. In practice, this leads to using the same initial values for computing the exact value of  $c^*$  as were used in the equivalence case for computing the left-hand bound  $c_1$  of the critical interval. Correspondingly, fixing the left-hand limit of the equivalence range for  $\vartheta$  at the same value as in a two-sided equivalence problem relating to the same functional, the exact values of  $c^*$  and  $c_1$  can be expected to differ only slightly.

Revisiting briefly the first of the examples used above for illustration, we find this expectation largely confirmed. As before, let us specify  $\Delta = .20$  and  $\Delta = \log(7/3) = .847$  for the test relating to  $\vartheta_{MWW}$  and  $\vartheta_{LMRR}$ , respectively. With these values of the equivalence margin, the critical lower bound to  $\tilde{S}_N^{MWW}$  and  $\tilde{S}_N^{LMRR}$  to be used in the Wilcoxon and the Savage test for noninferiority at exact level  $\alpha = .05$  is computed to be  $c_{MWW}^*(\alpha) = 148$  and  $c_{LMRR}^*(\alpha) = 11.2304$ , respectively. On the other hand, the power against the specific alternative  $\vartheta = 0$  being of primary interest for the majority of practical applications, increases to  $.5338$  [ $\leftrightarrow$  MMW] and  $.6236$  [ $\leftrightarrow$  LMRR] when proceeding from the equivalence to the noninferiority version of the tests. Moreover, the power function as a whole is no longer concave from below but monotonically increasing (decreasing) in  $\vartheta$  ( $\theta$ ) and takes every value in the open unit interval, as is generally the case with tests of one-sided hypotheses.

## 6.4 A distribution-free two-sample equivalence test allowing for arbitrary patterns of ties

The assumptions underlying the construction of a nonparametric two-sample test for equivalence described in § 6.2 imply that the occurrence of ties between observations from different samples can be excluded with probability one. Accordingly, the procedure is unsuitable for analyzing data sets made up of independent observations from two different discrete distributions.

In this section we show how to generalize the Mann-Whitney test in such a way that arbitrary patterns of ties can be admitted without affecting the (asymptotic) validity of the procedure. In elaborating this idea, basic analogies with the derivation of a generalized signed rank test for equivalence dealt with in § 5.5 will become apparent. Obviously, the definition of the functional  $\pi_+ [ \rightarrow (6.11), (6.24) ]$  makes sense for any type of distributions which the  $X$ 's and  $Y$ 's are taken from. However, in the noncontinuous case, identity of both distributions no longer implies that  $\pi_+$  takes on value 1/2. Instead, in the general case, the relationship  $F = G \Rightarrow \pi_+ = 1/2$  has to be replaced with

$$F = G \Rightarrow \pi_+ = (1/2) \cdot (1 - \pi_0) \quad (6.46)$$

provided we define  $\pi_0$  as the functional

$$\pi_0 = P[X_i = Y_j] = \int (G - G_-) dF \quad (6.47)$$

assigning each pair of cdf's as its value the probability of a tie between an observation from  $F$  and an observation from  $G$  independent of the former. (On the right-hand side of the second of the equalities (6.47),  $G_-$  stands for the left-continuous version of the cdf  $G$  as defined by  $G_-(y) = P[Y_j < y] \quad \forall y \in \mathbb{R}$ .) In view of (6.46), it seems natural to use the distance of  $\pi_+/(1 - \pi_0)$  from the point 1/2 as a measure for the degree of disparity of any two cdf's  $F$  and  $G$  from which our two samples  $(X_1, \dots, X_m)$  and  $(Y_1, \dots, Y_n)$  are taken. Adopting this as our basic measure of distance, we are led to formulate our testing problem as

$$\begin{aligned} H : \pi_+/(1 - \pi_0) &\leq 1/2 - \varepsilon'_1 \quad \text{or} \quad \pi_+/(1 - \pi_0) \geq 1/2 + \varepsilon'_2 \\ \text{versus} \quad K : 1/2 - \varepsilon'_1 &< \pi_+/(1 - \pi_0) < 1/2 + \varepsilon'_2. \end{aligned} \quad (6.48)$$

For both of the basic functionals appearing in (6.48), there exists an  $U$ -statistic estimator with a kernel which is just the indicator function of the respective event. Of course, the first of these estimators (for  $\pi_+$ ) is the same as that introduced in the previous section [recall (6.13)]. Analogously, a  $U$ -statistic estimating  $\pi_0$  is given by

$$W_0 = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n I_{\{0\}}(X_i - Y_j), \quad (6.49)$$

with  $I_{\{0\}}(u) = 1$  for  $u = 0$  and  $I_{\{0\}}(u) = 0$  for  $u \in (-\infty, 0) \cup (0, \infty)$ . Furthermore, it is natural to estimate the target functional in terms of which the hypotheses under assessment now have been formulated, by just plugging in (6.13) and (6.49) into the expression  $\pi_+/(1 - \pi_0)$ . Denoting this latter estimator by  $Q$ , we have by definition

$$Q = W_+/(1 - W_0). \quad (6.50)$$

Now, we proceed by studying the large-sample distribution of (6.50) which is readily obtained as soon as we know the joint asymptotic distribution of the two individual  $U$ -statistics involved. Assuming in the sequel that the sequence  $(m/N)_{N \in \mathbb{N}}$  of relative sizes of the first sample converges to some nondegenerate limit  $\lambda \in (0, 1)$ , say, it follows from the asymptotic distribution theory for two-sample  $U$ -statistics (see, e.g., Randles and Wolfe, 1979, §3.6) that  $\sqrt{N}(W_+ - \pi_+, W_0 - \pi_0)$  converges in law to a centered bivariate normal distribution. The covariance matrix  $\Sigma_N = \begin{pmatrix} \sigma_{+N}^2 & \sigma_{+0;N} \\ \sigma_{+0;N} & \sigma_{0N}^2 \end{pmatrix}$ , say, of this distribution can be computed exactly. The exact formulae contain also terms of order  $O(1/N)$ :

(i) Variance of  $\sqrt{N}(W_+ - \pi_+)$ :

$$\sigma_{+N}^2 = \frac{N}{mn} [\pi_+ - (N-1)\pi_+^2 + (m-1)\Pi_{XXY} + (n-1)\Pi_{XYY}] \quad (6.51a)$$

where

$$\Pi_{XXY} = P[X_{i_1} > Y_j, X_{i_2} > Y_j], \quad (6.51b)$$

$$\Pi_{XYY} = P[X_i > Y_{j_1}, X_i > Y_{j_2}]. \quad (6.51c)$$

(ii) Variance of  $\sqrt{N}(W_0 - \pi_0)$ :

$$\sigma_{0N}^2 = \frac{N}{mn} [\pi_0 - (N-1)\pi_0^2 + (m-1)\Psi_{XXY} + (n-1)\Psi_{XYY}] \quad (6.52a)$$

where

$$\Psi_{XXY} = P[X_{i_1} = Y_j, X_{i_2} = Y_j], \quad (6.52b)$$

$$\Psi_{XYY} = P[X_i = Y_{j_1}, X_i = Y_{j_2}]. \quad (6.52c)$$

(iii) Covariance of  $\sqrt{N}(W_+ - \pi_+)$  and  $\sqrt{N}(W_0 - \pi_0)$ :

$$\sigma_{+0;N} = \frac{N}{mn} [(m-1)\Lambda_{XXY} + (n-1)\Lambda_{XYY} - (N-1)\pi_+\pi_0] \quad (6.53a)$$

where

$$\Lambda_{XXY} = P[X_{i_1} > Y_j, X_{i_2} = Y_j], \quad (6.53b)$$

$$\Lambda_{XYY} = P[X_i > Y_{j_1}, X_i = Y_{j_2}]. \quad (6.53c)$$

Now let us define the function  $q(\cdot)$  by  $q(\pi_+, \pi_0) = \pi_+/(1 - \pi_0)$  and denote the result of evaluating the quadratic form  $\nabla q(\pi_+, \pi_0)\Sigma_N(\nabla q(\pi_+, \pi_0))'$  [with  $\nabla q(\pi_+, \pi_0)$  as the gradient (row) vector of  $q(\cdot)$ ] by  $\nu_N^2$ . Then, after straightforward calculations and rearrangements of terms we obtain the expression:

$$\nu_N^2 = \frac{\sigma_{+N}^2}{(1 - \pi_0)^2} + \frac{\pi_+^2 \sigma_{0N}^2}{(1 - \pi_0)^4} + \frac{2\pi_+ \sigma_{+0;N}}{(1 - \pi_0)^3}. \quad (6.54)$$

Resorting to the so-called  $\delta$ -method (cf. Bishop et al., 1975, §14.6) once more,

we can conclude that  $\sqrt{N}(Q - \pi_+/(1-\pi_0)) = \sqrt{N}(q(W_+, W_0) - q(\pi_+, \pi_0))$  converges weakly to a random variable following a centered normal distribution with variance  $\nu^2 = \lim_{N \rightarrow \infty} \nu_N^2$ .

The first step of a natural approach to deriving an estimator for  $\nu_N^2$  consists in plugging in  $U$ -statistic estimators of all functionals of the underlying cdf's  $F$  and  $G$  appearing on the right-hand side of equations (6.51)–(6.53). The  $U$ -statistic estimators for the functionals indexed by subscript  $XXY$  are given by

$$\begin{aligned} \widehat{\Pi}_{XXY} &= \frac{2}{m(m-1)n} \sum_{i_1=1}^{m-1} \sum_{i_2=i_1+1}^m \sum_{j=1}^n \left[ I_{(0,\infty)}(X_{i_1} - Y_j) \cdot \right. \\ &\quad \left. I_{(0,\infty)}(X_{i_2} - Y_j) \right] \end{aligned} \quad (6.55a)$$

$$\begin{aligned} \widehat{\Psi}_{XXY} &= \frac{2}{m(m-1)n} \sum_{i_1=1}^{m-1} \sum_{i_2=i_1+1}^m \sum_{j=1}^n \left[ I_{\{0\}}(X_{i_1} - Y_j) \cdot \right. \\ &\quad \left. I_{\{0\}}(X_{i_2} - Y_j) \right] \end{aligned} \quad (6.55b)$$

$$\begin{aligned} \widehat{\Lambda}_{XXY} &= \frac{1}{m(m-1)n} \sum_{i_1=1}^m \sum_{i_2 \in \{1, \dots, m\} \setminus \{i_1\}} \sum_{j=1}^n \left[ I_{(0,\infty)}(X_{i_1} - Y_j) \cdot \right. \\ &\quad \left. I_{\{0\}}(X_{i_2} - Y_j) \right]. \end{aligned} \quad (6.55c)$$

The analogous estimators of  $\Pi_{XYY}$ ,  $\Psi_{XYY}$  and  $\Lambda_{XYY}$  admit the representation:

$$\begin{aligned} \widehat{\Pi}_{XYY} &= \frac{2}{m(n-1)n} \sum_{i=1}^m \sum_{j_1=1}^{n-1} \sum_{j_2=j_1+1}^n \left[ I_{(0,\infty)}(X_i - Y_{j_1}) \cdot \right. \\ &\quad \left. I_{(0,\infty)}(X_i - Y_{j_2}) \right] \end{aligned} \quad (6.56a)$$

$$\begin{aligned} \widehat{\Psi}_{XYY} &= \frac{2}{m(n-1)n} \sum_{i=1}^m \sum_{j_1=1}^{n-1} \sum_{j_2=j_1+1}^n \left[ I_{\{0\}}(X_i - Y_{j_1}) \cdot \right. \\ &\quad \left. I_{\{0\}}(X_i - Y_{j_2}) \right] \end{aligned} \quad (6.56b)$$

$$\begin{aligned} \widehat{\Lambda}_{XYY} &= \frac{1}{m(n-1)n} \sum_{i=1}^m \sum_{j_1=1}^n \sum_{j_2 \in \{1, \dots, n\} \setminus \{j_1\}} \left[ I_{(0,\infty)}(X_i - Y_{j_1}) \cdot \right. \\ &\quad \left. I_{\{0\}}(X_i - Y_{j_2}) \right]. \end{aligned} \quad (6.56c)$$

Since each of the estimators defined in (6.13), (6.49), (6.55) and (6.56) is (strongly) consistent for its expectation (cf. Lee, 1990, §3.4.2), it follows that the plug-in estimator  $\hat{\sigma}_{+N}^2$  obtained from (6.51a) by replacing  $\pi_+$ ,  $\Pi_{XXY}$  and  $\Pi_{XYY}$  with  $W_+$ ,  $\hat{\Pi}_{XXY}$  and  $\hat{\Pi}_{XYY}$ , respectively, is consistent for  $\sigma_{+N}^2$ , and so on for  $\hat{\sigma}_{0N}^2$  and  $\hat{\sigma}_{+0;N}$ . Finally, consistency of  $W_+$ ,  $W_0$ ,  $\hat{\sigma}_{+N}^2$ ,  $\hat{\sigma}_{0N}^2$  and  $\hat{\sigma}_{+0;N}$  ensures that the asymptotic variance  $\nu_N^2$  of the statistic  $\sqrt{NW_+}/(1 - W_0)$  (and hence *a fortiori* its limiting variance  $\nu^2$ ) can be consistently estimated by

$$\hat{\nu}_N^2 = \frac{\hat{\sigma}_{+N}^2}{(1 - W_0)^2} + \frac{W_+^2 \hat{\sigma}_{0N}^2}{(1 - W_0)^4} + \frac{2W_+ \hat{\sigma}_{+0;N}}{(1 - W_0)^3}. \quad (6.57)$$

What is left to do then in order to complete the construction of a Mann-Whitney test for equivalence allowing for arbitrary patterns of ties in the data, is to apply the general approach of §3.4 to the specific case that  $k = 2$ ,  $T_N = W_+/(1 - W_0)$ ,  $\theta = \pi_+/(1 - \pi_0)$ ,  $\sigma = \nu$  and  $\hat{\tau}_N = \hat{\nu}_N/\sqrt{N}$ . Accordingly, we end up with the decision rule

$$\text{Reject } H : \frac{\pi_+}{1 - \pi_0} \leq \frac{1}{2} - \varepsilon'_1 \text{ or } \frac{\pi_+}{1 - \pi_0} \geq \frac{1}{2} + \varepsilon'_2 \quad \text{if and only if} \\ \sqrt{N} \left| W_+/(1 - W_0) - (1 - \varepsilon'_1 + \varepsilon'_2)/2 \right| / \hat{\nu}_N < C_{MW}^*(\alpha; \varepsilon'_1, \varepsilon'_2), \quad (6.58)$$

with

$$C_{MW}^*(\alpha; \varepsilon'_1, \varepsilon'_2) = \left\{ \begin{array}{l} \text{100}\alpha-\text{percentage point of the } \chi^2 \text{ - distri-} \\ \text{bution with } df = 1 \text{ and } \lambda_{nc}^2 = N(\varepsilon'_1 + \varepsilon'_2)^2 / 4\hat{\nu}_N^2 \end{array} \right\}^{1/2} \quad (6.59)$$

By Theorem A.3.4, it is a fact that the corresponding test is asymptotically valid over the whole class of all pairs  $(F, G)$  of distribution functions on the real line such that the limiting variance  $\nu^2$  of  $\sqrt{NW_+}/(1 - W_0)$  does not vanish. Inspecting specific families of discrete distributions corroborates the impression that this regularity condition is actually very mild. Even in the extreme case that both the  $X_i$  and the  $Y_j$  are Bernoulli variables, the requirement that  $\nu^2$  has to be positive rules out merely those constellations in which at least one of the two underlying distributions is degenerate on its own.

In order to reduce the practical effort entailed by carrying out the generalized Mann-Whitney test for equivalence to a minimum, another special computer program is provided at the URL accompanying this book, again both as a SAS macro and a R function. The program for which the name `mwtie_xy` has been chosen, performs all necessary computational steps from scratch, i.e., upon just providing the  $m + n$  raw data values as input.

### Computational methods for grouped data

Suppose that the set of values taken on by any of the observations in the pooled sample is given by  $\{w_1, \dots, w_K\}$  such that  $w_1 < w_2 < \dots < w_K$  and the number  $K$  of different groups of values is small compared to both sample sizes. Then, computations can be made considerably faster by reducing the raw data to group frequencies before calculating the various counting statistics involved. For elaborating on this idea, we need some additional notation.

Let  $M_k$  and  $N_k$  be the number of  $X$ 's and  $Y$ 's, respectively, taking on value  $w_k$  ( $k = 1, \dots, K$ ). Furthermore, define corresponding cumulative frequencies  $M_k^c$  and  $N_k^c$  by  $M_0^c = N_0^c = 0$ ,  $M_k^c = \sum_{l=1}^k M_l$ ,  $N_k^c = \sum_{l=1}^k N_l$ ,  $k = 1, \dots, K$ . Then, it is easy to verify that the individual  $U$ -statistic estimators required for computing the test statistic of (6.58) admit the following representations:

$$W_+ = \frac{1}{mn} \sum_{k=1}^K M_k N_{k-1}^c , \quad W_0 = \frac{1}{mn} \sum_{k=1}^K M_k N_k ; \quad (6.60)$$

$$\hat{\Pi}_{XXY} = \frac{1}{m(m-1)n} \sum_{k=1}^K (m - M_k^c)(m - M_k^c - 1) N_k , \quad (6.61a)$$

$$\hat{\Psi}_{XXY} = \frac{1}{m(m-1)n} \sum_{k=1}^K M_k (M_k - 1) N_k , \quad (6.61b)$$

$$\hat{\Lambda}_{XXY} = \frac{1}{m(m-1)n} \sum_{k=1}^K M_k (m - M_k^c) N_k ; \quad (6.61c)$$

$$\hat{\Pi}_{XYY} = \frac{1}{mn(n-1)} \sum_{k=1}^K M_k N_{k-1}^c (N_{k-1}^c - 1) , \quad (6.62a)$$

$$\hat{\Psi}_{XYY} = \frac{1}{mn(n-1)} \sum_{k=1}^K M_k N_k (N_k - 1) , \quad (6.62b)$$

$$\hat{\Lambda}_{XYY} = \frac{1}{mn(n-1)} \sum_{k=1}^K M_k N_k N_{k-1}^c . \quad (6.62c)$$

The above identities give rise to a simplified version of the code for the procedure to be found in the **WKTSEQ2 Source Code Package** under the program name **mwtie\_fr**. The program likewise processes the primary set of raw data. The (cumulative) frequencies appearing in expressions for the  $U$ -statistics involved are determined automatically and need not be available as entries in a user-supplied  $2 \times K$  contingency table.

*Example 6.2*

In a study of a N-methyl-D-aspartate receptor 2B (NR2B) gene variant as a possible risk factor for alcohol dependence (Schumann et al., 2003), a single nucleotide polymorphism (SNP) located at position 2873 of the gene was used in genotyping  $m = 204$  patients and  $n = 258$  unrelated healthy controls. The polymorphism leads to a C (cytosine) to T (thymine) exchange. Table 6.11 shows the frequencies of the 3 possible genotypes found in both samples.

Table 6.11 *Observed distribution of SNP C2873T NR2B genotypes in patients with alcohol dependence and healthy controls.*

Group	Genotype			$\sum$
	CC	CT	TT	
Patients	103 (50.5)	84 (41.2)	17 (8.3)	204 = $m$ (100.0)
Controls	135 (52.3)	105 (40.7)	18 (7.0)	258 = $n$ (100.0)

Since the location of the mutation was suspected to be outside the gene region encoding the domain of an antagonist of the NMDA receptor assumed to play a major role in mediating the reinforcing effects of alcohol abuse, the authors of the study aimed at excluding the existence of a substantial association between SNP C2873T NR2B genotype and alcohol dependence.

In a traditional study of this type being launched with the intention to establish the association, the standard inferential procedure is Armitage's (1955) test for an up- or downward trend in a  $2 \times K$  contingency table with ordered column categories and fixed row margins (for more recent reviews of this method focusing on applications to genetic epidemiology see Sasieni, 1997; Devlin and Roeder, 1999). Experience shows that the decision to which this procedure leads rarely differs from that to be taken in a Mann-Whitney test corrected for the large number of ties occurring in a set of data taking on only three different values, namely 0, 1 and 2 (giving simply the number of mutated alleles in an individual's genotype). This fact suggests analyzing the data set underlying Table 6.11 by means of the equivalence counterpart of the tie-corrected Mann-Whitney test derived above.

Let the significance level be chosen as usual, i.e.,  $\alpha = .05$ , and fix the equivalence limits to  $\pi_+/(1 - \pi_0) - 1/2$  at  $\mp .10$ . Running the grouped-data version `mwtie_fr` of the respective program we obtain for the genotype frequencies shown above the following estimates:

$$W_+ = .29298, \quad W_0 = .43759, \quad \hat{\nu}_N = .918898.$$

With these values and tolerances  $\varepsilon'_1 = \varepsilon'_2 = .10$ , the test statistic to be used according to (6.58) is computed to be

$$\frac{\sqrt{N} |W_+/(1 - W_0) - (1 - \varepsilon'_1 + \varepsilon'_2)/2|}{\hat{\nu}_N} = \frac{\sqrt{462} |.52093 - .5|}{.918898} = .489579.$$

As the critical upper bound which  $\sqrt{N} |W_+/(1 - W_0) - (1 - \varepsilon'_1 + \varepsilon'_2)/2|/\hat{\nu}_N$  has to be compared to at the 5%-level, the program yields the value 0.70545 which clearly exceeds the observed value of the test statistic. Thus, the results of the study warrant rejection of the null hypothesis that there is a nonnegligible association between the polymorphism under study and alcohol dependence.

### Simulation Study

*Samples from rounded normal distributions.* In numerous applications, occurrence of ties between quantitative observations can suitably be accounted for by numerical rounding processes which the associated “latent” values on some underlying continuous measurement scale are subject to. If the underlying continuous variables have Gaussian distributions with unit variance and rounding is done to the nearest multiple of some fixed rational number  $r$ , say, then the distribution of the observable discrete variables is given by a probability mass function of the form

$$p_{\mu;r}(z) = \int_{z-r/2}^{z+r/2} \frac{1}{\sqrt{2\pi}} \exp\{-(u - \mu)^2/2\} du, \quad z = j r, \quad j \in \mathbb{Z}. \quad (6.63)$$

In the sequel, we will call any distribution of this form a rounded normal distribution and write  $Z \sim \mathcal{N}_r(\mu, 1)$  for a discrete random variable  $Z$  having mass function (6.63).

In a first set of simulation experiments on the generalized Mann-Whitney test for equivalence with discretized normally distributed data, rounding was done to nearest multiples of  $r = 1/4$ . Whereas the  $Y_j$  were generated by rounding standard normal random numbers, three different specifications of  $\mu$  were used for generating the  $X_i$ :

$$(a) \quad \mu = -.695; \quad (b) \quad \mu = +.695; \quad (c) \quad \mu = 0.$$

The number  $-.695$  was determined by means of a numerical search algorithm as that value of  $\mu$  which ensures that  $X_i \sim \mathcal{N}_{1/4}(\mu, 1)$ ,  $Y_j \sim \mathcal{N}_{1/4}(0, 1) \Rightarrow \pi_+/(1 - \pi_0) = .30$ . By symmetry, it follows that  $(\mathcal{N}_{1/4}(.695, 1), \mathcal{N}_{1/4}(0, 1))$  is a specific pair of distributions belonging to the right-hand boundary of the hypotheses (6.48) for the choice  $\varepsilon'_1 = \varepsilon'_2 = .20$ . Of course, generating both the  $X_i$  and the  $Y_j$  as samples from  $\mathcal{N}_{1/4}(0, 1)$  provides a basis for studying the power of the test against the specific alternative that  $\pi_+/(1 - \pi_0) = 1/2$ .

As another setting involving samples from rounded normal distributions we investigated the case  $r = 1$ , i.e., rounding to the nearest integer, which of

course produces a much higher rate  $\pi_0$  of ties between  $X$ 's and  $Y$ 's. For  $r = 1$ , the nonnull  $\mu$ 's corresponding to  $\pi_+/(1 - \pi_0) = .30$  and  $\pi_+/(1 - \pi_0) = .70$  are computed to be  $\mp .591$ . The mass functions of the three elements of the family  $\{\mathcal{N}_1(\mu, 1) \mid \mu \in \mathbb{R}\}$  corresponding to  $\mu = -.591$ ,  $\mu = 0$  and  $\mu = +.591$  are depicted in Figure 6.3.

Tables 6.12a,b show the rejection probabilities of the generalized Mann-Whitney test for equivalence at nominal level  $\alpha = .05$  at both boundaries of the hypotheses ( $\rightarrow$  size) and the center of the equivalence range ( $\rightarrow$  power) found in 100,000 replications of the simulation experiments performed with data from rounded normal distributions. The results suggest that the test keeps the nominal significance level even when both sample sizes are fairly small. Furthermore, for given sample sizes the power of the test decreases markedly with the proportion of ties between observations from different populations.

Table 6.12a *Simulated rejection probabilities of the test defined by (6.58) with data from normal distributions rounded to the nearest multiple of .25. (From Wellek and Hampel, 1999, with kind permission by Wiley-VCH.)*

$m$	$n$	Rejection Prob.	Rejection Prob.	Power
		at $\pi_+/(1 - \pi_0) = .30$	at $\pi_+/(1 - \pi_0) = .70$	at $\pi_+/(1 - \pi_0) = .50$
20	20	.04429	.04374	.36408
40	40	.04337	.04279	.79949
60	60	.04245	.04253	.95070
10	90	.04830	.04959	.23361

Table 6.12b *Simulated rejection probabilities of the test defined by (6.58) with data from normal distributions rounded to the nearest integer. (From Wellek and Hampel, 1999, with kind permission by Wiley-VCH.)*

$m$	$n$	Rejection Prob.	Rejection Prob.	Power
		at $\pi_+/(1 - \pi_0) = .30$	at $\pi_+/(1 - \pi_0) = .70$	at $\pi_+/(1 - \pi_0) = .50$
20	20	.04277	.04378	.20246
40	40	.04212	.04216	.54873
60	60	.04247	.04248	.80174
10	90	.04664	.04681	.18361

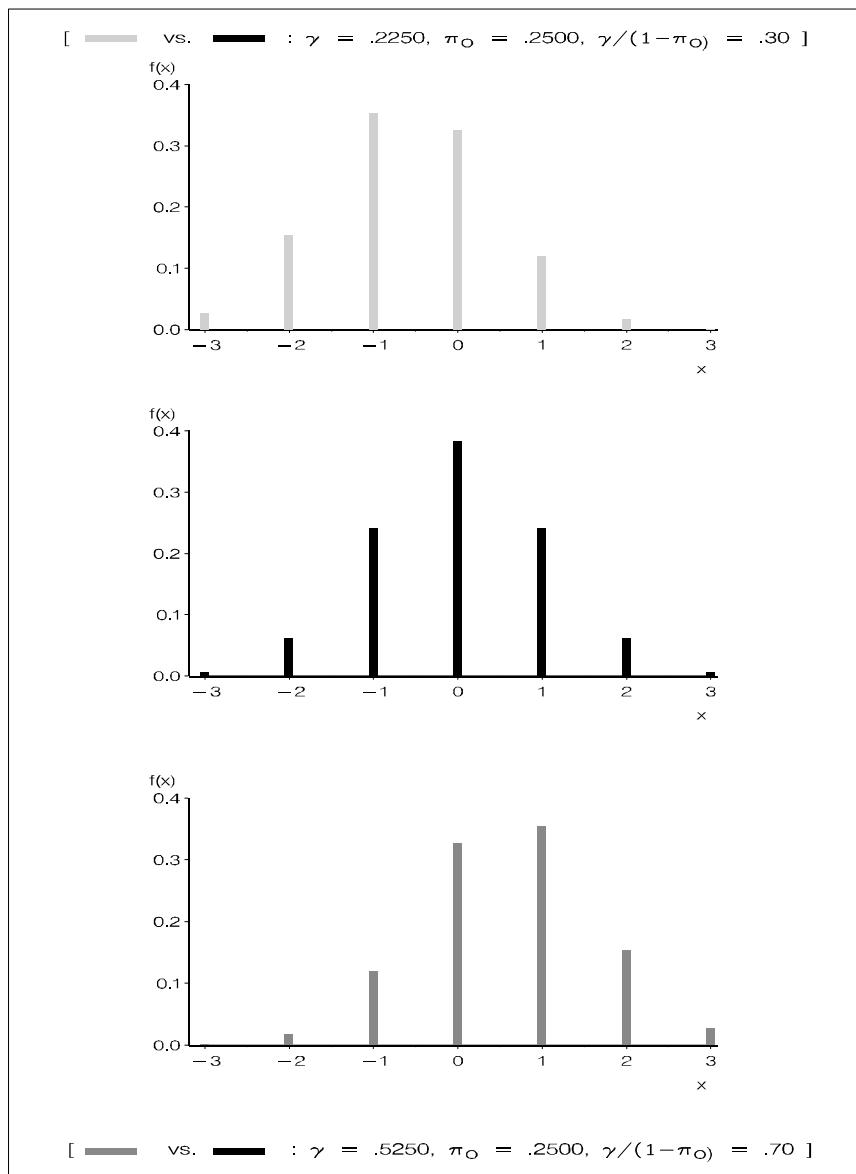


Figure 6.3 Normal distributions rounded to the nearest integer with  $\mu = -.695$  (—),  $\mu = 0$  (—), and  $\mu = .695$  (—).  
(From Wellek and Hampel, 1999, with kind permission by Wiley-VCH.)

*Lehmann alternatives to discrete uniform distributions.* Another parametric submodel of considerable interest for purposes of investigating the small sample properties of the test under consideration is given by the class of Lehmann alternatives to the uniform distribution on the set  $\{1, \dots, k\}$  of the first  $k$  natural numbers ( $k \in \mathbb{N}$ , fixed). By definition, each distribution belonging to this class has a probability mass function of the form

$$p_{\theta;k}(j) = (j/k)^{\theta} - ((j-1)/k)^{\theta}, \quad j = 1, \dots, k, \quad \theta > 0. \quad (6.64)$$

The symbol  $\mathcal{U}_k(\theta)$  is to denote any specific distribution of that form.

In the first part of the simulations referring to the model (6.64), the number of mass points was fixed at  $k = 6$ , and the parameter  $\theta$  varied over the set  $\{\theta_1, \theta_2, \theta_0\}$  with

$$(a) \quad \theta_1 = .467; \quad (b) \quad \theta_2 = 2.043; \quad (c) \quad \theta_0 = 1.$$

These values were determined in such a way that for  $\varepsilon'_1 = \varepsilon'_2 = .20$ , the pair  $(\mathcal{U}_6(\theta_1), \mathcal{U}_6(\theta_0))$  and  $(\mathcal{U}_6(\theta_2), \mathcal{U}_6(\theta_0))$  belongs to the left- and the right-hand boundary of the hypotheses (6.48), respectively. Bar charts of the corresponding probability mass functions are displayed in Figure 6.4.

The second half of the results shown in Table 6.13 for data generated from probability mass functions of the form (6.64) relates to Lehmann alternatives to the uniform distribution on the set  $\{1, 2, 3\}$ . In the case of these much coarser distributions, the values ensuring that  $X_i \sim \mathcal{U}_3(\theta_1), Y_j \sim \mathcal{U}_3(1) \Rightarrow \pi_+/(1 - \pi_0) = .30$  and  $X_i \sim \mathcal{U}_3(\theta_2), Y_j \sim \mathcal{U}_3(1) \Rightarrow \pi_+/(1 - \pi_0) = .70$  were computed to be  $\theta_1 = .494$  and  $\theta_2 = 1.860$ , respectively. Again, the nominal significance level was set to  $\alpha = .05$  throughout, and all rejection probabilities listed in the tables are based on 100,000 replications. Essentially, we are led to the same conclusions as suggested by the results for the rounding error model with underlying Gaussian distributions: For balanced designs, the generalized Mann-Whitney test for equivalence guarantees the nominal significance level even when both sample sizes are as small as 20. However, the rejection probability found in the case  $k = 6$ ,  $(m, n) = (10, 90)$  at the right-hand boundary of the hypotheses shows that the test is not in general strictly conservative.

*Remark.* It is worth noticing that ranking the four tables 6.12a – 6.13b with respect to the power of the test at the center of the equivalence interval yields for every combination of sample sizes the same ordering, namely (6.13b)  $\ll$  (6.12b)  $\ll$  (6.13a)  $\ll$  (6.12a). This is exactly the ordering obtained by arranging the associated values of the tie probability  $\pi_0$  into a decreasing sequence: Straightforward computations show that  $\pi_0$  takes on value .333, .250, .167 and .062 for  $\mathcal{U}_3, \mathcal{N}_1, \mathcal{U}_6$  and  $\mathcal{N}_{1/4}$ , respectively.

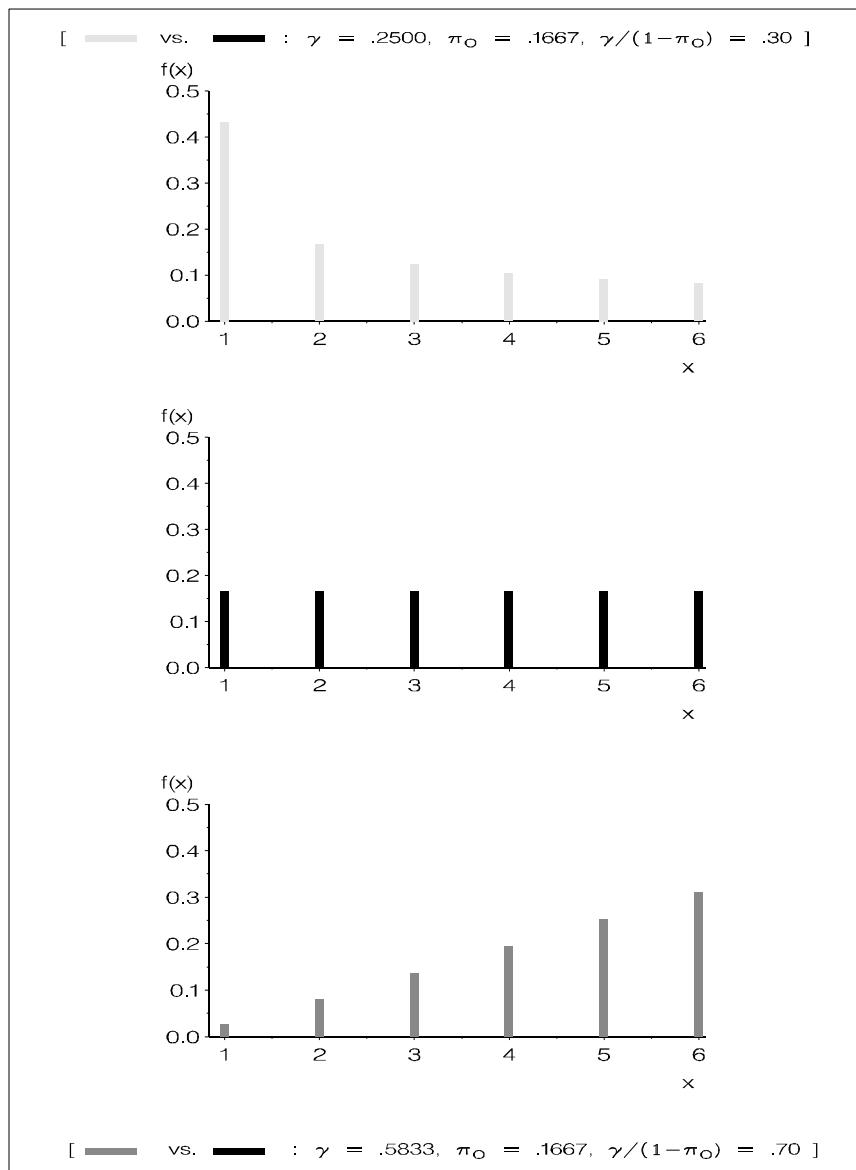


Figure 6.4 Lehmann alternatives to the uniform distribution on  $\{1, \dots, 6\}$  with  $\theta = .467$  () ,  $\theta = 1$  () , and  $\theta = 2.043$  () . (From Wellek and Hampel, 1999, with kind permission by Wiley-VCH.)

Table 6.13a *Simulated rejection probabilities of the test defined by (6.58) under Lehmann alternatives to the uniform distribution on {1, 2, ..., 6}. (From Wellek and Hampel, 1999, with kind permission by Wiley-VCH.)*

$m$	$n$	Rejection Prob.	Rejection Prob.	Power
		at $\pi_+/(1 - \pi_0) = .30$	at $\pi_+/(1 - \pi_0) = .70$	at $\pi_+/(1 - \pi_0) = .50$
20	20	.04500	.04312	.26757
40	40	.04266	.04202	.68698
60	60	.04184	.04201	.89465
10	90	.04838	.05230	.25194
90	10	.04370	.04544	.24936

Table 6.13b *Simulated rejection probabilities of the test defined by (6.58) under Lehmann alternatives to the uniform distribution on {1, 2, 3}. (From Wellek and Hampel, 1999, with kind permission by Wiley-VCH.)*

$m$	$n$	Rejection Prob.	Rejection Prob.	Power
		at $\pi_+/(1 - \pi_0) = .30$	at $\pi_+/(1 - \pi_0) = .70$	at $\pi_+/(1 - \pi_0) = .50$
20	20	.04531	.04566	.17858
40	40	.04270	.04321	.45012
60	60	.04149	.03984	.72160
10	90	.04690	.04772	.15050
90	10	.04203	.04212	.15188

### Generalized Mann-Whitney test for noninferiority

The noninferiority problem associated with (6.48) reads

$$H_1 : \pi_+/(1 - \pi_0) \leq 1/2 - \varepsilon' \text{ versus } K_1 : \pi_+/(1 - \pi_0) > 1/2 - \varepsilon'. \quad (6.65)$$

In view of the weak convergence result underlying (6.58) we can apply again the general result of § 2.3. Specifying  $k = 2$ ,  $T_N = W_+/(1 - W_0)$ ,  $\theta = \pi_+/(1 - \pi_0)$ ,  $\sigma = \nu$  and  $\hat{\tau}_N = \hat{\nu}_N/\sqrt{N}$ , this implies that using the rejection region

$$\left\{ \sqrt{N} \left( W_+/(1 - W_0) - 1/2 + \varepsilon' \right) / \hat{\nu}_N > u_{1-\alpha} \right\} \quad (6.66)$$

yields an asymptotically valid testing procedure for (6.65) which we propose to call generalized Mann-Whitney test for noninferiority. The entries in Tables 6.14a through 6.15b were obtained by simulation under the same models used before for studying the small-sample behavior of the equivalence version of the test. The major changes as compared to the two-sided case relate to power which turns out to increase by up to 40% when the equivalence range specified by the alternative hypothesis is left unbounded on the right.

Table 6.14a *Simulated rejection probabilities of the generalized Mann-Whitney test for noninferiority with data from normal distributions rounded to the nearest multiple of .25.*

$m$	$n$	Rejection Prob.	Power
		at $\pi_+/(1 - \pi_0) = .30$	at $\pi_+/(1 - \pi_0) = .50$
20	20	.04482	.66571
40	40	.04309	.89985
60	60	.04367	.97340
10	90	.04780	.63065

Table 6.14b *Simulated rejection probabilities of the generalized Mann-Whitney test for noninferiority with data from normal distributions rounded to the nearest integer.*

$m$	$n$	Rejection Prob.	Power
		at $\pi_+/(1 - \pi_0) = .30$	at $\pi_+/(1 - \pi_0) = .50$
20	20	.04415	.52300
40	40	.04238	.77213
60	60	.04086	.90060
10	90	.04733	.49468

Table 6.15a *Simulated rejection probabilities of the generalized Mann-Whitney test for noninferiority under Lehmann alternatives to the uniform distribution on {1, 2, ..., 6}.*

$m$	$n$	Rejection Prob.	Power
		at $\pi_+/(1 - \pi_0) = .30$	at $\pi_+/(1 - \pi_0) = .50$
20	20	.04431	.59250
40	40	.04205	.84452
60	60	.04363	.94805
10	90	.04844	.55951
90	10	.04356	.56142

Table 6.15b *Simulated rejection probabilities of the generalized Mann-Whitney test for noninferiority under Lehmann alternatives to the uniform distribution on {1, 2, 3}.*

$m$	$n$	Rejection Prob.	Power
		at $\pi_+/(1 - \pi_0) = .30$	at $\pi_+/(1 - \pi_0) = .50$
20	20	.04391	.46618
40	40	.04189	.71891
60	60	.04162	.85743
10	90	.04695	.44668
90	10	.04138	.44649

## 6.5 Testing for dispersion equivalence of two Gaussian distributions

In routine applications of the most frequently used inferential procedures, the preliminary check on the homogeneity of variances of two Gaussian distributions as a prerequisite for the validity of the classical two-sample  $t$ -test consists of an “inverted” conventional test of the null hypothesis of equality of both variances. In other words, the usual  $F$ -test is carried out and homogeneity of variances taken for granted whenever the result turns out insignificant. In this section we derive an optimal testing procedure tailored for establishing the

alternative hypothesis that the variances, say  $\sigma^2$  and  $\tau^2$ , of two normal distributions from which independent samples  $X_1, \dots, X_m$  and  $Y_1, \dots, Y_n$  have been taken, coincide up to irrelevant discrepancies. The region of still tolerable heterogeneity between  $\sigma^2$  and  $\tau^2$  is defined as a sufficiently short interval around 1 which has to cover the true value of the ratio  $\sigma^2/\tau^2$ . Accordingly, the present section deals with the problem of testing

$$\begin{aligned} H : 0 < \sigma^2/\tau^2 &\leq \omega_1^2 \quad \text{or} \quad \omega_2^2 \leq \sigma^2/\tau^2 < \infty \\ \text{versus} \quad K : \omega_1^2 < \sigma^2/\tau^2 &< \omega_2^2 \end{aligned} \quad (6.67)$$

by means of two independent samples  $(X_1, \dots, X_m), (Y_1, \dots, Y_n)$  such that

$$X_i \sim \mathcal{N}(\xi, \sigma^2), \quad i = 1, \dots, m, \quad Y_j \sim \mathcal{N}(\eta, \tau^2), \quad j = 1, \dots, n. \quad (6.68)$$

Of course, the tolerances  $\omega_1^2, \omega_2^2$  making up the common boundary of the two hypotheses of (6.67) are assumed to be fixed positive numbers satisfying  $\omega_1^2 < 1 < \omega_2^2$ .

Obviously, (6.67) remains invariant under the same group of transformations as the corresponding one-sided testing problem treated in Lehmann and Romano (2005, p. 220, Example 6.3.4). By the results established there we know that a test for (6.67) which is uniformly most powerful among all level- $\alpha$  tests invariant under transformations of the form  $(X_1, \dots, X_m, Y_1, \dots, Y_n) \mapsto (a + bx_1, \dots, a + bx_m, c + dy_1, \dots, c + dy_n)$  with  $-\infty < a, c < \infty, b, d \neq 0, |b| = |d|$ , can be obtained in the following way: In a first step, the raw data  $X_1, \dots, X_m, Y_1, \dots, Y_n$  are reduced to the statistic

$$Q = S_X^2/S_Y^2 \equiv \frac{n-1}{m-1} \sum_{i=1}^m (X_i - \bar{X})^2 / \sum_{j=1}^n (Y_j - \bar{Y})^2 \quad (6.69)$$

which the ordinary  $F$ -test of the null hypothesis  $\sigma^2/\tau^2 = 1$  is based upon. Under the Gaussian model (6.68), the distribution of  $Q$  is an element of the scale family associated with the central  $F$ -distribution with  $\nu_1 = m-1$ ,  $\nu_2 = n-1$  degrees of freedom, and the scale parameter  $\varrho$ , say, is related to the variances we want to compare simply by  $\varrho = \sigma^2/\tau^2$ . Hence, the testing problem (6.67) can be rewritten as

$$\tilde{H} : 0 < \varrho \leq \varrho_1 \vee \varrho_2 \leq \varrho < \infty \quad \text{versus} \quad \tilde{K} : \varrho_1 < \varrho < \varrho_2 \quad (\widetilde{6.67})$$

with  $\varrho_k = \omega_k^2$ ,  $k = 1, 2$ . If an UMP level- $\alpha$  test exists for  $(\widetilde{6.67})$  and is carried out through  $Q$  in terms of the original observations  $X_1, \dots, X_m, Y_1, \dots, Y_n$ , then the result is the desired UMPI test at the same level  $\alpha$  for (6.67).

Let  $R_\circ$  denote a generic random variable following a central  $F$ -distribution with  $(\nu_1, \nu_2)$  degrees of freedom, and  $f_{\nu_1, \nu_2}(\cdot; \varrho)$  the density function of  $\varrho R_\circ$

for arbitrary  $\varrho > 0$ . As is rigorously shown in the Appendix [see p. 370], the family  $(f_{\nu_1, \nu_2}(\cdot; \varrho))_{\varrho > 0}$  (sometimes called family of stretched  $F$ -densities in the literature – cf. Witting, 1985, p. 217) is strictly totally positive of any order and hence in particular STP<sub>3</sub>. Furthermore, continuity of the standard  $F$ -density  $f_{\nu_1, \nu_2}(\cdot; 1)$  with the respective numbers of degrees of freedom obviously implies that the function  $(q, \varrho) \mapsto f_{\nu_1, \nu_2}(q; \varrho)$  is continuous in both of its arguments. Thus, the family  $(f_{\nu_1, \nu_2}(\cdot; \varrho))_{\varrho > 0}$  of densities satisfies all conditions of Theorem A.1.5 [→ p. 371]. Consequently, the desired UMP level- $\alpha$  test for (6.67) exists indeed and is given by the rejection region

$$\left\{ \tilde{C}_{\alpha; \nu_1, \nu_2}^{(1)}(\varrho_1, \varrho_2) < Q < \tilde{C}_{\alpha; \nu_1, \nu_2}^{(2)}(\varrho_1, \varrho_2) \right\}. \quad (6.70)$$

The critical constants  $\tilde{C}_{\alpha; \nu_1, \nu_2}^{(k)}(\varrho_1, \varrho_2), k = 1, 2$ , have to be determined by solving the equations

$$\begin{aligned} F_{\nu_1, \nu_2}(\tilde{C}_2/\varrho_1) - F_{\nu_1, \nu_2}(\tilde{C}_1/\varrho_1) &= \alpha = \\ F_{\nu_1, \nu_2}(\tilde{C}_2/\varrho_2) - F_{\nu_1, \nu_2}(\tilde{C}_1/\varrho_2), \quad 0 < \tilde{C}_1 < \tilde{C}_2 < \infty, \end{aligned} \quad (6.71)$$

with  $F_{\nu_1, \nu_2}(\cdot)$  denoting the cdf of the standard central  $F$ -distribution with  $(\nu_1, \nu_2)$  degrees of freedom.

The algorithm for computing the optimal critical constants admits considerable simplification along the lines described on p. 44 provided the equivalence range for  $\varrho$  has been chosen as symmetric on the log scale, and the underlying parallel-group design is balanced with respect to the sample sizes so that the distribution of  $Q$  has the same number of degrees of freedom for numerator and denominator, respectively. In fact, for  $\nu_1 = \nu_2 = \nu$ , we obviously have  $Q^{-1} \stackrel{d}{=} Q$  under  $\varrho = 1$ , and this implies that the distribution of  $Q$  under any  $\varrho > 0$  coincides with that of  $Q^{-1}$  under  $\varrho^{-1}$ . Hence, for  $\nu_1 = \nu_2 = \nu$ ,  $\varrho_1 = \varrho_o^{-1}$ ,  $\varrho_2 = \varrho_o > 1$ , specifying  $T(\mathbf{X}) = Q$ ,  $\theta = \log \varrho$ ,  $\varepsilon = \log \varrho_o$  yields a setting which satisfies both of the conditions (3.10a), (3.10b) for symmetrizing the critical region of the corresponding uniformly most powerful test. Thus, by (3.11), we may reduce (6.70) and (6.71) in the case of a balanced design and symmetric equivalence limits to

$$\left\{ 1/\tilde{C}_{\alpha; \nu}^{(o)}(\varrho_o) < Q < \tilde{C}_{\alpha; \nu}^{(o)}(\varrho_o) \right\}, \quad (6.72)$$

and

$$F_{\nu, \nu}(\tilde{C}/\varrho_o) - F_{\nu, \nu}(\tilde{C}^{-1}/\varrho_o) = \alpha, \quad \tilde{C} > 1, \quad (6.73)$$

respectively.

Of course, in (6.72),  $\tilde{C}_{\alpha; \nu}^{(o)}(\varrho_o)$  stands for the (unique) solution to equation (6.73).

Reducibility of (6.70) to (6.72) in the special case  $\nu_1 = \nu_2 = \nu$ ,  $\varrho_1 = \varrho_o^{-1}$ ,  $\varrho_2 = \varrho_o$  means that we may write both for  $k = 1$  and  $k = 2$ :  $\tilde{C}_{\alpha; \nu, \nu}^{(k)}(\varrho_o^{-1}, \varrho_o)$

$= 1/\tilde{C}_{\alpha; \nu, \nu}^{(3-k)}(\varrho_{\circ}^{-1}, \varrho_{\circ})$ . In view of the obvious fact that for any random variable  $Z_{\nu_1, \nu_2}$  following a central  $F$ -distribution with  $(\nu_1, \nu_2)$  degrees of freedom, we have  $Z_{\nu_1, \nu_2} \stackrel{d}{=} 1/Z_{\nu_2, \nu_1}$ , this reciprocity relation generalizes to

$$\tilde{C}_{\alpha; \nu_1, \nu_2}^{(k)}(\varrho_1, \varrho_2) = 1/\tilde{C}_{\alpha; \nu_2, \nu_1}^{(3-k)}(\varrho_2^{-1}, \varrho_1^{-1}), \quad k = 1, 2. \quad (6.74)$$

Table 6.16 shows the value of the critical constant  $\tilde{C}_{.05; \nu}^{(\circ)}(\varrho_{\circ})$  to be used in the  $F$ -test for equivalence at level  $\alpha = .05$  in symmetric settings with  $\nu_1 = \nu_2 = \nu$  and  $\varrho_2 = 1/\varrho_1 = \varrho_{\circ}$ , for  $\nu = 10(5)75$  and  $\varrho_{\circ} = 1.50(.25)2.50$ . The corresponding values of the power attained against the alternative that the population variances under comparison are equal ( $\leftrightarrow \varrho = 1$ ), are given in Table 6.17.

Table 6.16 *Critical constant  $\tilde{C}_{.05; \nu}^{(\circ)}(\varrho_{\circ})$  of the  $F$ -test for dispersion equivalence of two Gaussian distributions at level  $\alpha = 5\%$  with common sample size  $m = n = \nu + 1$  and symmetric equivalence range  $(\varrho_{\circ}^{-1}, \varrho_{\circ})$  for  $\varrho = \sigma^2/\tau^2$ , for  $\nu = 10(5)75$  and  $\varrho_{\circ} = 1.50(.25)2.50$ .*

$\nu$	$\varrho_{\circ} =$				
	1.50	1.75	2.00	2.25	2.50
10	1.05113	1.06165	1.07593	1.09448	1.11796
15	1.04572	1.06048	1.08259	1.11432	1.15830
20	1.04362	1.06328	1.09561	1.14600	1.21867
25	1.04307	1.06849	1.11396	1.18805	1.29055
30	1.04346	1.07566	1.13724	1.23644	1.36110
35	1.04450	1.08461	1.16457	1.28538	1.42349
40	1.04605	1.09529	1.19424	1.33075	1.47701
45	1.04803	1.10756	1.22424	1.37118	1.52302
50	1.05040	1.12122	1.25297	1.40685	1.56300
55	1.05314	1.13590	1.27963	1.43839	1.59816
60	1.05623	1.15116	1.30399	1.46647	1.62939
65	1.05967	1.16654	1.32612	1.49167	1.65740
70	1.06344	1.18165	1.34626	1.51445	1.68272
75	1.06754	1.19620	1.36464	1.53518	1.70576

Table 6.17 Power attained at  $\varrho = 1$  when using the critical constants shown in Table 6.16 for the  $F$ -statistic.

$\nu$	$\varrho_0 =$				
	1.50	1.75	2.00	2.25	2.50
10	.06129	.07350	.08986	.11071	.13650
15	.06786	.08905	.12011	.16328	.22032
20	.07511	.10778	.15978	.23646	.33743
25	.08313	.13024	.21057	.33006	.47162
30	.09198	.15699	.27303	.43531	.59665
35	.10174	.18852	.34530	.53850	.69906
40	.11250	.22510	.42279	.62999	.77824
45	.12433	.26666	.49986	.70663	.83793
50	.13732	.31260	.57194	.76907	.88235
55	.15156	.36179	.63666	.81917	.91509
60	.16710	.41267	.69332	.85903	.93904
65	.18402	.46361	.74219	.89053	.95644
70	.20234	.51317	.78396	.91529	.96902
75	.22207	.56028	.81944	.93466	.97805

In contrast to (6.73) which can be solved more or less by trial and error since the expression on the left-hand side is obviously increasing in  $\tilde{C}$ , solving the system (6.71) to be treated in the general case requires appropriate software tools. The program **fstretch** to be found in the **WKTSEQ2 Source Code Package**, provides for arbitrary choices of  $\alpha, \nu_k$  and  $\varrho_k$  ( $k = 1, 2$ ) both the critical constants  $\tilde{C}_{\alpha; \nu_1, \nu_2}^{(k)}(\varrho_1, \varrho_2)$  determining the UMPI critical region (6.70), and the power against  $\varrho = 1$  ( $\Leftrightarrow \sigma^2 = \tau^2$ ). Basically, the algorithm used by the program is an adaptation of the iteration scheme described in general terms on pp. 42–3, to the specific case that the STP<sub>3</sub> family which the distribution of the test statistic belongs to is generated by rescaling a standard central  $F$ -distribution with the appropriate numbers of degrees of freedom, in all possible ways.

### Example 6.3

In a laboratory specialized in testing basic physical and chemical properties of innovative materials for potential use in dental medicine, the existing device for taking measurements of breaking strengths of solid plastic elements had to be replaced with technically modernized equipment of the same kind. Two systems  $A$  and  $B$ , say, were taken into final consideration with  $A$  being priced markedly lower than  $B$ . In view of this, the head of the laboratory decided to take his own data in order to assess the equivalence of both devices with respect to precision of the measurements they provide. For that purpose, a

total of 100 sample elements made of the same material were prepared. For one half of the elements, breaking strength was to be measured by means of  $A$ , for the other by means of  $B$ , respectively. Since all sample elements could be considered identical in physical properties, it was reasonable to regard pure measurement error as the only source of variability found in both sequences of values. Out of the 50 measurements to be taken by means of  $A$ , only  $m = 41$  came to a regular end since the remaining 9 elements broke already during insertion into the device's clamp. By the same reason, the number of values eventually obtained by means of device  $B$  was  $n = 46$  rather than 50. All measurement values were assumed to be realizations of normally distributed random variables with parameters  $(\xi, \sigma^2)$  [ $\leftrightarrow$  device  $A$ ] and  $(\eta, \tau^2)$  [ $\leftrightarrow$  device  $B$ ], respectively. Precision equivalence of both devices was defined through the condition  $.75 < \sigma/\tau < 1.33$  on the standard deviations of the underlying distributions. From the raw data [given in kiloponds], the following values of the sample means and variances were computed:

$$\begin{aligned} \text{Device } A: \quad \bar{X} &= 5.0298, \quad S_X^2 = .0766; \\ \text{Device } B: \quad \bar{Y} &= 4.8901, \quad S_Y^2 = .0851. \end{aligned}$$

Running the program **fstretch**, one reads from the output list that in the UMPI test at level  $\alpha = .05$  for  $\sigma^2/\tau^2 \leq .75^2$  or  $\sigma^2/\tau^2 \geq 1.33^2$  versus  $.75^2 < \sigma^2/\tau^2 < 1.33^2$ , the statistic  $Q = S_X^2/S_Y^2$  has to be checked for inclusion between the critical bounds  $\tilde{C}_{.05;40,45}^{(1)}(.5625, 1.7689) = .8971$ ,  $\tilde{C}_{.05;40,45}^{(2)}(.5625, 1.7689) = 1.1024$ . Since the observed value was  $Q = .0766/.0851 = .9001$ , the data satisfy the condition for rejecting nonequivalence. However, the power to detect even strict equality of variances was only 26.10% under the specifications made in this example.

#### *Testing for one-sided dispersion equivalence of two Gaussian distributions*

In some applications, it will be sufficient to establish the hypothesis that the variability of the observations  $X_i$  obtained under the experimental treatment, exceeds that of the reference population from which the  $Y_j$  are taken, by a practically irrelevant amount at most. Put in formal terms, the testing problem we then are interested in is given by the following pair of hypotheses:

$$H_1 : \sigma^2/\tau^2 \geq 1 + \varepsilon \text{ versus } K_1 : \sigma^2/\tau^2 < 1 + \varepsilon. \quad (6.75)$$

Similar arguments as those which were used above to show that (6.70) is the critical region of a UMPI level- $\alpha$  test for (6.67) allow us to conclude that the UMPI level- $\alpha$  test for (6.75) rejects the null hypothesis of a relevantly increased variability of the  $X_i$  if it turns out that

$$S_X^2/S_Y^2 < (1 + \varepsilon)F_{m-1,n-1;\alpha} \quad (6.76)$$

For a proper use of (6.76), it is essential to note that  $F_{m-1,n-1;\alpha}$  stands for the lower  $100\alpha$  percentage point of a central  $F$ -distribution with  $m - 1, n - 1$  degrees of freedom.

The critical upper bounds to the observed variance ratio which have to be used according to (6.76) are shown in Table 6.18 for the same choice of the  $df$ 's (assumed to be common to numerator and denominator) as covered by Table 6.16, and the equivalence margin  $\varepsilon$  ranging through  $\{.10, .25, .50, 1.0, 3.0\}$ . Remarkably, for small sample sizes, a decision in favor of the hypothesis of nonexistence of a relevant increase in variability in the population underlying the  $X_i$  as compared to the distribution of the  $Y_j$  requires that the observed ratio of variances falls below unity even when the upper limit to the theoretical variance ratio allowed by that hypothesis is as large as 2.00. For  $\varepsilon = .50$  and  $\varepsilon = 1.00$ , the entries in Table 6.18 are the noninferiority counterparts of the critical constants appearing in Table 6.16 for the choice  $\varrho_0 = 1.50$  and  $\varrho_0 = 2.00$ , respectively. In accordance with the construction of both tests, the latter are always larger than the former, but the difference vanishes asymptotically as  $\nu$  increases to infinity.

The power attained in the  $F$ -test determined by the critical upper bounds listed in Table 6.18 against the null alternative  $\sigma^2 = \tau^2$  can be read from Table 6.19. Given the sample sizes and the upper limit of the equivalence range specified under the hypothesis, the power of the test for one-sided equivalence

Table 6.18 *Critical upper bound of the  $F$ -test for one-sided dispersion equivalence of two Gaussian distributions at level  $\alpha = 5\%$  with common sample size  $m = n = \nu + 1$  and equivalence range  $(0, 1 + \varepsilon)$  for  $\varrho = \sigma^2/\tau^2$ , for  $\nu = 10(5)75$  and  $\varepsilon \in \{.10, .25, .50, 1.0, 3.0\}$ .*

$\nu$	$\varepsilon =$				
	0.10	0.25	0.50	1.00	3.00
10	0.36935	0.41971	0.50365	0.67154	1.34308
15	0.45768	0.52009	0.62410	0.83214	1.66428
20	0.51785	0.58847	0.70616	0.94155	1.88310
25	0.56253	0.63924	0.76709	1.02278	2.04557
30	0.59754	0.67903	0.81483	1.08644	2.17288
35	0.62602	0.71138	0.85366	1.13821	2.27643
40	0.64981	0.73842	0.88611	1.18148	2.36295
45	0.67011	0.76149	0.91379	1.21839	2.43677
50	0.68772	0.78150	0.93780	1.25039	2.50079
55	0.70319	0.79908	0.95889	1.27852	2.55705
60	0.71693	0.81470	0.97764	1.30351	2.60703
65	0.72926	0.82870	0.99444	1.32592	2.65185
70	0.74040	0.84136	1.00963	1.34618	2.69235
75	0.75053	0.85288	1.02345	1.36461	2.72921

uniformly exceeds that of the test for dispersion equivalence in the two-sided sense. This increase in power, which is most marked for larger sample sizes and a tight equivalence margin, reflects the fact that under  $\sigma^2/\tau^2 = 1 + \varepsilon$ , the probability mass of the interval between the upper critical bounds of both tests is small as compared to that of the left-hand part of the acceptance region for the equivalence case which belongs to the rejection region of the test for noninferiority.

*Exploiting the results of Section 6.5 for testing for equivalence of hazard rates in the two-sample setting with exponentially distributed data*

There is a close formal relationship between the testing problems treated in the core part of this section, and that of testing for equivalence of the hazard rates of two exponential distributions from which independent samples are obtained. In order to make this relationship precise, let us assume that instead of following normal distributions, the observations making up the two samples under analysis satisfy

$$X_i \sim \mathcal{E}(\sigma), \quad i = 1, \dots, m, \quad Y_j \sim \mathcal{E}(\tau), \quad j = 1, \dots, n, \quad (6.77)$$

where, as before [recall p. 55],  $\mathcal{E}(\theta)$  symbolizes a standard exponential distribution with hazard rate  $\theta^{-1} > 0$ . Let us further assume that the hypotheses

Table 6.19 *Power of the UMPI test (6.76) at level  $\alpha = .05$  against  $\sigma^2 = \tau^2$  attained with the sample sizes and equivalence margins appearing in Table 6.18.*

$\nu$	$\varepsilon =$				
	0.10	0.25	0.50	1.00	3.00
10	.06594	.09355	.14735	.27023	.67512
15	.07071	.10852	.18571	.36327	.83274
20	.07485	.12216	.22172	.44710	.91719
25	.07862	.13501	.25612	.52223	.96022
30	.08212	.14730	.28922	.58906	.98134
35	.08543	.15919	.32116	.64806	.99142
40	.08859	.17077	.35202	.69978	.99612
45	.09162	.18208	.38184	.74483	.99827
50	.09455	.19318	.41064	.78386	.99924
55	.09740	.20409	.43843	.81749	.99967
60	.10018	.21484	.46524	.84633	.99986
65	.10289	.22543	.49107	.87097	.99994
70	.10555	.23589	.51594	.89193	.99997
75	.10816	.24622	.53985	.90970	.99999

about  $(\sigma, \tau)$  between which we want to decide by means of these samples, read

$$H : 0 < \sigma/\tau < \varrho_1 \text{ or } \varrho_2 \leq \sigma/\tau < \infty \text{ versus } K : \varrho_1 < \sigma/\tau < \varrho_2, \quad (6.78)$$

and

$$H_1 : \sigma/\tau \geq 1 + \varepsilon \text{ versus } K_1 : \sigma/\tau < 1 + \varepsilon. \quad (6.79)$$

Another application of the general results proved in § 6.5 of the book of Lehmann and Romano (2005) shows that carrying out the UMP level  $\alpha$ -test for (6.67) in terms of  $m^{-1} \sum_{i=1}^m X_i / n^{-1} \sum_{j=1}^n Y_j$  yields a UMPI test for (6.78) at the same level, provided the numbers of degrees of freedom are specified as  $\nu_1 = 2m$ ,  $\nu_2 = 2n$  in (6.70) and (6.71). The group of transformations with respect to which the invariance property holds, consists this time of all homogeneous dilations of any point in  $(m+n)$ -dimensional real space  $\mathbb{R}^{m+n}$ . In other words, the practical implementation of the UMPI level- $\alpha$  test for equivalence of two exponential distributions with respect to the hazard rates differs from that of the modified  $F$ -test for dispersion equivalence of two Gaussian distributions only with respect to the rule for computing the test statistic from the raw observations and for counting the numbers of degrees of freedom. Thus, it is in particular clear that the program `fstretch` also enables us to carry out an optimal two-sample equivalence test for exponentially distributed data. Likewise, checking the observed value of  $m^{-1} \sum_{i=1}^m X_i / n^{-1} \sum_{j=1}^n Y_j$  against the critical upper bound appearing in (6.76) as computed with  $\nu_1 = 2m$ ,  $\nu_2 = 2n$  yields an UMPI level- $\alpha$  test for (6.79) under the model (6.77). Since the exponential scale-parameter family of distributions is the most widely used special case of a model which satisfies the proportional hazards assumption (6.23), the UMPI tests being available for two samples from exponential distributions can be viewed as natural parametric competitors to the exact semiparametric tests based on the linear rank statistic with Savage scores as have been derived in § 6.3.

## 6.6 Equivalence tests for two binomial samples

### 6.6.1 Exact Fisher type test for noninferiority with respect to the odds ratio

One of the most obvious and important examples of a setting which is definitely not accessible to the translation-based approach to one-sided equivalence testing discussed in § 2.1, is that of a two-arm trial with a binary endpoint, i.e., an outcome variable taking on only two possible values labeled

“response” and “nonresponse” in the sequel. Without loss of statistical information, the data obtained from a trial of this type can be reduced to the frequencies of responders and nonresponders observed in both treatment arms. The latter are conveniently arranged in a table of the well-known  $2 \times 2$  layout. In order to fix notation, a generic contingency table of that type is displayed below.

Table 6.20 *Contingency table for the analysis of a two-arm trial with binary outcome-variable.*

Treatment	Response		
	+	-	$\Sigma$
$A$	$X$ ( $p_1$ )	$m - X$ ( $1 - p_1$ )	$m$ (1.00)
$B$	$Y$ ( $p_2$ )	$n - Y$ ( $1 - p_2$ )	$n$ (1.00)
$\Sigma$	$S$	$N - S$	$N$

By assumption,  $X$  has a binomial distribution with parameters  $m, p_1$  [short-hand notation:  $X \sim \mathcal{B}(m, p_1)$ ] and is independent of  $Y \sim \mathcal{B}(n, p_2)$ . As to parametrization of the corresponding family of joint distributions of  $(X, Y)$ , we start with adopting the view supported by the arguments given in § 1.6 that the odds ratio  $\rho = p_1(1-p_2)/(1-p_1)p_2$  provides a more adequate measure of dissimilarity of the two distributions under comparison than the difference  $\delta = p_1 - p_2$  between the responder rates in the underlying populations. Accordingly, we are seeking for a suitable test of

$$H_1 : \rho \leq 1 - \varepsilon \quad \text{versus} \quad K_1 : \rho > 1 - \varepsilon \quad (6.80)$$

with some fixed  $0 < \varepsilon < 1$ .

An optimal solution to this problem can be derived by way of generalizing the construction behind the exact test for homogeneity of  $\mathcal{B}(m, p_1)$  and  $\mathcal{B}(n, p_2)$  usually named after R.A. Fisher (1934, § 21.02). Like the latter, the optimal [precisely: uniformly most powerful unbiased (UMPU) — see Lehmann and Romano (2005, pp. 126–127)] test for (6.80) is based on the conditional distribution of  $X$  given the realized value  $s \in \{0, 1, \dots, N\}$  of the “column total”  $S = X + Y$ . The conditional p-value  $p(x|s)$  of the number  $x$  of responders counted in treatment arm  $A$  has to be computed by means of a probability distribution which, following Harkness (1965), is usually called extended hypergeometric distribution (cf. Johnson et al., 1992, § 6.11). In

this conditional distribution, any possible value of  $X$  is an integer  $x$  such that  $\max\{0, s - n\} \leq x \leq \min\{s, m\}$ . For any such  $x$  the precise formula for the conditional p-value reads

$$p(x|s) = \sum_{j=x}^m \binom{m}{j} \binom{n}{s-j} (1-\varepsilon)^j / \sum_{j=\max\{0, s-n\}}^{\min\{s, m\}} \binom{m}{j} \binom{n}{s-j} (1-\varepsilon)^j. \quad (6.81)$$

Except for trivial cases like  $s = 1$  or  $x = m$ , this formula is tractable for manual computations only if both sample sizes are very small. On the other hand, writing a program for the extended hypergeometric distribution is hardly more than a routine exercise, and the SAS system provides an intrinsic function for the exact computation of the corresponding cdf. As is the case for most other predefined SAS functions for special distributions, `probhypr` has a counterpart in R. It can be found in the package `BiasedUrn` and called using the function name `pFNCHypergeo`.

Fortunately, the computational tools which can be made use of when applying the UMPU test for one-sided equivalence of two binomial distributions with respect to the odds ratio, go considerably beyond the availability of an algorithm for computing significance probabilities. In fact, both the power of the test against arbitrary specific alternatives, and sample sizes required for maintaining some prespecified power, are accessible to exact computational methods as well. Naturally, both of these tasks are much more demanding conceptually no less than technically than mere computation of exact conditional p-values. An algorithm for exact power computations will be outlined first.

Let us keep both sample sizes  $m, n$  fixed and denote by  $(p_1^*, p_2^*) \in (0, 1)^2$  any point in the region corresponding to the alternative hypothesis  $K_1$  of (6.80). By definition, the power of the test based on (6.81) against  $(p_1^*, p_2^*)$  equals the *nonconditional* rejection probability, given the true distribution of  $X$  and  $Y$  is  $\mathcal{B}(m, p_1^*)$  and  $\mathcal{B}(n, p_2^*)$ , respectively. In view of this, for each possible value  $s = 0, 1, \dots, N$  of the column total  $S = X + Y$ , the critical value  $k_\alpha(s)$ , say, of the test has to be determined. Adopting the convention that, conditional on  $\{S = s\}$ , the rejection region be of the form  $\{X > k_\alpha(s)\}$ ,  $k_\alpha(s)$  is obviously given as the smallest integer  $x$  such that  $\max\{0, s - n\} \leq x \leq \min\{s, m\}$  and  $p(x+1|s) \leq \alpha$ . Furthermore, it is essential to observe that the test holds its (rather strong) optimality property only if in each conditional distribution of  $X$  the prespecified significance level  $\alpha \in (0, 1)$  is exactly attained rather than only maintained in the sense of not exceeding it. In view of the discreteness of all distributions involved, this typically requires that on the boundary of the critical region  $\{X > k_\alpha(s)\}$ , i.e., for  $X = k_\alpha(s)$  a randomized decision

in favor of the alternative hypothesis is taken with probability  $\gamma(s) = [\alpha - p(k_\alpha(s)+1|s)]/[p(k_\alpha(s)|s) - p(k_\alpha(s)+1|s)]$ . Although randomized decisions between statistical hypotheses are hardly acceptable for real applications, it seems reasonable to carry out the power calculations both for the randomized and the ordinary nonrandomized version of the test defined by accepting on the boundary of the rejection regions *always*  $H_1$ . In fact, given some specific alternative  $(p_1^*, p_2^*)$  of interest, the rejection probability of the randomized version of the test is an upper bound to the power attainable in any other unbiased test of (6.80). Moreover, knowledge of the power of the exact UMPU test allows us to estimate the maximum improvement which might be achieved by deriving more sophisticated nonrandomized versions of the test [see below, § 6.6.2].

The conditional power of the randomized test is given by

$$\beta(p_1^*, p_2^*|s) =$$

$$\frac{\sum_{j=k_\alpha(s)+1}^{\min\{s,m\}} \binom{m}{j} \binom{n}{s-j} \rho_*^j + \gamma(s) \binom{m}{k_\alpha(s)} \binom{n}{s-k_\alpha(s)} \rho_*^{k_\alpha(s)}}{\sum_{j=\max\{0,s-n\}}^{\min\{s,m\}} \binom{m}{j} \binom{n}{s-j} \rho_*^j}, \quad (6.82)$$

$$\text{with } \rho_* = p_1^*(1-p_2^*)/(1-p_1^*)p_2^*.$$

In order to compute its nonconditional power against the arbitrarily fixed specific alternative  $(p_1^*, p_2^*)$ , we have to integrate the functions  $s \mapsto \beta(p_1^*, p_2^*|s)$  with respect to the distribution of  $S$ . Unfortunately, the latter is simply of binomial form again only if  $(p_1^*, p_2^*)$  lies on the diagonal of the unit square. Whenever we select a parameter combination with  $p_1^* \neq p_2^*$ , the distribution of  $S$  must be computed from scratch by means of convolution.

The Fortran program `bi2ste1` supplied in the **WKTSHEQ2 Source Code Package** implements this algorithm for sample sizes  $m, n$  with  $m + n \leq 2000$  and arbitrary choices of  $\alpha, \varepsilon, p_1^*$  and  $p_2^*$ . In order to guarantee high numerical accuracy of the results, it uses (like the other Fortran programs to be found at the same URL) fourfold precision in floating-point arithmetic throughout. A double-precision version is made available as a DLL which can be called from R and produces results whose numerical accuracy is still fully satisfactory for practical applications. Unavoidably, the computational effort entailed in determining sample sizes required to guarantee some prespecified level  $\beta^*$  of power is much higher. The program `bi2ste2` which has been constructed for serving that purpose, is based on an iteration algorithm whose efficiency strongly depends on a sensible choice of the initial values we let the iteration

process start from. In determining the latter, the exact nonconditional power  $\beta(p_1^*, p_2^*)$  is approximated by the conditional power  $\beta(p_1^*, p_2^*)\tilde{s}(m, n)$  where  $\tilde{s}(m, n)$  denotes the integer closest to the expected value  $mp_1^* + np_2^*$  of  $S$ .

*Example 6.4*

To provide an illustration of the Fisher type exact test for one-sided equivalence of two binomial distributions with respect to the odds ratio, we will use some part of the data from a comparative phase-IV trial of two antibiotics administered to treat streptococcal pharyngitis (Scaglione, 1990). At the time the trial was launched, clarithromycin was an experimental drug of which one knew from animal experiments that its toxicity was considerably weaker than that of the classical macrolid antibiotic erythromycin used as the reference medication. Hence, with regard to therapeutic efficacy, one-sided equivalence with erythromycin could have been considered a sufficient argument for preferring the new macrolid in the treatment of streptococcal infections.

Specifying  $\varepsilon = .50$  in (6.80), we find with the data of Table 6.21 by means of the SAS function mentioned in connection with the general formula for computing the p-value in the test under consideration [ $\rightarrow$  (6.81)] that  $p(98|195) = 1 - \text{probhyp}(213, 106, 195, 97, .50) = .049933$ . Thus, at the 5% level of significance, the data of the trial allow the conclusion that the odds of a favorable response to clarithromycin do not fall by more than 50% as compared to those of a positive outcome after administering erythromycin. For the power of the UMPU test at level  $\alpha = .05$  against the observed alternative  $(p_1^*, p_2^*) = (.9245, .9065)$  [ $\leftarrow$  relative frequencies of favorable responses shown in Table 6.21], the Fortran program `bi2ste1` gives the result  $\beta = .579960 \approx 58\%$ . Omitting randomized decisions and keeping the nominal value of  $\alpha$  equal to

Table 6.21 *Successes and failures observed when treating  $\beta$  hemolytic streptococcal pharyngitis in patients less than 65 years of age by means of clarithromycin (A) and erythromycin (B), respectively.*

Treat- ment	Response			$\Sigma$
	favorable	nonfavor.		
A	98 (92.45%)	8 (7.55%)		106 (100.0%)
B	97 (90.65%)	10 (9.35%)		107 (100.0%)
$\Sigma$	195	18		213

the target significance level, the power drops to  $.505559 \approx 51\%$ . Accordingly, about every second two-arm trial using a binary endpoint and sample sizes  $m = 106, n = 107$  will lead to a rejection of the null hypothesis  $\varrho \leq .50$  if the true values of the binomial parameters are 90.65% ( $\leftrightarrow$  Treatment *B*) and 92.45% ( $\leftrightarrow$  Treatment *A*), and the test is applied in its conservative nonrandomized version. The sample sizes required in a balanced trial to raise the power of the randomized test against the same specific alternative to 80%, are computed to be  $m = n = 194$  by means of `bi2ste2`.

### 6.6.2 Improved nonrandomized tests for noninferiority with respect to the odds ratio

We start with exploring the possibility of reducing the conservatism entailed in incorporating the critical bounds  $k_\alpha(0), \dots, k_\alpha(N)$  of the exact Fisher type test into the acceptance region, through replacing the *target significance level  $\alpha$*  with a maximally increased nominal level  $\alpha^*$ . Despite the conceptual simplicity of this approach which for testing traditional one- or two-sided hypotheses about  $\varrho$  goes back to Boschloo (1970), the computational procedure is fairly demanding and requires a suitable software tool. The computational burden is markedly reduced by the fact that the maximum of the rejection probability of each test carried out through comparing the conditional p-value (6.81) with some fixed upper bound over the null hypothesis of (6.80) is taken on at the boundary. This follows from a result stated by Röhmel and Mannsmann (1999a) referring to Hájek and Havránek (1978) as a primary source for a rigorous derivation.

All values of  $\alpha^*$  displayed in Table 6.22 have been computed by means of another Fortran program named `bi2ste3` (which is again made available in the `WKTSEQ2 Source Code Package` also as a DLL accessible from R). The program can be used for determining maximally raised nominal levels for arbitrary additional combinations of  $\alpha, \varepsilon, m, n$  as long as the sum  $N$  of both sample sizes does not exceed 2000. Actually, each  $\alpha^*$  given in the above table represents a nondegenerate interval of nominal levels over which the size of the test remains constant. Of course, it makes no difference if one specific point in that interval is replaced with some other value producing a test of identical size. All what really matters is maximization of the size subject to the condition that the latter must not exceed the prespecified significance level (except for differences small enough for being neglected in practice).

Detailed insights about the effectiveness of Boschloo's technique are provided by comparing the power of all three tests (exact UMPU, conservative nonrandomized version, and nonrandomized conditional test at maximally increased nominal level) against all possible null alternatives specifying  $p_1 = p_2$  for arbitrary  $p_2 \in (0, 1)$ . Figure 6.5 shows these power curves for samples of size  $n = 100$  each, noninferiority margin  $\varepsilon = .50$ , and  $.05$  as the target significance level. For reference responder rates  $p_2$  between 25 and 75%, the

Table 6.22 *Nominal significance levels and sizes of improved conditional tests for (6.80) maintaining the 5%-level, for  $\varepsilon = 1/3, 1/2$  and  $m = n = 10, 25, 50, 100, 200$ . [Number in (): size of the nonrandomized test at nominal level  $\alpha = .05$ .]*

$\varepsilon$	$n$	$\alpha^*$	Size	
$1/3$	10	.08445	.04401	(.02118)
	25	.09250	.05083	(.02728)
	50	.07716	.04991	(.02790)
	75	.06295	.04997	(.03797)
	100	.06064	.05006	(.03915)
	200	.05974	.05015	(.04108)
$1/2$	10	.10343	.04429	(.02869)
	25	.07445	.05064	(.03049)
	50	.06543	.04999	(.03551)
	75	.06810	.05053	(.03584)
	100	.06468	.05040	(.03761)
	200	.05992	.04970	(.04344)

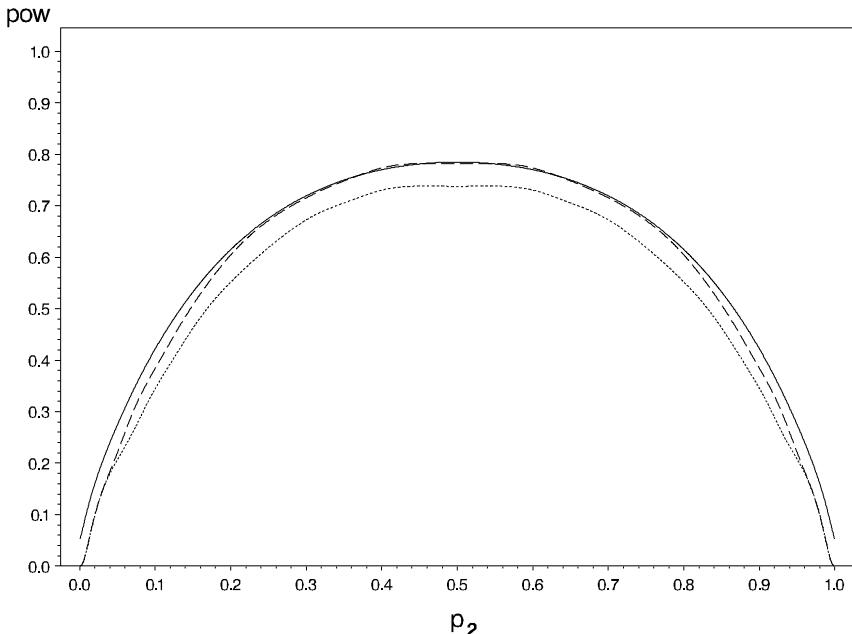


Figure 6.5 *Power function of the exact UMPU test (—) for noninferiority with respect to the odds ratio, its nonrandomized counterpart (· · ·), and the improved nonrandomized test at increased nominal level (---) [ $\varepsilon = .50, \alpha = .05, m = n = 100$ ].*

power curve of the nonrandomized conditional test at the nominal level  $\alpha^*$  to be used according to the results listed in Table 6.22 is practically indistinguishable from that of the exact randomized UMPU test, and the width of the interval for which this holds true, increases with increasing common sample size  $n$ . With the traditional conservative nonrandomized test, the improved nonrandomized test shares the behavior of the power function at the boundaries of the range of  $p_2$ : As  $p_2 \rightarrow 0$  or 1, the power converges to zero, reflecting lack of unbiasedness of both tests which cannot be avoided unless randomized decisions between both hypotheses are admitted.

The use of the *objective Bayesian approach* for constructing an improved nonrandomized test for one-sided equivalence of two binomial distributions with respect to various parametric functions including the odds ratio, was investigated by Wellek (2005). Generalizing the argument made explicit by Box and Tiao (1973, § 1.3.4) for the one-sample case to the binomial two-sample setting leads to defining a noninformative reference prior as the product of two beta distributions with parameters  $(1/2, 1/2)$  each. Denoting the joint posterior distribution of the basic parameters  $(p_1, p_2)$  associated with this prior by  $P_{(x,y)}^*(\cdot)$ , it can easily be shown that we have

$$P_{(x,y)}^*[\mathbf{p}_1 \leq p_1, \mathbf{p}_2 \leq p_2] = I_{p_1}(x + 1/2, m - x + 1/2) \cdot I_{p_2}(y + 1/2, n - y + 1/2) \quad (6.83)$$

where  $I.(a, b)$  stands for the beta distribution function with parameters  $(a, b)$  [for arbitrary  $a, b > 0$ ] and  $(x, y)$  the realized value of  $(X, Y)$ . The density corresponding to (6.83) is

$$f_{(x,y)}^*(p_1, p_2) = \frac{\Gamma(m+1)}{\Gamma(x+1/2)\Gamma(m-x+1/2)} p_1^{x-1/2} (1-p_1)^{m-x-1/2} \cdot \frac{\Gamma(n+1)}{\Gamma(y+1/2)\Gamma(n-y+1/2)} p_2^{y-1/2} (1-p_2)^{n-y-1/2}, \quad 0 < p_1, p_2 < 1. \quad (6.84)$$

Integrating  $f_{(x,y)}^*(\cdot, \cdot)$  over the set  $\mathcal{K}_1 = \{(p_1, p_2) \mid p_1(1-p_2)/(p_2(1-p_1)) > 1 - \varepsilon\}$  and comparing the result with the usual cutoff  $1 - \alpha$  [recall § 2.4] leads to the following critical inequality which the observed counts  $(x, y)$  of responders have to satisfy in order to allow rejection of the null hypothesis in the objective Bayesian test of (6.80):

$$\int_0^1 \left\{ \left[ 1 - I_{(1-\varepsilon)p_2/((1-p_2)+(1-\varepsilon)p_2)}(x + 1/2, m - x + 1/2) \right] \cdot \frac{\Gamma(n+1)}{\Gamma(y+1/2)\Gamma(n-y+1/2)} \cdot p_2^{y-1/2} (1-p_2)^{n-y-1/2} \right\} dp_2 \geq 1 - \alpha. \quad (6.85)$$

For evaluating the left-hand side of this inequality, we prefer to rely on Gauss type integration (see, e.g., Davis and Rabinowitz, 1975, § 2.7) with abscissas chosen as the roots of a Legendre polynomial of sufficiently high degree (c.f.

Abramowitz and Stegun, 1965, Table 25.4). Even when the number of abscissas is chosen as large as 96, the respective algorithm is fast enough for allowing one to search through the whole sample space  $\{0, 1, \dots, m\} \times \{0, 1, \dots, n\}$  of  $(X, Y)$  with regard to (6.85) in reasonably short computing time.

Once the rejection region of the Bayesian test has been determined, its basic frequentist properties can easily be investigated by means of exact computational methods. In particular, the size of the critical region can be calculated. If the latter turns out to exceed the target significance level  $\alpha$ , the nominal level can (and should be) adjusted by diminishing it as far as necessary to obtain a test which is exactly valid with regard to its level of significance. As before [recall, e.g., Tab. 5.9], we denote the largest nominal significance level to be used in (6.85) for obtaining a critical region of size  $\leq \alpha$ , by  $\alpha^*$ . Once the posterior probability of  $\mathcal{K}_1$  has been computed for each point  $(x, y)$  in the sample space of  $(X, Y)$ ,  $\alpha^*$  can be determined in the same way as described in § 5.2.3 for the problem of testing for one-sided equivalence of binomial proportions with correlated observations. Table 6.23 is the analogue of Table 6.22 for the objective Bayesian test, except for the entries put in parentheses. The latter are all larger than 5% and give the size of the critical region of the test at uncorrected nominal level. All values shown in Table 6.23 were computed by means of another program to be found in the WKTSEQ2 Source Code Package under the program name `bi2by_ni_or`.

Table 6.23 *Nominal significance levels for the objective Bayesian test of (6.80) maintaining the 5%-level, for the same equivalence margins and sample sizes covered by Table 6.22. [Number in (): size of the test at uncorrected nominal level  $\alpha = .05$ .]*

$\varepsilon$	$n$	$\alpha^*$	Size	
1/3	10	.03547	.04401	(.06206)
"	25	.03895	.04370	(.06580)
"	50	.03992	.04565	(.06684)
"	75	.04023	.04551	(.06724)
"	100	.04039	.04545	(.06744)
"	200	.04063	.04535	(.06771)
1/2	10	.03978	.04028	(.07340)
"	25	.04123	.04927	(.07177)
"	50	.04245	.04559	(.07122)
"	75	.04258	.04833	(.07104)
"	100	.04264	.04776	(.07094)
"	200	.04273	.04705	(.07081)

Figure 6.6 shows the result of overlaying the power curve for the Bayesian test at nominal level  $\alpha^* = .04264$  for  $p_1(1 - p_2)/(p_2(1 - p_1)) \leq 1/2$  vs.  $p_1(1 - p_2)/(p_2(1 - p_1)) > 1/2$  based on two samples of size 100 each, with the dashed line of Figure 6.5. From a practical point of view, the differences between both curves are negligible so that the objective Bayesian approach proves once more to be a promising way of constructing nonrandomized tests for discrete families of distributions.

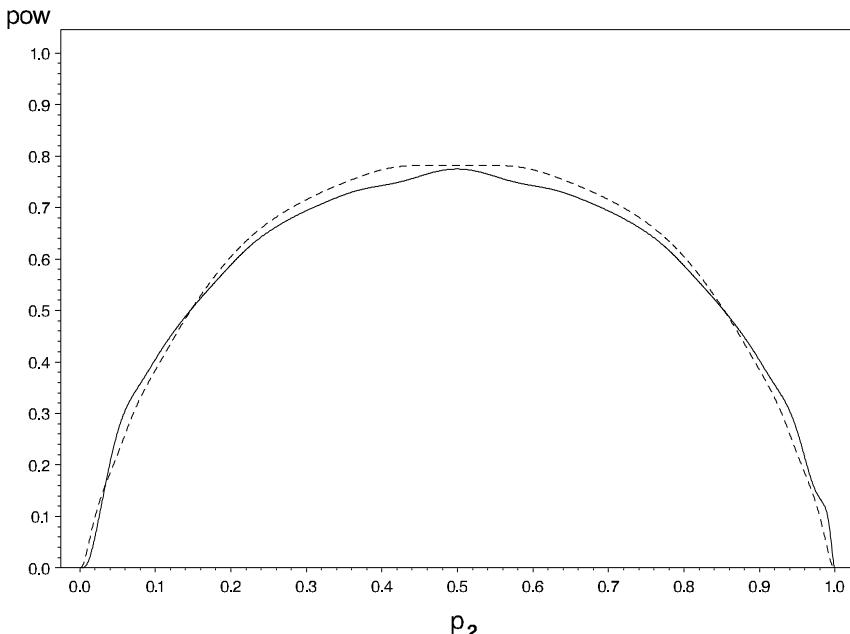


Figure 6.6 *Power function of the Bayesian test (—) for noninferiority with respect to the odds ratio as compared to that of the improved nonrandomized conditional test of Fig. 6.5 (---).* (From Wellek, 2005, with kind permission by Wiley-VCH.)

### 6.6.3 Tests for noninferiority using alternative parametrizations

Although there are large fields of application (epidemiology being perhaps the most important of them) where considerations of odds ratios play a predominant role, not a few users of equivalence testing methods prefer to look at the difference  $\delta = p_1 - p_2$  between both binomial parameters. Notwithstanding the logical difficulties encountered in defining equivalence in terms of this apparently more intuitive measure of treatment effect [recall p. 14], the literature contains numerous contributions dealing with the corresponding nonin-

feriority testing problem (among many others see Makuch and Simon, 1978; Blackwelder, 1982; Rodary et al., 1989; Roebruck and Kühn, 1995; Chan, 1998; Chan and Zhang, 1999). Another popular option for choosing the target parameter is the relative risk  $p_1/p_2$  (see, e.g., Miettinen and Nurminen, 1985; Farrington and Manning, 1990; Blackwelder, 1993; Chan, 1998; Nam, 1998). Each of these choices leads to a different form of the indifference zone by which the alternative hypothesis  $\{(p_1, p_2) \mid 0 < p_2 < p_1 < 1\}$  of the classical one-sided testing problem about the two binomial parameters is enlarged when the question of noninferiority is raised.

In principle, there are infinitely many ways of defining such an indifference zone. Preferably, the definition will be based on some parametric function  $g(\cdot, \cdot)$  with constant value  $g_0$  on the diagonal  $\{(p, p) \mid 0 < p < 1\}$  in the unit square and  $g(p_1, p_2) < g_0$  for  $p_1 < p_2$ . In order to keep the corresponding testing problem

$$H_1^g : g(p_1, p_2) \leq g_0 - \varepsilon \text{ vs. } K_1^g : g(p_1, p_2) > g_0 - \varepsilon \quad (6.86)$$

tractable, the real-valued transformation  $g(\cdot, \cdot)$  of the parameter space  $(0, 1) \times (0, 1)$  of  $(p_1, p_2)$  into the real line will usually be chosen to be continuous and strictly increasing (decreasing) in its first (second) argument.

Under the additional assumption that  $g$  is totally differentiable with  $g'_1, g'_2$  as its two first-order partial derivatives, it seems reasonable to base a test for (6.86) on Wald's statistic (cf. Rao, 1973, § 6e.3) which in the present setting admits the simple explicit representation

$$W_g(X, Y) = \frac{g(\hat{p}_1, \hat{p}_2) - g_0 + \varepsilon}{[(g'_1(\hat{p}_1, \hat{p}_2))^2 \hat{p}_1(1 - \hat{p}_1)/m + (g'_2(\hat{p}_1, \hat{p}_2))^2 \hat{p}_2(1 - \hat{p}_2)/n]^{1/2}}, \quad (6.87)$$

where  $\hat{p}_1$  and  $\hat{p}_2$  denotes as usual the relative frequency of “successes” observed under treatment  $A$  and  $B$ , respectively. In order to construct a nonconditional test of exact level  $\alpha$  based on (6.87) one has to determine the smallest  $w \in \mathbb{R}$  such that  $\sup_{(p_1, p_2) \in H_1^g} P_{p_1, p_2}[W_g(X, Y) > w] \leq \alpha$  and use this  $w$  as a critical constant. As long as simple explicit formulae are available for both partial derivatives of  $g$ , adopting this technique (the basic idea behind it goes back to Suissa and Shuster, 1985) is a straightforward exercise. However, even on a high-speed computer, determination of the critical constant  $w_g(\varepsilon, \alpha)$ , say, of a test of that type with satisfactory numerical accuracy might be a very time-consuming process except for very small sample sizes  $m$  and  $n$ . The reason is that it is in general not sufficient to restrict maximization of the rejection probability  $P_{p_1, p_2}[W_g(X, Y) > w_g(\varepsilon; \alpha)]$  to the boundary  $\partial H_1^g = \{(p_1, p_2) \mid g(p_1, p_2) = g_0 - \varepsilon\}$  of  $H_1^g$ . In fact, except for parametrizations  $g$  satisfying rather strong additional conditions (as derived by Röhmel and Mannsmann, 1999a), the full size  $\sup_{(p_1, p_2) \in H_1^g} P_{p_1, p_2}[W_g(X, Y) > w_g(\varepsilon; \alpha)]$  of the rejection region under consideration will be greater than the value found

by maximization over  $\partial H_1^g$  only. For some of the exact nonconditional tests for one-sided equivalence of two binomial distributions studied in the literature, it is only evident that they provide control over the significance level in this weaker sense whereas proving validity over the null hypothesis as a whole is a technically challenging problem. Perhaps the best known example of that kind is the test proposed by Chan for  $p_1 - p_2 \leq -\varepsilon$  vs.  $p_1 - p_2 > -\varepsilon$  (see Chan, 1998; Chan and Zhang, 1999; Chan, 2003) whose validity was initially questioned by Röhmel and Mannsmann (1999b) and later on proved by Röhmel (2005).

The objective Bayesian approach which we used in the previous subsection as an alternative method for constructing powerful nonrandomized tests for noninferiority with respect to the odds ratio, can likewise be adopted for any other parametric function  $g(\cdot, \cdot)$  such that  $p_1 \rightarrow g(p_1, p_2)$  has a well defined inverse  $g^{-1}(\cdot; p_2)$ , say, for arbitrarily fixed  $p_2 \in (0, 1)$ . Under this condition, the critical inequality of the objective Bayesian test for (6.86) can be written

$$\int_0^1 \left\{ \left[ 1 - I_{\max\{0, g^{-1}(g_0 - \varepsilon; p_2)\}}(x + 1/2, m - x + 1/2) \right] \cdot \frac{\Gamma(n+1)}{\Gamma(y+1/2)\Gamma(n-y+1/2)} \cdot p_2^{y-1/2} (1-p_2)^{n-y-1/2} \right\} dp_2 \geq 1 - \alpha, \quad (6.88)$$

generalizing (6.85) in the obvious way.

In the remainder of this subsection, we confine ourselves to apply both approaches to the special case that noninferiority shall be established in terms of the nonscaled difference  $g(p_1, p_2) = p_1 - p_2 \equiv \delta$  of proportions. With this choice of  $g(\cdot, \cdot)$  we obviously have  $g'_1 = g'_2 = 1$ , and (6.87) simplifies to

$$W_\delta(X, Y) = \frac{X/m - Y/n + \varepsilon}{\left[ (1/m)(X/m)(1-X/m) + (1/n)(Y/n)(1-Y/n) \right]^{1/2}}. \quad (6.89)$$

In view of the asymptotic normality of this test statistic, a convenient way of determining the corresponding lower critical bound  $w_\delta(\varepsilon; \alpha)$  is through iteration over the nominal significance level  $\alpha^*$  required in a test with rejection region  $\{W_\delta(X, Y) > u_{1-\alpha^*}\}$ . Denoting the probability of this region under any parameter constellation  $(p_1, p_2) \in (0, 1)^2$  by  $Q_{m,n}^{\varepsilon; \alpha^*}(p_1, p_2)$ , we can write

$$Q_{m,n}^{\varepsilon; \alpha^*}(p_1, p_2) = \sum_{y=0}^n \left\{ \sum_{x \in \mathcal{X}_m^{\varepsilon; \alpha^*}(y)} b(x; m, p_1) \right\} b(y; n, p_2), \quad (6.90)$$

where  $\mathcal{X}_m^{\varepsilon; \alpha^*}(y)$  stands for the  $y$ -section of  $\{(\tilde{x}, \tilde{y}) \mid 0 \leq \tilde{x} \leq m, 0 \leq \tilde{y} \leq n, W_\delta(\tilde{x}, \tilde{y}) > u_{1-\alpha^*}\}$ .

As to the form of the latter, we are in a similar situation as we encountered in §5.2.2 [recall the footnote on p. 81]: In contrast to the statistic obtained from  $W_\delta(X, Y)$  by replacing the unrestricted variance estimator with the pooled estimator,  $W_\delta(X, Y)$  itself does not satisfy the convexity condition

usually (see, e.g., Röhmel and Mannsmann, 1999a) named after Barnard from which it could be inferred that the  $\mathcal{X}_m^{\varepsilon;\alpha^*}(y)$  are left-bounded intervals in the sample space of  $X$ . However, in extensive numerical studies based on (6.90), we never found an exception from the rule that for each  $y = 0, 1, \dots, n$ , there is some integer  $k_y \in \{0, 1, \dots, n\}$  such that

$$\mathcal{X}_m^{\varepsilon;\alpha^*}(y) = \{x \in \mathbb{N}_0 \mid k_y \leq x \leq m\}. \quad (6.91)$$

From the basic properties of the binomial distribution, it follows that whenever (6.91) holds true, the maximum of the rejection probability of the test with critical region  $\{W_\delta(X, Y) > u_{1-\alpha^*}\}$  over the null hypothesis  $H_1 : p_1 - p_2 \leq -\varepsilon$  is taken on the boundary of the latter which allows one to restrict numerical search to the line  $\{(p_2 - \varepsilon, p_2) \mid \varepsilon < p_2 < 1\}$  in determining the size of the test. In other words, taking the validity of (6.91) for granted, we can write

$$SIZE_{m,n}^{\varepsilon;\alpha^*}(W_\delta) = \sup_{\varepsilon < p_2 < 1} Q_{m,n}^{\varepsilon;\alpha^*}(p_2 - \varepsilon, p_2) \quad (6.92)$$

for any choice of the nominal significance level  $0 < \alpha^* < 1$ . The SAS/IML program `bi2wld_ni_del` by means of which the entries in Table 6.24 have been obtained, evaluates (6.92) iteratively for a sequence of values of  $\alpha^*$ ,

Table 6.24 *Nominal significance level and exact size of a corrected version of the Wald type test for noninferiority in terms of  $\delta$  in the binomial two-sample setting with  $m = n = 25(25)200$  and equivalence margin  $\varepsilon \in \{.10, .15\}$ . [Number in (): size of the critical region of the noncorrected asymptotic testing procedure.]*

$\varepsilon$	$n$	$\alpha^*$	$SIZE_{m,n}^{\varepsilon;\alpha^*}(W_\delta)$
.10	25	.03560	(.08965)
"	50	.01519	(.11173)
"	75	.00400	(.11894)
"	100	.01089	(.11716)
"	125	.01619	(.07138)
"	150	.01956	(.06905)
"	175	.02146	(.09943)
"	200	.02233	(.09295)
.15	25	.00250	(.09307)
"	50	.01855	(.11211)
"	75	.01272	(.10816)
"	100	.01801	(.09945)
"	125	.02021	(.08990)
"	150	.02754	(.08069)
"	175	.03735	(.07315)
"	200	.03531	(.06816)

starting from  $\alpha^* = .05$  and ending after the maximum allowable number of iteration steps with the largest trial value  $\alpha^* < .05$  such that  $SIZE_{m,n}^{\varepsilon;\alpha^*}(W_\delta)$  does not exceed the target significance level.

Specialization of the critical inequality (6.88) of the objective Bayesian test to the case  $g(p_1, p_2) = p_1 - p_2$  leads, after some simplification, to the condition

$$\frac{I_\varepsilon(y + 1/2, n - y + 1/2) + \int_\varepsilon^1 \left\{ \left[ 1 - I_{p_2-\varepsilon}(x + 1/2, m - x + 1/2) \right] \cdot \frac{\Gamma(n+1)}{\Gamma(y+1/2)\Gamma(n-y+1/2)} \cdot p_2^{y-1/2} (1-p_2)^{n-y-1/2} \right\} dp_2}{\Gamma(y+1/2)\Gamma(n-y+1/2)} \geq 1 - \alpha. \quad (6.93)$$

From the properties of the incomplete beta integral (see, e.g., Johnson et al., 1995, p. 246, Eq. (25.72a,b)), it follows that  $I_q(x + 1/2, m - x + 1/2)$  is a decreasing function of  $x$  for any  $q \in (0, 1)$ . Thus, replacing the rejection region of the Wald type with that of the objective Bayesian test for  $\delta$ -noninferiority, the result that all  $y$ -sections of this region are left-bounded intervals in the sample space of  $X$  [recall (6.91)] can even be stated as a general mathematical fact. Hence, essentially the same algorithm can be used for determining the largest nominal level  $\alpha^*$  by which  $\alpha$  has to be replaced in (6.93) in order to ensure that the corresponding test for  $\delta$ -noninferiority maintains the target significance level in the usual, i.e., frequentist sense. The only modification required refers to the test statistic which now has to be computed through evaluating the left-hand side of (6.93). As before in the objective Bayesian construction of an improved nonrandomized test for noninferiority in terms of the odds ratio, Gauss-Legendre integration is a particularly efficient numerical tool for that purpose. At the accompanying website, the SAS/IML code of a program implementing this algorithm can be found in the file named `bi2by_ni_del`. All nominal levels shown in Table 6.25 being the analogue of Table 6.24 for the objective Bayesian approach, were obtained by running this program. As in the case of testing for noninferiority in terms of the odds ratio, it is tempting to compare the power curves of the objective Bayesian and the proposed frequentist solution to the problem of testing for  $\delta$ -noninferiority in the binomial two-sample setting. Of course, exact computation is possible for both tests also at that stage of analysis. Figure 6.7 shows the exact power functions for an balanced design with 100 observations per group and noninferiority margin  $\varepsilon = .15$  to  $\delta = p_1 - p_2$ . The most remarkable conclusion to be taken from this graph is that the level-adjusted objective Bayesian test turns out to be uniformly more powerful as compared with the Wald type test maintaining the same target significance level. The relevance of the fact that the power function of the former converges to unity at both boundaries of the parameter space of  $p_2$  whereas the power of the test based on the maximum likelihood statistics drops to arbitrary small values as  $p_2 \rightarrow 0+$  should not be overestimated. Actually, it simply reflects noninclusion of a single point in the sample space, viz.,  $(X, Y) = (0, 0)$ , in the rejection region of the Wald type test.

Table 6.25 *Nominal significance levels for the objective Bayesian test of  $p_1 - p_2 \leq -\varepsilon$  vs.  $p_1 - p_2 > -\varepsilon$  maintaining the 5%-level, for the same values of the noninferiority margin  $\varepsilon$  and sample sizes covered by Table 6.24. [Number in (): size of the test at uncorrected nominal level  $\alpha = .05$ .]*

$\varepsilon$	$n$	$\alpha^*$	$SIZE_{m,n}^{\varepsilon;\alpha^*}(BAY_\delta)$	
.10	25	.01400	.02377	(.07179)
"	50	.04433	.04997	(.11132)
"	75	.02214	.03027	(.05700)
"	100	.02920	.03593	(.06185)
"	125	.03307	.03782	(.06310)
"	150	.03492	.04054	(.06236)
"	175	.03550	.04355	(.06051)
"	200	.03525	.04238	(.06280)
.15	25	.03154	.03906	(.09307)
"	50	.03233	.04605	(.06538)
"	75	.02815	.03311	(.06650)
"	100	.03135	.04028	(.06369)
"	125	.03182	.03891	(.05923)
"	150	.03628	.04900	(.05913)
"	175	.04304	.04944	(.07171)
"	200	.04292	.04810	(.06529)

#### 6.6.4 Exact test for two-sided equivalence with respect to the odds ratio

In the remaining parts of this section, we keep supposing that we are given a  $2 \times 2$  contingency table of the form shown in Table 6.20 [ $\rightarrow$  p. 173]. But now we are interested in establishing two-sided equivalence of the underlying binomial distributions rather than noninferiority of  $\mathcal{B}(m, p_1)$  to  $\mathcal{B}(n, p_2)$ . Measuring, as in § 6.6.1, the distance between both distributions in terms of  $|\varrho - 1|$  with  $\varrho = p_1(1 - p_2)/(1 - p_1)p_2$ , and  $p_1$  and  $p_2$  as the probability of a favorable response to treatment  $A$  and  $B$ , respectively, the problem of testing for equivalence in the strict sense reads

$$H : 0 < \varrho \leq \varrho_1 \text{ or } \varrho_2 \leq \varrho < \infty \quad \text{versus} \quad K : \varrho_1 < \varrho < \varrho_2. \quad (6.94)$$

Of course, the limits  $\varrho_1, \varrho_2$  of the equivalence range for the odds ratio  $\varrho$  have to be specified as positive real numbers satisfying  $\varrho_1 < 1 < \varrho_2$ . If we keep denoting the number of experimental units responding to treatment  $A$  and  $B$

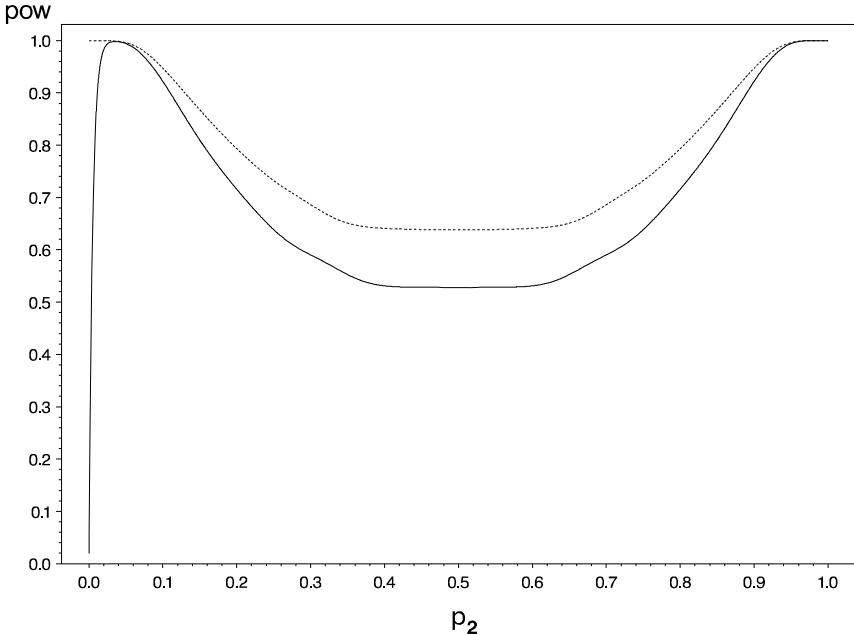


Figure 6.7 *Power function of the asymptotic test for one-sided  $\delta$ -equivalence at corrected nominal level  $\alpha_0 = .01801$  [—] as compared with that of the objective Bayesian test for of the same hypothesis [---].* (From Wellek, 2005, with kind permission by Wiley-VCH.)

by  $X$  and  $Y$ , respectively, then independence of  $X$  and  $Y$  implies that the joint probability mass function of both variables can be written

$$\begin{aligned} P[X = x, Y = y] &= \binom{m}{x} \binom{n}{y} (1 - p_1)^m (1 - p_2)^n \cdot \\ &\quad \exp \left\{ x \log \left[ \frac{p_1(1 - p_2)}{(1 - p_1)p_2} \right] + (x + y) \log[p_2/(1 - p_2)] \right\}, \\ &\quad (x, y) \in \{0, \dots, m\} \times \{0, \dots, n\}. \end{aligned} \quad (6.95)$$

Letting  $\theta = \log \varrho = \log[p_1(1 - p_2)/(1 - p_1)p_2]$ ,  $\vartheta = \log[p_2/(1 - p_2)]$ , it is easily verified that we have  $1 - p_1 = (1 + e^{\theta+\vartheta})^{-1}$ ,  $1 - p_2 = (1 + e^\vartheta)^{-1}$  and hence  $(1 - p_1)^m (1 - p_2)^n = (1 + e^{\theta+\vartheta})^{-m} (1 + e^\vartheta)^{-n}$ . If the joint density of  $(X, Y)$  is defined with respect to the measure  $B^{(2)} \mapsto \sum_{(x,y) \in B^{(2)} \cap \{0, \dots, m\} \times \{0, \dots, n\}} \binom{m}{x} \binom{n}{y}$  on the  $\sigma$ -algebra of Borel sets in  $\mathbb{R}^2$  rather than the counting measure of the set  $\{0, 1, \dots, m\} \times \{0, 1, \dots, n\}$ , we thus obtain the expression

$$p_{\theta, \vartheta}(x, y) = (1 + e^{\theta+\vartheta})^{-m} (1 + e^\vartheta)^{-n} \exp\{x\theta + (x + y)\vartheta\}. \quad (6.96)$$

Obviously, (6.96) defines the generic element of a family of densities of the form considered in § A.2 [see in particular equation (A.16)] with  $k = 1, T = X, S = X + Y$ . In other words, the densities (6.96) make up a two-parameter exponential family in  $\theta = \log \varrho, \vartheta = \log[p_2/(1 - p_2)]$  with  $X$  and  $S = X + Y$  as a statistic sufficient for  $\theta$  and  $\vartheta$ , respectively.

Hence, Theorem A.2.2 [ $\rightarrow$  p. 375] does apply implying that a uniformly most powerful unbiased (UMPU) level- $\alpha$  test for (6.94) is obtained by proceeding for each of the possible values  $s \in \{0, 1, \dots, m+n\}$  of the first column total  $S$  according to the following decision rule:

$$\left\{ \begin{array}{ll} \text{Rejection of } H & \text{if } C_{1;\alpha}^{m,n}(s; \varrho_1, \varrho_2) < \\ & X < C_{2;\alpha}^{m,n}(s; \varrho_1, \varrho_2) \\ \text{Reject with prob. } \gamma_{1;\alpha}^{m,n}(s; \varrho_1, \varrho_2) & \text{if } X = C_{1;\alpha}^{m,n}(s; \varrho_1, \varrho_2) \\ \text{Reject with prob. } \gamma_{2;\alpha}^{m,n}(s; \varrho_1, \varrho_2) & \text{if } X = C_{2;\alpha}^{m,n}(s; \varrho_1, \varrho_2) \\ \text{Acceptance of } H & \text{if } X < C_{1;\alpha}^{m,n}(s; \varrho_1, \varrho_2) \\ & \text{or } X > C_{2;\alpha}^{m,n}(s; \varrho_1, \varrho_2) \end{array} \right. . \quad (6.97)$$

The critical constants  $C_{\nu;\alpha}^{m,n}(s; \varrho_1, \varrho_2), \gamma_{\nu;\alpha}^{m,n}(s; \varrho_1, \varrho_2), \nu = 1, 2$ , have to be determined for each  $s$  by solving the equations

$$\sum_{x=C_1+1}^{C_2-1} h_s^{m,n}(x; \varrho_1) + \sum_{\nu=1}^2 \gamma_{\nu} h_s^{m,n}(C_{\nu}; \varrho_1) = \alpha = \sum_{x=C_1+1}^{C_2-1} h_s^{m,n}(x; \varrho_2) + \sum_{\nu=1}^2 \gamma_{\nu} h_s^{m,n}(C_{\nu}; \varrho_2), \quad 0 \leq C_1 \leq C_2 \leq m, \quad 0 \leq \gamma_1, \gamma_2 < 1. \quad (6.98)$$

For any value  $\varrho \in \mathbb{R}_+$  taken on by the true population odds ratio  $p_1(1 - p_2)/(1 - p_1)p_2$ ,  $h_s^{m,n}(\cdot; \varrho)$  stands for the probability mass function of the distribution of  $X$  conditional on the event  $\{S = s\}$ . As was already stated in § 6.6.1, this conditional distribution is a so-called extended hypergeometric distribution given by

$$h_s^{m,n}(x; \varrho) = \binom{m}{x} \binom{n}{s-x} \varrho^x / \sum_{j=\max\{0, s-n\}}^{\min\{s, m\}} \binom{m}{j} \binom{n}{s-j} \varrho^j, \\ \max\{0, s-n\} \leq x \leq \min\{s, m\}. \quad (6.99)$$

In the balanced case of equal sample sizes  $m = n$ , it is easy to verify that the conditional distributions of  $X$  satisfy the symmetry relation

$$h_s^{n,n}(s-x; \varrho^{-1}) = h_s^{n,n}(x; \varrho), \\ \max\{0, s-n\} \leq x \leq \min\{s, n\}, \quad s = 0, 1, \dots, 2n. \quad (6.100)$$

In view of (6.100) and Corollary A.1.7 [ $\rightarrow$  p. 372–3], in the case of a balanced design with common size  $n$  of both samples and a symmetric equivalence

range, say  $(\varrho_0^{-1}, \varrho_0)$  for the population odds ratio  $\varrho$ , the decision rule (6.97) defining a UMPU level- $\alpha$  test for (6.94) can be simplified to

$$\begin{cases} \text{Rejection of } H & \text{if } |X - s/2| < C(n, s; \varrho_0, \alpha) \\ \text{Reject with prob. } \gamma(n, s; \varrho_0, \alpha) & \text{if } |X - s/2| = C(n, s; \varrho_0, \alpha) \\ \text{Acceptance of } H & \text{if } |X - s/2| > C(n, s; \varrho_0, \alpha) \end{cases}. \quad (6.101)$$

Of the two critical constants appearing in (6.101), the first one has to be determined as

$$C(n, s; \varrho_0, \alpha) = C^* = \max \left\{ C \left| \begin{array}{l} s/2 - C, s/2 + C \in \mathbb{N}_0, s/2 - C \geq \\ \max\{0, s - n\}, s/2 + C \leq \min\{s, n\}, \sum_{x=s/2-C+1}^{s/2+C-1} h_s^{n,n}(x; \varrho_0) \leq \alpha \end{array} \right. \right\}. \quad (6.102)$$

As soon as the correct value of the critical upper bound  $C^* \equiv C(n, s; \varrho_0, \alpha)$  to  $|X - s/2|$  has been found, the randomization probability  $\gamma(n, s; \varrho_0, \alpha)$  can be computed by means of the explicit formula

$$\gamma(n, s; \varrho_0, \alpha) = \left( \alpha - \sum_{x=s/2-C^*+1}^{s/2+C^*-1} h_s^{n,n}(x; \varrho_0) \right) / \left( h_s^{n,n}(s/2 - C^*; \varrho_0) + h_s^{n,n}(s/2 + C^*; \varrho_0) \right). \quad (6.103)$$

Furthermore, the condition for sure rejection of the null hypothesis given in (6.101) can alternatively be expressed in terms of a conditional p-value to be computed as

$$p_{n; \varrho_0}(x|s) = \sum_{j=s-\tilde{x}_s}^{\tilde{x}_s} h_s^{n,n}(j; \varrho_0) \quad (6.104)$$

where  $h_s^{n,n}(x; \varrho_0)$  is obtained by specializing (6.99) in the obvious way, and the upper summation limit is given by

$$\tilde{x}_s = \max\{x, s - x\}. \quad (6.105)$$

In fact, it is readily verified that the observed value  $x$  of  $X$  satisfies  $|x - s/2| < C(n, s; \varrho_0, \alpha)$  if and only if the conditional p-value (6.104) does not exceed  $\alpha$ .

The system (6.98) of equations which determines the critical constants of the exact Fisher type equivalence test in the general case, is of the same form as that which had to be treated in a previous chapter [see p. 60, (4.19)] in the context of the one-sample binomial test for equivalence. Correspondingly, the algorithm proposed for solving the latter can be adopted for handling (6.98) with comparatively few modifications. Of course, the most conspicuous change is that the functions  $x \mapsto h_s^{m,n}(x; \varrho_\nu)$  have to replace the binomial probability mass functions  $x \mapsto b(x; n, p_\nu)$  ( $\nu = 1, 2$ ). Since the equivalence

range for  $\varrho$  was assumed to be an interval covering unity, we can expect that an interval  $(C_1, C_2)$  satisfying (6.98) when combined with a suitable  $(\gamma_1, \gamma_2) \in [0, 1]^2$ , contains that point  $x_o \in [\max\{0, s - n\}, \min\{s, m\}]$  which, as an entry into a contingency table with first column total  $s$ , yields an observed odds ratio exactly equal to 1. Elementary algebra shows that this number is  $x_o = ms/(m + n)$ . For the purpose of searching iteratively for the solution of (6.98),  $C_1^o = \lceil x_o \rceil + 5$  proved to be a suitable choice of an initial value of the left-hand critical bound to  $X$  in numerous trials of the algorithm. Once more, both SAS and R allow fairly easy implementation of such a numerical procedure. The reason is that in both systems predefined functions for the cdf of any extended hypergeometric distribution with probability mass function given by (6.99) are available [recall p. 188]. The source codes of a program computing for any  $s \in \{0, 1, \dots, m + n\}$  the critical constants of the optimal conditional test for equivalence of  $\mathcal{B}(m, p_1)$  and  $\mathcal{B}(n, p_2)$  with respect to the odds ratio, can be found in the **WKTSEQ2 Source Code Package** under the main file name **bi2st**.

The computational issues to be tackled in connection with power analysis and sample size determination for the exact Fisher type test for equivalence are basically the same as those we had to consider in § 6.6.1 in connection with the noninferiority version of the procedure. The major modification to be effected in the two-sided case is that in the numerator of the expression for the conditional power [ $\rightarrow$  p. 175, (6.82)],  $\{j \geq k_\alpha(s) + 1\}$  has to be replaced with the summation region  $\{C_{1; \alpha}^{m, n}(s; \varrho_1, \varrho_2) + 1 \leq j \leq C_{2; \alpha}^{m, n}(s; \varrho_1, \varrho_2) - 1\}$  bounded on both sides, and the properly weighted probability of the single point  $k_\alpha(s)$  with a weighted sum of the probabilities of the two boundary points of the  $s$ -section of the critical region. As before, the nonconditional power is obtained by integrating the function  $s \mapsto \beta(p_1^*, p_2^*|s)$  assigning to each  $s \in \{0, 1, \dots, m + n\}$  the conditional power given  $\{S = s\}$ , with respect to the distribution of  $S = X + Y$  under the specific alternative  $(p_1^*, p_2^*) \in (0, 1)^2$  of interest. In view of the large number of steps involving computation of extended hypergeometric and binomial probabilities, respectively, the results obtained by means of this algorithm will be numerically reliable only if the precision of each intermediate result is kept as high as possible. This is again a reason for having recourse to maximum floating-point precision being available within the programming environment used for its implementation. Primarily, the programs provided in the **WKTSEQ2 Source Code Package** for computing power and minimally required sample sizes for the Fisher type test for equivalence of two binomial distributions were coded in Fortran using `real*16` floating-point arithmetic. The procedure named **bi2aeq1** performs the task of computing the power against any specific alternative  $(p_1^*, p_2^*) \in \{(p_1, p_2) \in (0, 1)^2 | (p_1(1 - p_2)/(1 - p_1)p_2) \in (\varrho_1, \varrho_2)\}$ . It accommodates total sample sizes  $N = m + n$  of up to 2000. The other program, **bi2aeq2**, determines for a given specific alternative  $(p_1^*, p_2^*)$  and prespecified power  $\beta^*$  to be attained against it, the smallest sample sizes  $(m^*, n^*)$  required to guarantee that the rejection probability of the test does not fall short of

$\beta^*$ , subject to the side condition that the ratio of both sample sizes equals some arbitrarily fixed positive value  $\lambda$ . Except for the iterative determination of the critical constants to be used at each possible value of the conditioning statistic  $S$  and the above-mentioned modifications referring to the computation of the conditional power, the Fortran program `bi2aeq1` and `bi2aeq2` is a direct analogue of the corresponding program `bi2ste1` and `bi2ste2` for the exact Fisher type test for noninferiority, respectively. As in the case of testing for one-sided equivalence, of both programs, a double-precision version has been made available as a DLL which can be called from R. Again, diminishing the numerical precision of the individual arithmetic operations affects the final results only at decimal places being largely irrelevant for practical applications.

### Example 6.5

In a test preliminary to the final confirmatory assessment of the data obtained from a controlled comparative multicenter trial of the calcium blocking agent verapamil and a classical diuretic drug with respect to antihypertensive efficacy, it was the aim to rule out the possibility that patients with and without some previous antihypertensive treatment differ to a relevant extent in the probability of a favorable response to the actual medication. Table 6.26 shows the frequencies of responders and nonresponders observed in both strata of the study population during an 8 weeks' titration period. The equivalence range for the odds ratio  $\varrho$  characterizing the underlying (sub-)populations was set to  $(\varrho_1, \varrho_2) = (.6667, 1.5000)$ , following the tentative guidelines of § 1.7 for specifying the tolerances in formulating equivalence hypotheses about target parameters and functionals frequently studied in practice. The significance level was chosen as usual specifying  $\alpha = .05$ .

Table 6.26 Responder rates observed in the VERDI trial (Holzgreve et al., 1989) during an 8 weeks' titration period in patients with and without previous antihypertensive treatment.

Previous treatment with anti-hypertensive drugs	Response to trial medication			$\Sigma$
	+	-		
yes	108 (48.00%)	117 (52.00%)	225 (100.0%)	
no	63 (52.94%)	56 (47.06%)	119 (100.0%)	
$\Sigma$	171	173	344	

With  $m = 225$ ,  $n = 119$ ,  $s = 171$  and the specified values of the  $\varrho_{\nu}$ , the program **bi2st** outputs the critical interval (110, 113) and the probabilities .0214, .6322 of a randomized decision in favor of equivalence to be taken if it happens that  $X = 110$  and  $X = 113$ , respectively. Since the observed value of  $X$  remained below the left-hand limit of the critical interval, the data shown in Table 6.26 do not allow to reject the null hypothesis that there are relevant differences between patients with and without previous antihypertensive treatment with respect to the probability of a positive response to the study medication. Running program **bi2aeq1**, the power of the UMPU test against the specific alternative that the true responder rates  $p_1$  and  $p_2$  coincide with the observed proportions .4800 and .5294, is computed to be as small as 16.19%. If the nonrandomized conservative version of the test is used this value drops even to 10.70%.

A selection of results on sample sizes required in the exact Fisher type test for equivalence to maintain some prespecified power against fixed alternatives on the diagonal of the unit square are shown in Table 6.27. All entries into the rightmost columns have been computed by means of program **bi2aeq2**. One interesting fact becoming obvious from these results is that the efficiency of the test decreases with the distance of the common value of  $p_1$  and  $p_2$  from the center of the unit interval. Furthermore, the order of magnitude of the sample sizes appearing in the lower half of the table corroborates the view that the requirement  $2/3 < \varrho < 3/2$  corresponds to a rather strict criterion

Table 6.27 *Sample sizes required in the exact Fisher type test for equivalence at level  $\alpha = .05$  to maintain a prespecified power against selected alternatives  $p_1 = p_2 = p_*$  on the diagonal of the unit square, for the balanced [ $\leftrightarrow \lambda = m/n = 1$ ] and various unbalanced designs [ $\leftrightarrow \lambda > 1$ ].*

$(\varrho_1, \varrho_2)$	$p_*$	POW	$\lambda$	$m$	$n$	$N = m + n$
(.4286, 2.3333) <sup>†</sup>	.50	.80	1.00	98	98	196
"	.40	"	1.00	102	102	204
"	.30	"	1.00	117	117	234
"	.20	"	1.00	156	156	312
"	.10	"	1.00	281	281	562
(.6667, 1.5000)	.50	.60	1.00	302	302	604
"	.50	"	1.25	340	272	612
"	.50	"	1.50	377	251	628
"	.50	"	2.00	454	227	681
"	.50	"	3.00	606	202	808

<sup>†</sup>  $\approx (3/7, 7/3)$  – cf. pp. 16-7

of equivalence of two binomial distributions, as suggested in § 1.7 [recall Table 1.1 (iii)]. Finally, it is worth noticing that keeping both the equivalence range  $(\varrho_1, \varrho_2)$  and the specific alternative of interest fixed, the total sample size  $N = m + n$  increases with the ratio  $\lambda$  of the larger over the smaller of both group sizes.

### 6.6.5 An improved nonrandomized version of the UMPU test for two-sided equivalence

Nonrandomized versions of UMPU tests for discrete families of distributions obtained by incorporating the whole boundary of the rejection into the acceptance region, are notoriously conservative except for unusually large sample sizes. This is in particular true for the exact Fisher type test for equivalence in the strict, i.e., two-sided sense of two binomial distributions as presented in the previous subsection. Fortunately, it will turn out that the conceptually very simple trick of raising the nominal conditional significance level by the maximum allowable amount, which leads to considerable improvements to the nonrandomized Fisher type test for the noninferiority setting [recall § 6.6.2], works in the present context as well. Apart from showing also the power of both the conventional and the improved nonrandomized version of the UMPU test, Table 6.28 is the direct two-sided analogue of Table 6.22 of § 6.6.2. It refers likewise to balanced designs with common sample sizes ranging over the whole spectrum of what is realistic for practical applications. The upper and the lower half of the table refers to what has been proposed in § 1.7 as a strict and liberal choice of a symmetric equivalence range for the odds ratio, respectively.

All values appearing in the third column of Table 6.28 as a maximally raised nominal significance level  $\alpha^*$  have been computed by means of another Fortran program named `bi2aeq3` which is the analogue of `bi2ste3` for equivalence testing in the strict sense. The program enables its user to find maximally increased nominal levels for arbitrary combinations of values of  $\alpha, \varrho_0, m$  and  $n$ , provided the total sample size  $N = m + n$  does not exceed 2000. As to the numerical accuracy attained in computing a maximally raised nominal level for any given configuration  $(\alpha, \varrho_0, m, n)$ , the situation is once more strictly analogous to that described in connection with the noninferiority case: there is always a whole interval of nominal levels over which the exact size of the nonrandomized conditional test remains constant so that it suffices to ensure that the resulting value is contained in that interval. Furthermore, a tolerance of 0.1% was specified throughout for a practically negligible anticonservatism of the improved nonrandomized test.

Table 6.28 *Nominal significance level, size and power against  $p_1 = p_2 = 1/2$  of an improved nonrandomized Fisher type test for the problem (6.94) maintaining significance level  $\alpha = 5\%$ , for  $\varrho_1 = \varrho_0^{-1}$ ,  $\varrho_2 = \varrho_0 = 1.5000, 2.3333$ , and  $m = n = 25(25)100(50)250$ . [Numbers in parentheses refer to the ordinary nonrandomized version of the exact Fisher type test using nominal level .05.]*

$\varrho_0$	$n$	$\alpha^*$	Size	Power against	
				$p_1 = p_2 = 1/2$	$p_1 = p_2 = 1/2$
1.5000	25	.17300	.03717 (.00000)	.04804 (.00000)	
	"	.12000	.05017 (.00000)	.07959 (.00000)	
	"	.12000	.04582 (.00000)	.06504 (.00000)	
	"	.08188	.05052 (.03915)	.13916 (.05635)	
	"	.06500	.05079 (.04059)	.22695 (.13747)	
	"	.05875	.05041 (.04153)	.34462 (.27361)	
	"	.05891	.05005 (.04199)	.47284 (.43801)	
2.3333	25	.14989	.04893 (.00000)	.13629 (.00000)	
	"	.07500	.05054 (.02910)	.38233 (.23560)	
	"	.06875	.05060 (.03576)	.63079 (.53742)	
	"	.06250	.05020 (.03834)	.81869 (.77034)	
	"	.05875	.04941 (.04064)	.95357 (.94298)	
	"	.06000	.04972 (.04104)	.98929 (.98580)	
	"	.05812	.04968 (.04352)	.99771 (.99721)	

Going through the fourth and fifth column of Table 6.28, one finds that even with sample sizes as large as 250 there remains a substantial margin for adjusting the size of the nonrandomized conditional test. In many cases, the gain in power achieved by raising the nominal over the target significance level exhibits an order of magnitude in view of which one has clearly to discourage from using the nonrandomized test in its conventional form for real applications.

### 6.6.6 Tests for two-sided equivalence with respect to the difference of success probabilities

In this subsection we consider the two-sided version of the testing problem upon which we focussed in § 6.6.3. In other words, we aim at establishing a satisfactory solution for the problem of testing

$$H : -1 < \delta \leq -\delta_1 \text{ or } \delta_2 \leq \delta < 1 \text{ versus } K : -\delta_1 < \delta < \delta_2 , \quad (6.106)$$

where

$$\delta = p_1 - p_2 , \quad (6.107)$$

and  $\delta_1, \delta_2$  denote fixed numbers to be chosen from the open unit interval.

In contrast to the odds ratio  $\varrho = p_1(1-p_2)/(1-p_1)p_2$ , the simple difference  $\delta$  of both success probabilities is not a mathematically natural parameter for the underlying class of products of two binomial distributions. Accordingly, we will not be able to exploit one of the major theoretical results presented in the Appendix for deriving an optimal solution to (6.106). Instead, we will adopt the basic idea behind the approach taken in § 5.2.2 to constructing a valid level- $\alpha$  test for equivalence with respect to  $\delta$  of two binomial distributions from which a sample of paired observations is available. In the case of two independent samples we have now to deal with, a test statistic for (6.106) which satisfies the conditions of Theorem A.3.4 is given by

$$\frac{|T_N - (\delta_2 - \delta_1)/2|}{\hat{\tau}_N} = \frac{|(X/m - Y/n) - (\delta_2 - \delta_1)/2|}{[(1/m)(X/m)(1 - X/m) + (1/n)(Y/n)(1 - Y/n)]^{1/2}}, \quad (6.108)$$

where the subscript  $N$  to which all limiting operations apply, denotes this time the total sample size  $m + n$ . Accordingly, the two-sample analogue of the decision rule (5.9) obtained in the McNemar setting, reads as follows:

Reject nonequivalence if and only if

$$\frac{|(X/m - Y/n) - (\delta_2 - \delta_1)/2|}{[(1/m)(X/m)(1 - X/m) + (1/n)(Y/n)(1 - Y/n)]^{1/2}} < C_\alpha \left( \frac{(\delta_1 + \delta_2)/2}{[(1/m)(X/m)(1 - X/m) + (1/n)(Y/n)(1 - Y/n)]^{1/2}} \right). \quad (6.109)$$

The function  $C_\alpha(\cdot)$  to be evaluated for determining the critical upper bound to the test statistic is defined as before so that the expression on the right-hand side of (6.109) stands for the square root of the  $\alpha$ -quantile of a  $\chi^2$ -distribution with 1 degree of freedom and (random) noncentrality parameter  $(\delta_1 + \delta_2)^2 / (4[(1/m)(X/m)(1 - X/m) + (1/n)(Y/n)(1 - Y/n)])$ . By Theorem A.3.4, we know that the testing procedure corresponding to (6.109) has asymptotic significance level  $\alpha$  whenever it can be taken for granted that the variance of the statistic  $\sqrt{N}T_N = \sqrt{N}(X/m - Y/n)$  converges to a positive limit as  $N \rightarrow \infty$ . Since we obviously have that  $\text{Var}[\sqrt{N}(X/m - Y/n)] = (N/m)p_1(1 - p_1) + (N/n)p_2(1 - p_2)$ , this simply requires that the relative size  $m/N$  of sample 1 converges to some nondegenerate limit and at least one of the two binomial distributions under comparison be nondegenerate. In other words, (6.109) defines a test of asymptotic level  $\alpha$  provided there exists some  $\lambda \in (0, 1)$  such that  $m/N \rightarrow \lambda$  as  $N \rightarrow \infty$ , and at least one of the primary parameters  $p_1$  and  $p_2$  is an interior point of the unit interval.

A straightforward approach to *transforming the asymptotic testing procedure (6.109) into an exactly valid test* for the problem (6.106) starts from establishing an algorithm for the exact computation of its rejection probability under any fixed parameter constellation  $(p_1, p_2)$ . For brevity, let us

write

$$\begin{aligned} Q_{\alpha;m,n}^{\delta_1,\delta_2}(p_1, p_2) &= \\ P_{(p_1, p_2)} \left[ \frac{|X/m - Y/n - (\delta_2 - \delta_1)/2|}{[(1/m)(X/m)(1 - X/m) + (1/n)(Y/n)(1 - Y/n)]^{1/2}} \right. \\ &\quad \left. < C_\alpha \left( \frac{(\delta_1 + \delta_2)/2}{[(1/m)(X/m)(1 - X/m) + (1/n)(Y/n)(1 - Y/n)]^{1/2}} \right) \right] . \end{aligned} \quad (6.110)$$

A representation of  $Q_{\alpha;m,n}^{\delta_1,\delta_2}(p_1, p_2)$  which proves particularly convenient for computational purposes reads

$$Q_{\alpha;m,n}^{\delta_1,\delta_2}(p_1, p_2) = \sum_{x=0}^m \left\{ \sum_{y \in \mathcal{U}_{\alpha,n}^{\delta_1,\delta_2}(x)} b(y; n, p_2) \right\} \cdot b(x; m, p_1) , \quad (6.111)$$

where

$$\begin{aligned} \mathcal{U}_{\alpha;m,n}^{\delta_1,\delta_2}(x) &= \left\{ y \in \mathbb{N}_0 \mid y \leq n, \right. \\ &\quad \frac{|x/m - y/n - (\delta_2 - \delta_1)/2|}{[(1/m)(x/m)(1 - x/m) + (1/n)(y/n)(1 - y/n)]^{1/2}} \\ &\quad \left. < C_\alpha \left( \frac{(\delta_1 + \delta_2)/2}{[(1/m)(x/m)(1 - x/m) + (1/n)(y/n)(1 - y/n)]^{1/2}} \right) \right\} . \end{aligned} \quad (6.112)$$

With regard to the form of the sets  $\mathcal{U}_{\alpha;m,n}^{\delta_1,\delta_2}(x)$  over which the inner sum has to be extended in computing exact rejection probabilities of the asymptotic test, the situation is quite the same as found in the analogous setting with paired observations: By explicit construction, the  $\mathcal{U}_{\alpha;m,n}^{\delta_1,\delta_2}(x)$  turn out to be intervals in the sample space of the second of the binomial variables involved. Since there is no argument in sight which allows to assert that this observation reflects a general mathematical fact, both computer programs **bi2diffac**, **bi2dipow** involving evaluations of exact rejection probabilities  $Q_{\alpha;m,n}^{\delta_1,\delta_2}(p_1, p_2)$  by means of (6.111) are organized as their paired-observations counterparts [cf. the footnote on p. 81].

Now, we are ready to determine once more the largest nominal significance level  $\alpha^*$  which has to be substituted for  $\alpha$  in order to ensure that the exact size of the asymptotic testing procedure does not exceed the target significance level. For the sake of avoiding excessively long execution times of the program **bi2diffac** provided for accomplishing that task,

$$\begin{aligned} SIZE_\alpha^B(\delta_1, \delta_2; m, n) &= \max \left\{ \sup_{0 < p_1 < 1 - \delta_1} Q_{\alpha;m,n}^{\delta_1,\delta_2}(p_1, p_1 + \delta_1) , \right. \\ &\quad \left. \sup_{\delta_2 < p_1 < 1} Q_{\alpha;m,n}^{\delta_1,\delta_2}(p_1, p_1 - \delta_2) \right\} \end{aligned} \quad (6.113)$$

rather than

$$\text{SIZE}_{\alpha}(\delta_1, \delta_2; m, n) = \sup \left\{ Q_{\alpha; m, n}^{\delta_1, \delta_2}(p_1, p_2) \mid 0 < p_1, p_2 < 1, \right. \\ \left. p_1 - p_2 \leq -\delta_1 \text{ or } p_1 - p_2 \geq \delta_2 \right\} \quad (6.114)$$

is used in the iteration process as primary objective function of  $\alpha$ . Obviously, the corrected nominal significance level  $\alpha^*$  obtained in this way has the desired property only if it satisfies  $\text{SIZE}_{\alpha^*}^B(\delta_1, \delta_2; m, n) = \text{SIZE}_{\alpha^*}(\delta_1, \delta_2; m, n)$ . Although up to now no mathematical proof has been made available for the proposition that the trivial relationship  $\text{SIZE}_{\alpha^*}^B(\delta_1, \delta_2; m, n) \leq \text{SIZE}_{\alpha^*}(\delta_1, \delta_2; m, n)$  can always be replaced with the corresponding straight equality under the present circumstances, in extensive numerical investigations not a single counter-example came to light. Nevertheless, `bi2diffac` checks on this condition and provides for putting out an error code rather than a numerical value for  $\alpha^*$  in case of finding that the maximum of the probability of a wrong decision in favor of  $-\delta_1 < p_1 - p_2 < \delta_2$  is attained at an interior point of the subspace  $\{(p_1, p_2) \in (0, 1)^2 \mid p_1 - p_2 \leq -\delta_1 \text{ or } p_1 - p_2 \geq \delta_2\}$ .

For a selection of balanced two-sample designs with binomially distributed data and two specifications of a symmetric equivalence range for  $\delta = p_1 - p_2$ , Table 6.29 shows the result of reducing the nominal significance level as far as necessary for making an exactly valid test of the asymptotic procedure (6.109). The exact size of the corrected test is also shown and compared to that of the uncorrected procedure using the target significance level of 5% as the nominal level. All entries in columns 3 to 5 have been computed running `bi2diffac` with grid span .001 in all steps of searching through the respective parameter subspace for the maximum rejection probability. Qualitatively speaking, the conclusions to be drawn from studying the exact size of the asymptotic test for equivalence with respect to the difference of success rates are quite similar in the two-sample as compared to the McNemar setting: Even with sample sizes located in the extreme upper tail of the distribution of sample sizes available for well-planned clinical trials, the nominal level has to be curtailed substantially in order to obtain a test for equivalence with respect to  $\delta$  which really maintains the target significance level of 5%. We can only speculate about the reasons why convergence of the exact size to the nominal significance level turns out to be so slow in these two specific applications of the asymptotic theory developed in §3.4. It seems as if the mathematically unnatural parametrization of the underlying family of distributions through a raw difference of probabilities would be of greater impact than the binary structure of the individual observations.

Table 6.29 *Nominal significance level and exact size of a corrected version of the testing procedure (6.109) maintaining the 5% level in finite samples of common size  $m = n = 25(25)200$  for equivalence ranges  $(-\delta_1, \delta_2) = (-\delta_o, \delta_o)$  with  $\delta_o = .2, .4$ . [Number in (): size of the critical region of the noncorrected asymptotic testing procedure.]*

$\delta_o$	$n$	$\alpha^*$	$SIZE_{\alpha^*}(\delta_o, \delta_o; n, n)$	
.20	25	.013476	.04046	(.09445)
"	50	.022753	.04801	(.10338)
"	75	.016406	.03029	(.09275)
"	100	.033984	.04823	(.08044)
"	125	.035937	.04999	(.07000)
"	150	.032324	.04332	(.06247)
"	175	.027832	.03732	(.07501)
"	200	.038867	.04758	(.06568)
.40	25	.030468	.03822	(.07356)
"	50	.029296	.03177	(.05868)
"	75	.039746	.04478	(.06170)
"	100	.043164	.04593	(.06190)
"	125	.043164	.04675	(.05939)
"	150	.041601	.04426	(.05676)
"	175	.039257	.04272	(.05514)
"	200	.044824	.04930	(.05309)

Provided one is certain about the reduced nominal level  $\alpha^*$  to be substituted for  $\alpha$  in (6.109), it is of course of considerable practical interest to know with what power the corresponding level-corrected asymptotic test for the equivalence problem (6.106) is able to detect an arbitrarily specified alternative  $(p_1, p_2) \in (0, 1)^2$  such that  $\delta = p_1 - p_2 \in (-\delta_1, \delta_2)$ . The program named `bi2dipow` enables its user to readily find the correct answer to any question of that type. It returns the exact value of the rejection probability of the test (6.109) at any nominal level  $\alpha \in (0, 1)$  under any specified parameter constellation  $(p_1, p_2) \in (0, 1)^2$ . Since each such computation amounts to just a single evaluation of the double sum appearing on the right-hand side of equation (6.111), execution times are very short even if both sample sizes run far beyond the upper limit of the range covered by Table 6.29. Given the sample sizes, the equivalence range and the nominal level, the power of the test not only depends on the true value of the target parameter  $\delta$ , but at the same time on the “baseline” success probability (i.e.,  $p_1$  or  $p_2$ ) as a nuisance parameter. For  $m = n = 50$  and  $\delta_1 = \delta_2 = .20$ , Figure 6.8 represents the changes in power of the level-corrected asymptotic test for equivalence with

respect to  $\delta$  occurring when the true parameter point  $(p_1, p_2)$  moves along the main diagonal of the unit square. Apart from being symmetric about  $1/2$ , the curve is sharply peaked in the neighborhood of both boundaries of the parameter space flattening out to a minimum of less than  $1/3$  of the height of these peaks when approaching the center from either side.

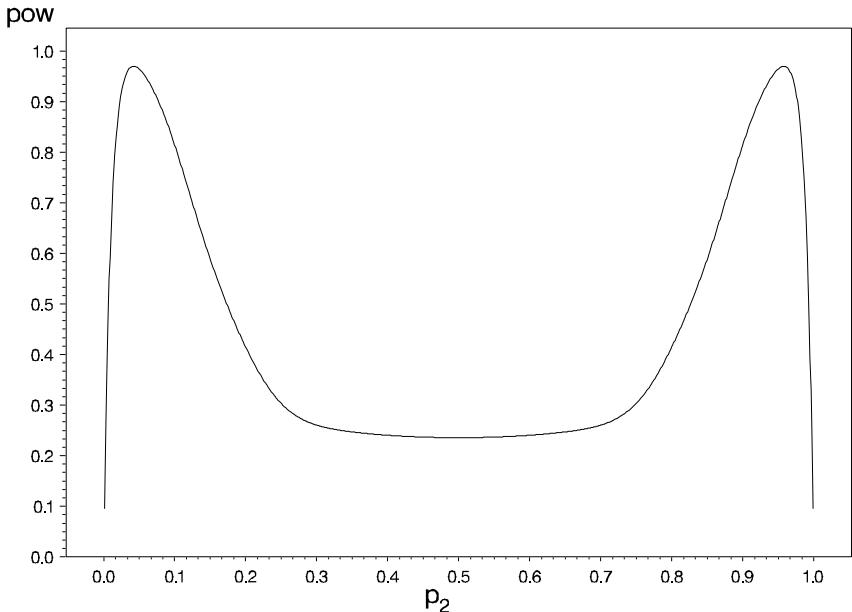


Figure 6.8 *Exact power of the level-corrected version of (6.109) against the alternative  $\delta = 0$  [ $\Leftrightarrow p_1 = p_2$ ] as a function of  $p_2$  as the remaining nuisance parameter, for  $m = n = 50$ ,  $\delta_1 = \delta_2 = \delta_o = .20$  and  $\alpha^* = .0228$ .*

### Discussion

On the one hand, there can be little doubt that it is reasonable to base a test for equivalence of the two binomial distributions  $\mathcal{B}(m, p_1)$  and  $\mathcal{B}(n, p_2)$  with respect to the raw difference  $\delta$  of both success probabilities, on the natural estimator of the distance between the true value of  $\delta$  and the center of the interval specified under the hypothesis to be established. On the other, inspecting the graph shown in the above figure one can hardly help but concede that the power function of the (level-corrected) testing procedure (6.109) looks fairly strange. Actually, it would be desirable to have an option of replacing the asymptotic procedure with an alternative test giving rise to power curves of much more equalized form, provided the price to pay for this advantage had not been a uniform loss in efficiency.

The most promising competitor to the asymptotic procedure (6.109) and

its level-corrected modification we know of is obtained through adopting an idea explained in full detail in the German precursor edition to this book (see Wellek, 1994, § 6.4.2). The key fact one has to exploit in approaching the problem from this alternative point of view is as follows: Given the equivalence limits  $-\delta_1$  and  $\delta_2$  to  $\delta$ , one can find equivalence limits  $\varrho_1^*(\delta_1, \delta_2)$ ,  $\varrho_2^*(\delta_1, \delta_2)$ , say, to the odds ratio such that the associated null hypothesis of nonequivalence with respect to  $\varrho$  is the smallest null hypothesis of the kind treated in § 6.6.4 containing the null hypothesis to be tested now. More precisely speaking, by rather elementary analytical arguments it can be shown that defining

$$\varrho_1^*(\delta_1, \delta_2) = \left( \frac{1 - \delta_1}{1 + \delta_1} \right)^2, \quad \varrho_2^*(\delta_1, \delta_2) = \left( \frac{1 + \delta_2}{1 - \delta_2} \right)^2 \quad (6.115)$$

and

$$H^* = \left\{ (p_1, p_2) \in (0, 1)^2 \mid p_1(1 - p_2)/(1 - p_1)p_2 \leq \varrho_1^*(\delta_1, \delta_2) \vee p_1(1 - p_2)/(1 - p_1)p_2 \geq \varrho_2^*(\delta_1, \delta_2) \right\} \quad (6.116)$$

ensures the validity of the following two statements:

- (i)  $H^* \supseteq H = \left\{ (p_1, p_2) \in (0, 1)^2 \mid p_1 - p_2 \leq -\delta_1 \vee p_1 - p_2 \geq \delta_2 \right\}$ ;
- (ii) the common boundary of  $H$  and  $K$  in the sense of (6.106) contains points which lie also on the common boundary of  $H^*$  and the associated alternative hypothesis  $K^*$ .

Clearly, part (i) of the result implies that the exact Fisher type level- $\alpha$  test of  $H^*$  versus  $K^*$  is a *a fortiori* a test for equivalence of  $\mathcal{B}(m, p_1)$  and  $\mathcal{B}(n, p_2)$  with respect to  $\delta = p_1 - p_2$  which will never exceed the prespecified significance level  $\alpha$ . Since this test is in particular unbiased for  $H^*$  versus  $K^*$ , we can conclude from (ii) that even when reinterpreted as a test for equivalence in the sense of  $-\delta_1 < p_1 - p_2 < \delta_2$ , it exhausts the nominal significance level exhibiting a critical region of a size not falling short of  $\alpha$ .

Figure 6.9 contrasts the power function (again restricted to the diagonal of the unit square) of the UMPU test for  $H^*$  versus  $K^*$  with that of the  $\alpha$ -corrected asymptotic test for  $H$  versus  $K$  in sense of (6.106), for the same specification of the sample sizes and the equivalence range for  $\delta$  which underlies Figure 6.8. At first glance, the conclusion to be drawn from the graph seems undisputable: The UMPU test of the enlarged null hypothesis  $H^*$  looks much too conservative since there are specific alternatives detected by means of the corrected asymptotic test with a power of more than 95% under which

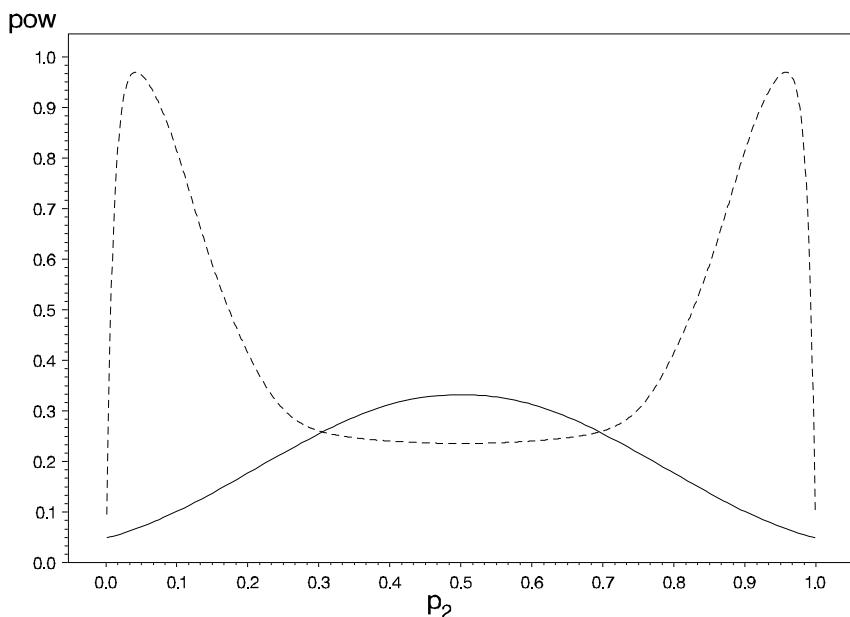


Figure 6.9 *Power of the conservative exact Fisher type test for equivalence with respect to  $\delta$  against alternatives of the form  $p_1 = p_2$ , for  $\alpha = .05$ ,  $m = n = 50$  and  $\delta_1 = \delta_2 = \delta_0 = .20$ . [Dashed line: Power function of the level-corrected asymptotic test for the same problem, reproduced from Figure 6.8.]*

its rejection probability fails to substantially exceed the nominal significance level. However, a bit of further reflection reveals reasons for relativizing this unequivocal preference for the modified asymptotic testing procedure: The alternatives under which the corrected asymptotic test so impressively dominates the conservative UMPU test are restricted to rather extreme constellations being of relevance only for the planning of trials such that the baseline responder rate has to be assumed either exceptionally low or near to 100%. In the majority of real clinical trials, prior information suggests to locate the baseline level at a point of the unit interval being nearer to its center than one of its boundaries, and against alternatives specifying a common response rate in a range of about (.25, .75), the UMPU test of the minimally enlarged null hypothesis in terms of  $\varrho$  is at least as powerful as the corrected asymptotic test. Thus, relying on the experience that extreme alternatives are rarely realistic, the conservative approach to testing for equivalence with respect to  $\delta$  through enclosing the null hypothesis of primary interest by a hypothesis which states inequivalence with respect to the odds ratio, can be regarded as a true alternative to the modified large-sample solution of the testing problem (6.106).

For the sake of completeness, it should be mentioned that the power curve shown in Figure 6.9 for the conservative conditional test, refers to the exact randomized version of the Fisher type test. However, by the means provided in § 6.6.5, the loss in power entailed by dispensing with randomized decisions between the hypotheses can be largely avoided. Hence, the comparative evaluation of the merits of the conservative relative to the direct approach through correcting the large-sample procedure will be little affected by replacing the UMPU test of  $H^*$  versus  $K^*$  with its improved nonrandomized counterpart.

---

## 6.7 Log-rank test for equivalence of two survivor functions

Following Wellek (1993b), we show in this section that the methodology of equivalence and noninferiority testing also extends to situations in which the data are subject to random right-censoring. For this purpose, we adopt the standard framework of statistical survival analysis assuming that the complete information on an arbitrary observational unit is contained in a pair, say  $(T, C)$ , of random variables such that  $T$  denotes the waiting time to some event of interest (usually labeled “death,” irrespective of its concrete meaning), and  $C$  stands for the time until some censoring event occurs. Of these two variables, only one becomes exactly known to the observer whereas the exact value of the time until the other, later occurring event remains latent. In addition to recording the value of the smaller of both time variables, the observer is also enabled to identify the type of event (death versus censorship) which occurred at the respective time. As usual, information on the event type is represented by an indicator variable, say  $\tilde{D}$ , taking on value 1 and 0 if  $T \leq C$  and  $T > C$ , respectively.

For convenience, let us arrange the pooled set of all pairs of time variables underlying a two-sample setting with survival data subject to random right-censoring, into a single sequence  $((T_k, C_k))_{1 \leq k \leq N}$ . Then, the data effectively available from the first sample are given by the  $m$  random pairs  $(X_1, \tilde{D}_1^{(1)}), \dots, (X_m, \tilde{D}_m^{(1)})$  with

$$X_i = \min(T_i, C_i), \quad \tilde{D}_i^{(1)} = \begin{cases} 1 & \text{for } T_i \leq C_i \\ 0 & \text{for } T_i > C_i \end{cases}, \quad 1 \leq i \leq m. \quad (6.117a)$$

Likewise, the censoring process reduces the second sample from  $(T_{m+1}, C_{m+1}), \dots, (T_{m+n}, C_{m+n})$  to  $(Y_1, \tilde{D}_1^{(2)}), \dots, (Y_n, \tilde{D}_n^{(2)})$  where

$$Y_j = \min(T_{m+j}, C_{m+j}), \quad \tilde{D}_j^{(2)} = \begin{cases} 1 & \text{for } T_{m+j} \leq C_{m+j} \\ 0 & \text{for } T_{m+j} > C_{m+j} \end{cases}, \quad 1 \leq j \leq n. \quad (6.117b)$$

Of course, the distributions whose equivalence we want to establish are those of the true survival times  $T_1, \dots, T_m$  ( $\leftrightarrow$  Sample 1) and  $T_{m+1}, \dots, T_{m+n}$  ( $\leftrightarrow$  Sample 2). The survivor functions characterizing these distributions will be denoted by  $S_1(\cdot)$  and  $S_2(\cdot)$  so that we have by definition:

$$S_1(t) = P[T_i \geq t], \quad t \geq 0, \quad i \in \{1, \dots, m\}; \quad (6.118a)$$

$$S_2(t) = P[T_{m+j} \geq t], \quad t \geq 0, \quad j \in \{1, \dots, n\}. \quad (6.118b)$$

Both  $S_1(\cdot)$  and  $S_2(\cdot)$  are assumed to be absolutely continuous and to satisfy an ordinary proportional hazards model (see, e.g., Kalbfleisch and Prentice, 2002, § 2.3.2). With no loss of generality, we treat  $S_2(\cdot)$  as the baseline survivor function underlying the model. With this convention, the proportional hazards assumption is equivalent to the condition [cf. (6.23)]

$$S_1(t) = [S_2(t)]^\theta \text{ for all } t \geq 0 \text{ and some } \theta > 0. \quad (6.119)$$

A nonparametric test for one-sided equivalence of survivor functions has been developed by Freitag (2005) (see also Freitag et al., 2006). Although this test does not have to rely on the proportional hazards assumption, it has other limitations. In particular, it requires restriction to some fixed time window to be specified *a priori*, and the critical bounds to the test statistic have to be determined by means of bootstrap resampling methods.

### 6.7.1 Rationale of the log-rank test for equivalence in the two-sided sense

Generally, it seems intuitively plausible to define equivalent survivor functions by requiring the uniform distance between  $S_1(\cdot)$  and  $S_2(\cdot)$  to be sufficiently small. In other words, we can take equivalence of  $S_1(\cdot)$  and  $S_2(\cdot)$  for granted if we have

$$\|S_1 - S_2\| \equiv \sup_{t>0} |S_1(t) - S_2(t)| < \delta \text{ for some } \delta > 0. \quad (6.120)$$

In view of the continuity of  $S_2(\cdot)$ , (6.119) implies that

$$\|S_1 - S_2\| = \sup_{0 < u < 1} |u - u^\theta|, \quad (6.121)$$

and from (6.121) it follows by means of some elementary calculus that

$$\|S_1 - S_2\| = |\theta^{1/(1-\theta)} - \theta^{\theta/(1-\theta)}|. \quad (6.122)$$

Since, as a function of  $\theta$ , the right-hand side of equation (6.122) is strictly decreasing (increasing) on  $(0, 1)$  (on  $(1, \infty)$ ), and remains invariant under

the reparametrization  $\theta \mapsto \theta^{-1}$ , we can conclude that condition (6.120) is (logically) equivalent to

$$(1 + \varepsilon)^{-1} < \theta < 1 + \varepsilon \text{ for some suitable } \varepsilon > 0. \quad (6.123)$$

Here the constant  $\varepsilon$  has to be determined from  $\delta$  by solving the equation

$$(1 + \varepsilon)^{-1/\varepsilon} - (1 + \varepsilon)^{-(1+\varepsilon)/\varepsilon} = \delta. \quad (6.124)$$

Consequently, under the proportional hazards assumption, equivalent survivor functions in the sense of (6.120) are characterized by the fact that the relative risk belongs to a sufficiently small interval around 1 being symmetric on the log scale. Table 6.30 shows the correspondence between  $\delta$  and  $\varepsilon$  by (6.124) for a selection of numerical examples, together with the right-hand limit of the equivalence range for  $\log \theta$  which in the proportional hazards model corresponds to (6.120) for the specified value of  $\delta$ . Figure 6.10 visualizes a section through the equivalence region (6.120) obtained by fixing the baseline survivor function  $S_2(\cdot)$  and letting  $S_1(\cdot)$  vary over the set  $\{S_2^\theta(\cdot) \mid \theta > 0, \|S_2^\theta - S_2\| < \delta\}$ .

Table 6.30 *Numerical correspondence between equivalence limits referring to  $\|S_1 - S_2\|$  and the parameter  $\theta$  relating both survivor functions under the proportional hazards model.*

$\delta$	.05	.075	.10	.15	.20	.25
$\varepsilon$	.1457	.2266	.3135	.5077	.7341	1.0000
$\log(1 + \varepsilon)$	.1360	.2042	.2727	.4106	.5505	.6931

Since both survivor functions under comparison have been assumed absolutely continuous, each of the true survival times  $T_k$  ( $k = 1, \dots, N$ ) has a well-defined hazard function, say  $\lambda_1(\cdot)$  (for  $k = 1, \dots, m$ ) and  $\lambda_2(\cdot)$  (for  $k = m + 1, \dots, m + n$ ), where

$$\lambda_\nu(t) = \frac{d}{dt} [-\log S_\nu(t)], \quad t \geq 0, \quad \nu \in \{1, 2\}. \quad (6.125)$$

By the basic model assumption (6.119), these functions admit the representation

$$\lambda_\nu(t) = \lambda_2(t)e^{z_k \beta}, \quad t \geq 0, \quad \nu \in \{1, 2\}, \quad k \in \{1, \dots, N\} \quad (6.126)$$

where

$$\beta = \log \theta \text{ and } z_k = \begin{cases} 1 & \text{for } 1 \leq k \leq m \\ 0 & \text{for } m + 1 \leq k \leq N \end{cases}. \quad (6.127)$$

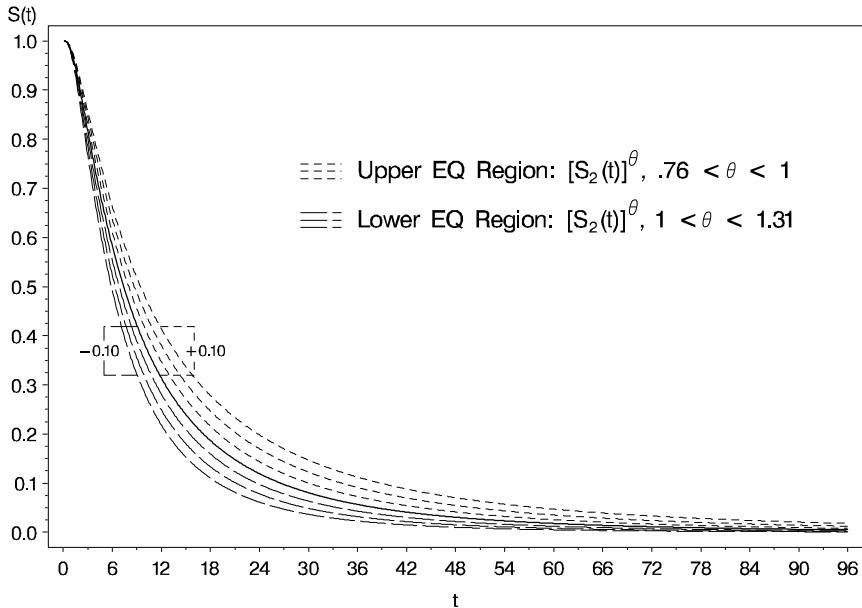


Figure 6.10 *Region of survival curves being equivalent to a given baseline survivor function (solid line) with respect to uniform vertical distance in the proportional hazards model.*

After these preparations, we are ready to construct an asymptotic testing procedure for

$$H : \|S_1 - S_2\| \geq \delta \text{ versus } K : \|S_1 - S_2\| < \delta. \quad (6.128)$$

As explained above, under the proportional hazards assumption (6.119), the problem of testing (6.128) is the same as that of testing

$$\tilde{H} : |\beta| \geq \log(1 + \varepsilon) \text{ versus } \tilde{K} : |\beta| < \log(1 + \varepsilon) \quad (\widetilde{6.128})$$

provided  $\beta$  is the real-valued regression coefficient appearing in (6.126) and  $\varepsilon$  is determined as the unique solution to equation (6.124).

It is easy to verify that in the special case of a proportional hazards model given by (6.126) and (6.127), the partial log-likelihood for  $\beta$  simplifies to

$$\log L(\beta) = \tilde{d}_+^{(1)} \beta - \sum_{q=1}^{\tilde{q}} \log(r_{q1} e^\beta + r_{q2}) \quad (6.129)$$

where

$r_{q\nu}$  = number of items at risk in the  $\nu$ th sample at the  $q$ th smallest failure time  $t_{(q)}$  ( $\nu = 1, 2$ ;  $q = 1, \dots, \tilde{q}$ ),

$$\tilde{d}_{\cdot}^{(1)} = \sum_{i=1}^m \tilde{d}_i^{(1)} = \text{total number of failures in Sample 1}. \quad (6.130)$$

Correspondingly, the maximum partial likelihood estimator  $\hat{\beta}$  is found by solving the equation

$$\sum_{q=1}^{\tilde{q}} \frac{r_{q1} e^{\beta}}{r_{q1} e^{\beta} + r_{q2}} = \tilde{d}_{\cdot}^{(1)}, \quad -\infty < \beta < \infty, \quad (6.131)$$

and the observed information at  $\beta = \hat{\beta}$  is obtained as

$$I_N(\hat{\beta}) = \sum_{q=1}^{\tilde{q}} \frac{r_{q1} r_{q2} e^{\hat{\beta}}}{(r_{q1} e^{\hat{\beta}} + r_{q2})^2}. \quad (6.132)$$

Fairly mild conditions which can be made mathematically precise by specializing the regularity assumptions formulated by Andersen and Gill (1982, §3) (for additional details see Andersen et al., 1993, § VII.2.2) ensure that there exists a finite positive constant,  $\sigma_*^2(\hat{\beta})$  say, such that

$$N^{-1} I_N(\hat{\beta}) \xrightarrow{P} 1/\sigma_*^2(\hat{\beta}) \text{ as } N \rightarrow \infty, \quad (6.133)$$

and

$$\sqrt{N}(\hat{\beta} - \beta)/\sigma_*(\hat{\beta}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) \text{ as } N \rightarrow \infty, \text{ for all } \beta. \quad (6.134)$$

In view of these properties of  $\hat{\beta}$  and  $I_N(\hat{\beta})$ , another application of Theorem A.3.4 shows that the rejection region of an asymptotically valid test for (6.128) and hence for (6.128) itself, is given by

$$\left\{ I_N^{1/2}(\hat{\beta}) |\hat{\beta}| < C_\alpha(I_N^{1/2}(\hat{\beta}) \log(1 + \varepsilon)) \right\}. \quad (6.135)$$

The function  $\psi \mapsto C_\alpha(\psi)$  to be evaluated at  $\psi = I_N^{1/2}(\hat{\beta}) \log(1 + \varepsilon)$  in order to determine the critical upper bound to the absolute value of the standardized estimate of the regression coefficient is the same as defined in the general description given in § 3.4 of the asymptotic approach to constructing tests for equivalence. In other words, for any  $\psi > 0$ ,  $C_\alpha(\psi)$  stands for the square root of the lower  $100\alpha$  percentage point of a  $\chi^2$ -distribution with  $df = 1$  and non-centrality parameter  $\psi^2$ . Thus, given the value of  $\hat{\beta}$  and its estimated standard error, the practical effort entailed in computing the (random) critical bound to be used in the log-rank test for equivalence with rejection region (6.135) reduces to writing a single line of programming code invoking the intrinsic SAS function `cinv` with arguments  $\alpha$ , 1 and  $I(\hat{\beta}) \log^2(1 + \varepsilon)$ , respectively, or

the R function `qchisq` with the same arguments to be plugged in, in the same order. The standard statistical software packages provide the user also with an easy to handle tool for finding the solution to the partial likelihood equation (6.131): It suffices to run a procedure tailored for analyzing standard proportional hazards models with time-independent covariates. In the output list one will find both the value of  $\hat{\beta}$  (e.g., under the label “Parameter Estimate” in the SAS procedure `phreg`) and  $I_N^{-1/2}(\hat{\beta})$  (as an entry into the column “se(coef)” in the standard output of the R function `coxph`). Altogether, these hints should make it sufficiently clear that the practical implementation of the log-rank test for equivalence as given by the rejection region (6.135) reduces to steps very easy to carry out by anybody having access to professional statistics software.

### *Example 6.6*

For the Medulloblastoma II Trial of the SIOP (for details see Bailey et al., 1995),  $N = 357$  children suffering from such malignoma of the cerebellum were enrolled. At some stage of the statistical analysis of the data obtained from this long-term trial, it was of interest to establish equivalence with respect to relapse-free survival between the subgroups of patients treated in big (8 or more cases accrued into the trial) and small centers, respectively. Sample 1 ( $\leftrightarrow$  patients from big centers) was of size  $m = 176$ , exhibiting a total of  $\tilde{d}^{(1)} = 75$  failure events. In the  $n = 181$  patients of Sample 2 ( $\leftrightarrow$  small centers),  $\tilde{d}^{(2)} = 70$  relapses were recorded. The complete relapse-free survival curves obtained for both groups by means of standard product-limit estimation are shown in Figure 6.11. As the maximum distance  $\delta$  considered compatible with equivalence of the underlying true survivor functions [recall (6.120)] let us choose  $\delta = .15$ , i.e., the value in the middle between the two alternative specifications recommended in § 1.7 for defining equivalence ranges for parameters taking values in  $(0, 1)$ . By (6.124) and (6.128), this is the same as specifying the alternative hypothesis  $|\beta| < .4106$  [ $\rightarrow$  Table 6.30] about the regression coefficient introduced in (6.127). With the data underlying Figure 6.11, solving equation (6.131) gives the ML estimate  $\hat{\beta} = .0944$ . Using this value in (6.132) yields  $I_N(\hat{\beta}) = 36.1981$  as the information observed at  $\hat{\beta}$  so that we get  $I_N^{1/2}(\hat{\beta})|\hat{\beta}| = .5680$ . On the other hand, with  $\hat{\psi}_N = \sqrt{36.1981} \cdot .4106 = 2.4704$  and  $\alpha = .05$ , for the critical bound which the test statistic  $I_N^{1/2}(\hat{\beta})|\hat{\beta}|$  has to be compared with according to (6.135), we obtain  $C_\alpha(\hat{\psi}_N) = \sqrt{.689235} = .8302$  where, in SAS, .689235 is the value assigned to the variable `csq`, say, as the result of getting executed the statement `csq=cinv(.05,1,2.4704**2)` (in R, the analogous statement reads `csq=qchisq(.05,1,2.4704^2)` yielding the value .6892414). Since the observed value of  $I_N^{1/2}(\hat{\beta})|\hat{\beta}|$  thus falls short of its critical upper bound, the test decides in favor of equivalence of the survivor functions under comparison.

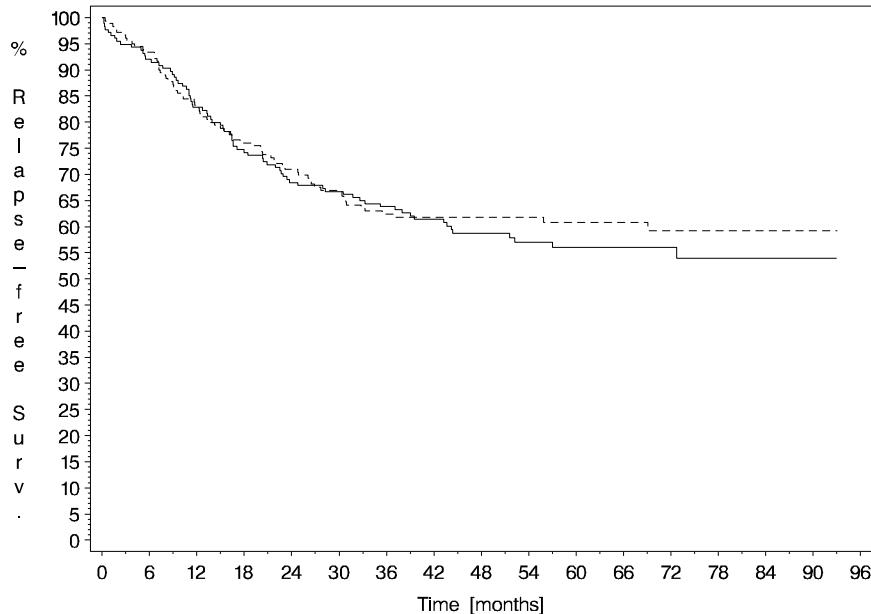


Figure 6.11 *Medulloblastoma Trial SIOP II: Observed relapse-free survival curves by size of treatment center. Solid line: Big centers having enrolled  $\geq 8$  patients. Dashed line: Small centers with total accrual < 8.* (From Wellek, 1993b, with kind permission by the International Biometric Society.)

#### *Simulation results on the size of the log-rank test for equivalence*

In order to study the size of the critical region (6.135), two sets of simulation experiments were performed. In the first set, the baseline survivor function was lognormal with parameters  $\mu = 2$  and  $\sigma^2 = 1$  as given by  $S_2(t) = \Phi(2 - \log t)$ , whereas the censoring distribution function was specified  $G(t) = 1 - e^{-t/50}$  irrespective of the true value of the hazard ratio  $\theta$ . In the second set of experiments, the lognormal distribution  $\mathcal{LN}(2, 1)$  was replaced by the Weibull distribution with form parameter  $c = 2$  and unit scale parameter. The associated censoring times were generated from two different rectangular distributions inducing baseline ( $\leftrightarrow$  Group 2) censoring rates of 20 and 80%, respectively. The results of studying the size of the critical region (6.135) under both types of models are summarized in Tables 6.31 a,b. Each entry is based on 40,000 replications of the respective simulation experiments and gives the fraction of samples for which the test decides in favor of equivalence. None of the entries into the rightmost columns of the tables indicates relevant anticonservatism so that the level properties of the test are fully satisfactory even when performed with very small samples.

Table 6.31a *Estimated rejection probabilities of the log-rank test for equivalence at nominal level  $\alpha = .05$  under the null hypothesis  $\|S_1 - S_2\| \geq .15$  ( $\leftrightarrow \varepsilon = .5077$ ) in a model with lognormal baseline survivor function and exponentially distributed censoring times.*<sup>†</sup>

$m$	$n$	$\theta = 1/(1 + \varepsilon)$	$\theta = 1 + \varepsilon$
25	25	.04945	.05103
50	25	.05078	.05158
50	50	.05108	.05003
75	50	.04778	.05238
75	75	.05085	.04913
100	75	.04800	.04963
100	100	.04943	.05025

<sup>†</sup> censoring rates: .188 for Group 2; .268 and .133 for Group 1 under  $\theta = 1/1.5077$  and  $\theta = 1.5077$ , respectively

Table 6.31b *Analogue of Table 6.31a for a Weibull baseline survivor function and uniform censoring distributions  $\mathcal{U}(0, \tau)$  inducing various group-specific censoring rates  $\zeta_1$  and  $\zeta_2$ .*

$m$	$n$	$\tau$	$\theta = 1/(1 + \varepsilon)$	$\theta = 1 + \varepsilon$
25	25	4.431133*	.04913	.05068
50	25	"	.04703	.05125
50	50	"	.05000	.04843
75	50	"	.04548	.05063
75	75	"	.04940	.05030
100	75	"	.04978	.05110
100	100	"	.04848	.05165
25	25	1.748709**	.04838	.04820
50	25	"	.04875	.05375
50	50	"	.04995	.04850
75	50	"	.04918	.04995
75	75	"	.05050	.05045
100	75	"	.04723	.05028
100	100	"	.05090	.04705

\*  $(\zeta_1, \zeta_2) = (.2450, .2000)$  and  $(.1633, .2000)$  under  $\theta = 1/(1 + \varepsilon)$  and  $\theta = 1 + \varepsilon$ , respectively

\*\*  $(\zeta_1, \zeta_2) = (.5938, .5000)$  and  $(.4127, .5000)$  under  $\theta = 1/(1 + \varepsilon)$  and  $\theta = 1 + \varepsilon$ , respectively

### 6.7.2 Power approximation and sample size calculation for the log-rank test for equivalence

As usual, for purposes of approximating the power and computing sample sizes required for attaining some prespecified power, we need an expression for the “theoretical” variance of the nonstandardized test statistic under any alternative of potential interest. In view of practical applications to the planning of equivalence trials with survival endpoints, it seems sufficient to restrict the class of alternatives of interest by the following assumptions:

- (I) Both groups are subject to the same independent censoring mechanism induced by a recruitment process running uniformly over some fixed accrual period  $[0, \tau]$ .
- (II) The study protocol specifies a minimum follow-up period of length  $t^\circ \geq 0$  for each patient. Since  $t^\circ$  might also be equal to zero, studies allowing the admission of new patients during the whole duration of the trial are covered as a special case.
- (III) The trial under consideration is to be a balanced one, in the sense that we have  $m = n, N = 2n$ .

Applying Theorem VII.2.2 of Andersen et al. (1993) it can be shown that under assumptions (I)–(III), the asymptotic variance of the normalized partial likelihood estimator  $\sqrt{N}\hat{\beta}$  is given by

$$\sigma_*^2(\hat{\beta}) = \frac{1}{v^2(\beta; \tau, t^\circ)} \quad (6.136)$$

where

$$v^2(\beta; \tau, t^\circ) = \frac{e^\beta}{2} \left\{ \int_0^{t^\circ} \left[ e^\beta + (S_2(t))^{1-e^\beta} \right]^{-1} f_2(t) dt + \int_{t^\circ}^{t^\circ + \tau} \left[ e^\beta + (S_2(t))^{1-e^\beta} \right]^{-1} \frac{t^\circ + \tau - t}{\tau} f_2(t) dt \right\} \quad (6.137)$$

and  $f_2(\cdot)$  denotes the density of the baseline survival distribution as assumed for Group 2.

In general, the integral appearing on the right-hand side of (6.137) admits no representation in closed form. However, it is easy to evaluate numerically (e.g., using *Mathematica*<sup>®</sup> or *Maple*<sup>®</sup>). In the specific case  $\beta = 0$  [ $\leftrightarrow$  alternative of exact coincidence of both survival distributions], (6.137) reduces to

$$v^2(0; \tau, t^\circ) = \frac{1}{4} P[T < C] \quad (6.138)$$

where  $T$  denotes a random variable with density  $f_2(\cdot)$ , and  $C$  a censoring variable independent of  $T$  and following the uniform distribution over the interval  $[t^\circ, t^\circ + \tau]$ .

Now, (6.133) and (6.134) suggest that for purposes of power approximation, we replace  $I_N^{1/2}(\hat{\beta})$  with  $\sqrt{N}/\sigma_*(\hat{\beta})$  everywhere in (6.135) and treat  $\sqrt{N}\hat{\beta}/\sigma_*(\hat{\beta})$  as normally distributed with unit variance and mean  $\sqrt{N}\beta_o/\sigma_*(\hat{\beta})$  where  $\beta_o \in \mathbb{R}$  denotes the value of the log-hazard ratio  $\beta$  [recall (6.126), (6.127)] specified as the alternative of interest. Denoting the power of the test given by (6.135) against  $\beta = \beta_o$  by  $\Pi_o$ , this leads to the relationship

$$\begin{aligned} \Pi_o \approx & \Phi\left(C_\alpha(\sqrt{N}\log(1+\varepsilon)/\sigma_*(\hat{\beta})) - \sqrt{N}\beta_o/\sigma_*(\hat{\beta})\right) - \\ & \Phi\left(-C_\alpha(\sqrt{N}\log(1+\varepsilon)/\sigma_*(\hat{\beta})) - \sqrt{N}\beta_o/\sigma_*(\hat{\beta})\right). \end{aligned} \quad (6.139)$$

A natural way of establishing a method for estimating minimally required sample sizes is now to treat (6.139) as an equation in  $N$  with all other quantities as given numbers. The task of formulating an algorithm for finding the solution to this equation can be considerably simplified by making use of another approximation, namely  $C_\alpha(\sqrt{N}\log(1+\varepsilon)/\sigma_*(\hat{\beta})) \approx \sqrt{N}\log(1+\varepsilon)/\sigma_*(\hat{\beta}) - u_{1-\alpha}$  [ $\rightarrow$  p. 65, (4.26) with  $\sqrt{N}\log(1+\varepsilon)/\sigma_*(\hat{\beta})$  being substituted for  $\sqrt{n}\varepsilon$ ]. Combining the latter with (6.139) and dropping the distinction between approximate and exact identities yields, after a few rearrangements of terms, the equation

$$\begin{aligned} \Phi\left(\frac{\sqrt{N}(\log(1+\varepsilon) - \beta_o)}{\sigma_*(\hat{\beta})} - u_{1-\alpha}\right) \\ + \Phi\left(\frac{\sqrt{N}(\log(1+\varepsilon) + \beta_o)}{\sigma_*(\hat{\beta})} - u_{1-\alpha}\right) = 1 + \Pi_o. \end{aligned} \quad (6.140)$$

In order to determine  $N$  from (6.140), it is convenient to proceed as follows: We first treat  $\sigma_*(\hat{\beta})/\sqrt{N}$  as an unknown positive number, say  $\tilde{\sigma}$ , and solve the equation

$$\Phi\left(\frac{\log(1+\varepsilon) - \beta_o}{\tilde{\sigma}} - u_{1-\alpha}\right) + \Phi\left(\frac{\log(1+\varepsilon) + \beta_o}{\tilde{\sigma}} - u_{1-\alpha}\right) = 1 + \Pi_o \quad (\widetilde{6.140})$$

for  $\tilde{\sigma}$  which can be done by elementary numerical techniques since the expression on the left-hand side is obviously decreasing in  $\tilde{\sigma}$ . Denoting the solution to (6.140) by  $\tilde{\sigma}(\beta_o; \alpha, \varepsilon, \Pi_o)$ , we have to choose  $N$  such that it satisfies  $\sigma_*(\hat{\beta})/\sqrt{N} = \tilde{\sigma}(\beta_o; \alpha, \varepsilon, \Pi_o)$ . In view of (6.136), this is equivalent to determining the approximate total sample size required for attaining power  $\Pi_o$  against the specific alternative  $\beta_o \in (-\log(1+\varepsilon), \log(1+\varepsilon))$  by means of

$$N = \left[ \tilde{\sigma}^2(\beta_o; \alpha, \varepsilon, \Pi_o) v^2(\beta_o; \tau, t^\circ) \right]^{-1}. \quad (6.141)$$

Through  $v^2(\beta_o; \tau, t^\circ)$  [recall (6.137)], the resulting minimum sample size depends both on the selected baseline survival distribution and the censoring

mechanism, as has necessarily to be the case.

In the special case of the “null alternative”  $\beta_0 = 0$  under which both survivor functions actually coincide, (6.140) reduces to  $\Phi\left(\frac{\log(1+\varepsilon)}{\tilde{\sigma}} - u_{1-\alpha}\right) = (1 + \Pi_0)/2$  so that  $\tilde{\sigma}(0; \alpha, \varepsilon, \Pi_0)$  is explicitly given by  $\frac{1}{\tilde{\sigma}(0; \alpha, \varepsilon, \Pi_0)} = \frac{\Phi^{-1}((1 + \Pi_0)/2) + u_{1-\alpha}}{\log(1 + \varepsilon)}$ . Plugging this expression into (6.141) gives

$$N = \frac{[\Phi^{-1}((1 + \Pi_0)/2) + u_{1-\alpha}]^2}{v^2(0; \tau, t^\circ) \log^2(1 + \varepsilon)} \quad (6.141^\circ)$$

as the null alternative version of (6.141). In view of (6.138), one can state that (6.141 $^\circ$ ) exhibits a strict formal analogy with the well-known sample size formula derived by Schoenfeld (1981) for the log-rank test for conventional one-sided hypotheses. In fact, according to Schoenfeld’s result, (6.141 $^\circ$ ) approximates the total sample size required to detect the alternative that the true hazard ratio equals  $1 + \varepsilon$ , with power  $1 - \alpha$  in a test at nominal significance level  $1 - (1 + \Pi_0)/2$ .

#### *Simulation results on power and sample size approximation*

In the majority of real applications of the log-rank test for equivalence, sample size calculation will be done under the assumption that the trial aims at detecting perfect coincidence of the survival distributions under comparison. Accordingly, the bulk of the simulation results presented below as a basis for assessing the accuracy of the sample size approximation (6.141) refers to the special case (6.141 $^\circ$ ). Table 6.32 contrasts the total sample size computed by means of our approximate formula with the simulated exact  $N$  required to attain prespecified power  $\Pi_0$  against  $\beta = \beta_0 = 0$  for various values of  $\Pi_0$  and common censoring rates  $\zeta$ . In the Monte Carlo experiments performed to generate the entries in the last two columns of this table, we again assumed both true survival distributions to be Weibull with shape parameter  $c = 2.0$  and unit scale parameter. Furthermore, the associated values of the censoring variables were generated to satisfy conditions (I)–(III) of p. 210 with  $t^\circ = 0$  [ $\leftrightarrow$  no minimum follow-up period, i.e., admission of new patients over the whole duration of the trial] and length  $\tau$  of accrual period being adjusted to match with the prespecified event rate  $1 - \zeta$  common to both arms of the trial. The equivalence range for the log-hazard rate  $\beta$  was chosen as in Tables 6.31 a,b so that under the hypothesis to be established, a vertical distance of  $\delta = .15$  at most was considered compatible with equivalence between two survivor functions. Each individual simulation experiment was run 10,000 times.

Table 6.32 *Simulated exact total sample sizes versus sample sizes approximated by means of formula (6.141°), for nominal power  $\Pi_o = .5, .8, .9$  and common censoring rate  $\zeta = .1, .2, .5, .8$ . [Survival distributions: Weibull with form parameter  $c = 2.0$ ; censoring distributions: uniform.]*

$\zeta$	$\Pi_o$	Approx. $N$	Simulated Exact Power	Simulated Exact $N$
.10	.50	140	.4779	144
"	.80	226	.7826	230
"	.90	286	.9008	= Approx. $N$
.20	.50	160	.4834	164
"	.80	254	.7871	258
"	.90	320	.8902	328
.50	.50	256	.5022	= Approx. $N$
"	.80	406	.7931	408
"	.90	514	.8966	= Approx. $N$
.80	.50	638	.4994	= Approx. $N$
"	.80	1016	.7923	1020

Qualitatively speaking, the accuracy of the suggested approximation to the total sample size required to detect null alternatives in the log-rank test with prespecified power, turns out surprisingly good. In fact, the discrepancy between the approximated and the simulated exact value, if existent at all, is found not to exceed 4 per group which seems negligible for all practical purposes.

The subsequent table suggests that this conclusion is warranted all the more if the alternative  $\beta_o$  of interest is chosen as some point between the center and the boundaries of the equivalence range specified for the target parameter under the alternative hypothesis. The simulation study performed to generate the entries in the rightmost column of Table 6.33 was designed as follows: Both survival distributions were assumed to be exponential with hazard rate  $\lambda_2 = .80$  ( $\leftrightarrow$  baseline) and  $\lambda_1 = \lambda_2(1 + \psi\varepsilon)$  ( $\leftrightarrow$  Group 1), respectively, with  $\varepsilon$  specified as before (i.e.,  $\varepsilon = .5077$ ) and  $0 < \psi < 1$ . The censoring distribution was assumed uniform over  $[0, \tau] = [0, 1.992030]$  for both arms of the trial ensuring that the censoring proportion had exact value  $\zeta_2 = .50$  under the baseline hazard rate. The entries in column 3 have been obtained by evaluating the second integral (the first one vanishes for  $t^o = 0$ ) on the right-hand side of (6.137) with  $S_2(t) = e^{-\lambda_2 t}$ ,  $f_2(t) = \lambda_2 e^{-\lambda_2 t}$ ,  $t > 0$ . Column 5 shows the values of  $\tilde{\sigma}$  which solve equation (6.140) for the respective combination of  $\beta_o$  and  $\Pi_o$  when  $(\alpha, \varepsilon)$  is fixed at (.05, .5077).

The main conclusion to be drawn from the results shown in Table 6.33 is that the exact power attained with samples of total size estimated by means of formula (6.141) approximates the nominal power even slightly better for nonnull alternatives as compared to settings where perfect coincidence of both survivor functions is the specific alternative of interest. Presumably, this just reflects the obvious fact that given the values of all other parameters involved, the total sample size required to detect the specific alternative  $\beta = \beta_0$  increases with the distance of  $\beta_0$  from the center of the equivalence range for  $\beta$ .

Table 6.33 *Exact versus nominal power against nonnull alternatives of the log-rank test for equivalence with total sample size estimated from (6.141). [Baseline survival distribution: exponential with hazard rate .80; censoring distribution: uniform, inducing event rates of  $1 - \zeta_2 = .50$  under baseline hazard.]*

$\psi$	$\beta_0$	$v^2(\beta_0; \tau, t^0)$	$\Pi_0$	$\tilde{\sigma}(\beta_0; \alpha, \varepsilon, \Pi_0)$	Simul.		
					Appr.	$N$	Ex. Pow.
.1	.049526	.126806	.50	.17497	258	.4995	
"	"	"	.80	.13640	424	.7951	
"	"	"	.90	.11988	548	.8949	
.25	.119499	.129242	.50	.16332	290	.5038	
"	"	"	.80	.11676	568	.8010	
"	"	"	.90	.09945	782	.8975	

### 6.7.3 Log-rank test for noninferiority

In the one-sided case, it is only of interest to rule out that the survivor function  $S_1(\cdot)$  associated with the experimental treatment, falls below the reference survivor function  $S_2(\cdot)$  by more than some given threshold  $\delta > 0$  at some point on the time axis. In other words, the experimental treatment is considered noninferior to the reference if there holds  $S_1(t) > S_2(t) - \delta$  for all  $t > 0$ . In the proportional hazards model (6.119), the corresponding testing problem

$$H_1 : S_1(t) \leq S_2(t) - \delta \quad \text{for at least one } t > 0 \\ \text{versus} \quad K_1 : S_1(t) > S_2(t) - \delta \quad \text{for all } t > 0 \quad (6.142)$$

can easily be shown to be equivalent to

$$\bar{H}_1 : \beta \geq \log(1 + \varepsilon) \quad \text{versus} \quad \bar{K}_1 : \beta < \log(1 + \varepsilon). \quad (6.143)$$

The sign reversal entailed by proceeding from  $(H_1, K_1)$  to  $(\tilde{H}_1, \tilde{K}_1)$  directly reflects the obvious fact that large hazards rates are undesirable.

In view of (6.133) and (6.134), it follows from the general result stated in § 2.3 that the rejection region of an asymptotically valid test for (6.143) consists of all points in the sample space of the primary observations  $(T_1^{(1)}, D_1^{(1)}), \dots, (T_m^{(1)}, D_m^{(1)}), (T_1^{(2)}, D_1^{(2)})$  for which it turns out that

$$\left\{ \left( \hat{\beta} - \log(1 + \varepsilon) \right) I_N^{1/2}(\hat{\beta}) < u_\alpha \right\}. \quad (6.144)$$

The practical implementation of the corresponding decision rule is still simpler than that of the log-rank test for two-sided equivalence and needs no additional explanation.

Under the assumptions made precise in the previous subsection, the power  $\Pi_o$  of the asymptotic test with rejection region (6.144) against any alternative  $\beta_o$  can be approximated through setting the second term on the right-hand side of Equation (6.140) equal to unity yielding

$$\Pi_o = \Phi \left( \frac{\sqrt{N}(\log(1 + \varepsilon) - \beta_o)}{\sigma_*(\hat{\beta})} - u_{1-\alpha} \right). \quad (6.145)$$

Recalling (6.136) and solving this equation for  $N$  yields after some trivial rearrangements of terms:

$$N = \frac{\left[ \Phi^{-1}(\Pi_o) + u_{1-\alpha} \right]^2}{[\log(1 + \varepsilon) - \beta_o]^2 v^2(\beta_o; \tau, t^*)}. \quad (6.146)$$

In the special case that interest is in controlling approximately the power against the null alternative specifying  $\beta_o = 0$ , (6.146) simplifies to

$$N = \frac{\left[ \Phi^{-1}(\Pi_o) + u_{1-\alpha} \right]^2}{\log^2(1 + \varepsilon) v^2(0; \tau, t^*)}. \quad (6.146^\circ)$$

According to (6.141°), this coincides with the approximate total sample size required for achieving power  $2\Pi_o - 1$  against  $\beta = 0$  in the log-rank test for two-sided equivalence.

### *Example 6.7*

For illustration, we present the results of sample-size calculation for a clinical trial to be run for the purpose of establishing one-sided equivalence of interstitial brachytherapy and radical surgery as treatments of localized carcinoma of the prostate with time to tumor progression as outcome criterion of primary interest. The study was projected to extend over a total of 6.5 years, including an accrual period of  $\tau = 5$  years' duration. Accordingly, minimum

follow-up time was limited to  $t^{\circ} = 1.5$  [yrs]. The probability of remaining progression-free for at least 5 yrs was assumed to be  $P_2 = .85$  after radical surgery, and the largest tolerable difference between this baseline survival rate and the corresponding value  $P_1 = S_1(\tau)$  for the brachytherapy arm defined to be  $\delta_{\tau} = .15$ . These specifications imply that the equivalence margin for the regression coefficient  $\beta$  of the dummy covariate [treatment-arm indicator]  $Z$  is given by

$$.85^{(1+\varepsilon)} = .70 \Leftrightarrow 1 + \varepsilon = 2.194667 \Leftrightarrow \log(1 + \varepsilon) = .786030. \quad (6.147)$$

Sample size calculation was based on the assumption that both survival distributions are exponential. The power to detect the null alternative  $\beta_0 = 0$  in a log-rank test for noninferiority at asymptotic level  $\alpha = .05$  was set at  $\Pi_0 = .80$ .

Starting from these assumptions and specifications, calculation of the asymptotic variance of the normalized partial likelihood estimator  $\sqrt{N} \hat{\beta}$  by means of (6.138) is a fairly easy task. Under the selected alternative, both survival distributions are exponential with hazard rate  $-\log(P_2)/\tau = -\log(.85)/5 = .0325$ . The probability that a random variable  $T$  with distribution  $\mathcal{E}(1/.0325)$  does not exceed an independent  $C$  uniformly distributed over  $[0, \tau] = [0, 5]$  is obtained by means of elementary calculations to be .120952 so that we have  $v^2(0; 5, 1.5) = (1/4)P[T < C] = .030238$ . Plugging in this value together with (6.147) and  $\alpha = .05$ ,  $\Pi_0 = .80$  in our sample size formula (6.146°) for the noninferiority case with  $\beta = 0$  as the alternative of interest, yields

$$N = \frac{[\Phi^{-1}(.80) + u_{.95}]^2}{.030238 \cdot .786030^2} = \frac{[.841621 + 1.644854]^2}{.018682} = 330.93 \approx 330.$$

Since the assumption that the experimental therapy leads to exactly the same tumor progression hazard as the reference treatment was deemed a bit too optimistic, the following nonnull alternative was alternatively taken into consideration:

$$\begin{aligned} P_1 = S_1(\tau) &= .85, \quad P_2 = S_2(\tau) = .825 \\ &\Rightarrow .85^{\exp\{\beta_0\}} = .825 \Leftrightarrow \beta_0 = .168636. \end{aligned}$$

Leaving all other specifications invariant, we obtain  $v^2(\beta; \tau, t^*) = .027529$  implying that the total sample size required for ensuring asymptotic power 80% has to be recomputed as

$$N = \frac{[\Phi^{-1}(.80) + u_{.95}]^2}{[.786030 - .168636]^2 \cdot .027529} = 589.19 \approx 590.$$

Comparing both results reflects the fact that the power of the log-rank test for (one-sided) equivalence is highly sensitive against misspecifications of the alternative.

*Simulation results on the log-rank test for noninferiority*

The simulation studies performed to investigate the small-sample behavior under specific alternatives of the noninferiority version of the log-rank were designed in exactly the same way as those which gave the results shown in Tables 6.32 and 6.33 for the case of testing for two-sided equivalence of survivor functions. Table 6.34 provides a basis for assessing the accuracy attainable by estimating total sample size through applying formula (6.146°). The models for generating the data by Monte Carlo simulation were exactly the same as in the two-sided case, and regarding the differences between approximated and simulated exact values of  $N$ , the same qualitative statements can be made. Again, the accuracy of the suggested approximation to the sample size required when the focus is on null alternatives, turns out very satisfactory.

Table 6.34 *Results of repeating the simulation experiments leading to Table 6.32, for the noninferiority version of the log-rank test.*

$\zeta$	$\Pi_0$	Approx. $N$	Simulated Exact Power	Simulated Exact $N$
.10	.50	72	.4920	74
"	.80	162	.7886	166
"	.90	226	.8952	= Approx. $N$
.20	.50	80	.4850	84
"	.80	184	.7950	= Approx. $N$
"	.90	254	.8906	256
.50	.50	128	.4966	= Approx. $N$
"	.80	294	.8022	= Approx. $N$
"	.90	406	.8995	= Approx. $N$
.80	.50	320	.4938	322
"	.80	734	.8035	= Approx. $N$

Finally, all simulation experiments whose results are given in Table 6.33 were likewise repeated with the log-rank test for noninferiority. Comparing the entries in columns 4 and 6 of Table 6.35 within the same row leads to the conclusion that the accuracy of the large sample approximation (6.145) to the power of the one-sided version of the log-rank test for equivalence keeps being remarkably good even when the alternative of interest is chosen from a neighborhood of the null model.

Table 6.35 *Exact versus nominal power against nonnull alternatives of the log-rank test for noninferiority with total sample size estimated from (6.146). [For details on notation and the models used to generate the data see the description of Table 6.33.]*

$\psi$	$\beta_o$	$v^2(\beta_o; \tau, t^o)$	$\Pi_o$	Simul.		
				Appr.	$N$	Ex. Pow.
.1	.049526	.126806	.50	164	.4993	
"	"	"	.80	374	.7952	
"	"	"	.90	518	.8969	
.25	.119499	.129242	.50	248	.4993	
"	"	"	.80	564	.7920	
"	"	"	.90	782	.8929	

---

## Multisample tests for equivalence

---

### 7.1 The intersection-union principle as a general solution to multisample equivalence problems

The existing literature on  $k$ -sample equivalence testing procedures is comparatively sparse. A reason for this fact might be that, at least for parametric settings, there is a straightforward yet often overconservative solution to problems of establishing equivalence of  $k$  distributions as soon as a test for the corresponding two-sample problem is available. The rationale behind this approach going back to Berger (1982) is a result which can appropriately be termed intersection-union principle because of its straight duality with Roy's (1953) well-known union-intersection approach to multiple testing and interval estimation. The scope of potential applications of the principle is fairly wide and goes far beyond the specific multisample settings discussed in the subsequent sections of this chapter.

In order to introduce the general idea, let us assume that we are considering an arbitrary but fixed number  $q \geq 2$ , say, of elementary null hypotheses  $H_1, \dots, H_q$  to be tested against associated alternatives  $K_1, \dots, K_q$ . For each  $\nu = 1, \dots, q$ , let  $H_\nu$  and  $K_\nu$  denote statements about some parametric function (in nonparametric settings: a functional of the distributions from which the data are taken)  $\eta_\nu$  such that there is some testing procedure  $\phi_\nu$ , say, valid at the selected nominal significance level  $\alpha \in (0, 1)$ . Finally, suppose that the rejection probability of test  $\phi_\nu$  does not exceed  $\alpha$ , under any possible parameter constellation  $(\eta_1, \dots, \eta_q)$  such that  $\eta_\nu \in H_\nu$ , *irrespective of how many and what other elementary null hypotheses are true*. Then, the intersection-union principle states that the following decision rule defines a valid level- $\alpha$  test for the “combined” problem

$$H \equiv H_1 \cup H_2 \cup \dots \cup H_q \quad \text{versus} \quad K \equiv K_1 \cap K_2 \dots \cap K_q : \quad (7.1)$$

Reject  $H$  in favor of  $K$  if and only if

each individual test  $\phi_\nu$  rejects the null hypothesis  $H_\nu$

which it has been constructed for ( $\nu = 1, \dots, q$ ). (7.2)

The *proof* of the result is almost trivial, at least if one is willing to adopt some piece of the basic formalism customary in expositions of the abstract

theory of statistical hypotheses testing methods. In fact, if we denote the combined test defined by (7.2)  $\phi$  and identify each test involved with its critical function, it is clear that  $\phi$  admits the representation  $\phi = \phi_1 \cdot \phi_2 \cdot \dots \cdot \phi_q$ . The condition we have to verify, reads then as follows:

$$E_{(\eta_1, \dots, \eta_q)}(\phi) \leq \alpha \quad \text{for all } (\eta_1, \dots, \eta_q) \in H \quad (7.3)$$

where  $E_{(\eta_1, \dots, \eta_q)}(\cdot)$  denotes the expected value computed under the parameter constellation  $(\eta_1, \dots, \eta_q)$ . Now, if  $(\eta_1, \dots, \eta_q)$  belongs to  $H$ , this means by definition that there is some  $\nu \in \{1, \dots, q\}$  with  $\eta_\nu \in H_\nu$ . Furthermore, every critical function  $\phi_1, \dots, \phi_q$  takes on values in  $[0, 1]$  only, which implies that  $\phi \leq \phi_\nu$  and thus  $E_{(\eta_1, \dots, \eta_q)}(\phi) \leq E_{(\eta_1, \dots, \eta_q)}(\phi_\nu)$ . By assumption, we have  $E_{(\eta_1, \dots, \eta_q)}(\phi_\nu) \leq \alpha$  for every  $(\eta_1, \dots, \eta_q)$  with  $\eta_\nu \in H_\nu$ , and hence the proof of (7.3) is already complete.

It is interesting to note that the principle of confidence interval inclusion introduced in § 3.1 as a general approach to the construction of tests for single equivalence hypotheses, can be viewed as a special case of the intersection-union principle. This follows simply from the fact pointed out on p.34 that every interval inclusion test admits a representation as a double one-sided testing procedure.

In order to apply the result to multisample equivalence testing problems, let  $\theta_j$  be the parameter of interest (e.g., the expected value) for the  $i$ th distribution under comparison, and require of a pair  $(i, j)$  of distributions equivalent to each other that the statement

$$K_{(i,j)} : \rho(\theta_i, \theta_j) < \varepsilon, \quad (7.4)$$

holds true with  $\rho(\cdot, \cdot)$  denoting a suitable measure of distance between two parameters. Suppose furthermore that for each  $(i, j)$  a test  $\phi_{(i,j)}$  of  $H_{(i,j)} : \rho(\theta_i, \theta_j) \geq \varepsilon$  versus  $K_{(i,j)} : \rho(\theta_i, \theta_j) < \varepsilon$  is available whose rejection probability is  $\leq \alpha$  at any point  $(\theta_1, \dots, \theta_k)$  in the full parameter space such that  $\rho(\theta_i, \theta_j) \geq \varepsilon$ . Then, by the intersection-union principle, deciding in favor of “global equivalence” if and only if equivalence can be established for all  $\binom{k}{2}$  possible pairs, yields a valid level- $\alpha$  test for

$$H : \max_{i < j} \{\rho(\theta_i, \theta_j)\} \geq \varepsilon \quad \text{versus} \quad K : \max_{i < j} \{\rho(\theta_i, \theta_j)\} < \varepsilon. \quad (7.5)$$

As a specific instance, we get a quick solution to the problem of testing  $k$  Gaussian distributions with common unknown variance  $\sigma^2$  for equivalence with respect to their unscaled means, say  $\mu_1, \dots, \mu_k$ , if we proceed in the following way: Apply for each  $(i, j)$  the interval inclusion rule with standard  $t$ -based  $(1-2\alpha)$ -confidence intervals for  $\mu_i - \mu_j$ . This yields a test for equivalence of  $\mathcal{N}(\mu_i, \sigma^2)$  and  $\mathcal{N}(\mu_j, \sigma^2)$  in the sense of  $|\mu_i - \mu_j| < \varepsilon$  whose validity is not affected by changes of any other functions of the whole vector  $(\mu_1, \dots, \mu_k)$  of means except for the difference within the given pair. Thus, deciding in favor of “global” equivalence of all  $k$  distributions if and only if all  $\binom{k}{2}$  confidence

intervals are included in  $(-\varepsilon, \varepsilon)$  yields a valid level- $\alpha$  test of  $H : |\mu_i - \mu_j| \geq \varepsilon$  for at least one  $(i, j)$  versus  $K : |\mu_i - \mu_j| < \varepsilon$  for all  $(i, j) \in \{1, \dots, k\}^2$ .

More often than not, the obvious advantages of this approach (conceptual simplicity of the underlying principle, ease of practical implementation of the resulting “global” decision rule) will be outweighed by the following drawbacks:

1. The test of  $H : \max_{i < j} \{\rho(\theta_i, \theta_j)\} \geq \varepsilon$  versus  $K : \max_{i < j} \{\rho(\theta_i, \theta_j)\} < \varepsilon$  obtained by applying the intersection-union principle with  $q = \binom{k}{2}$  and  $H_{ij} : \rho(\theta_i, \theta_j) \geq \varepsilon$  vs.  $K_{ij} : \rho(\theta_i, \theta_j) < \varepsilon$  ( $1 \leq i < j \leq k$ ) as the elementary testing problems can be markedly overconservative.
  2. Looking at the maximum distance between pairs of distributions might not adequately translate an investigator's notion of equivalence between all  $k$  distributions into a testable statistical hypothesis; for other distance measures, the intersection-union principle is typically no longer applicable.

In the following sections of the present chapter, we show for selected settings frequently encountered in practice, how to replace the intersection-union test with a less conservative specific multisample test for equivalence.

## 7.2 $F$ -test for equivalence of $k$ normal distributions

The considerations of this (as well as of the subsequent) section refer to the classical univariate ANOVA one-way layout with normally distributed observations of common (yet unknown) variance  $\sigma^2 \in \mathbb{R}_+$ . Thus we suppose that the pooled data set to be analyzed exhibits the following structure:

$$\begin{array}{ll} X_{11}, \dots, X_{1n_1} & \leftarrow \text{1st sample, from } \mathcal{N}(\mu_1, \sigma^2) \\ X_{21}, \dots, X_{2n_2} & \leftarrow \text{2nd sample, from } \mathcal{N}(\mu_2, \sigma^2) \\ \vdots & \vdots \quad \ddots, \quad \vdots \\ X_{k1}, \dots, X_{kn_k} & \leftarrow \text{kth sample, from } \mathcal{N}(\mu_k, \sigma^2) \end{array}$$

In the balanced case  $n_1 = \dots = n_k = n$ , a natural measure of the degree to which the true underlying distributions deviate from the “ideal” of perfect pairwise coincidence is given by the squared Euclidean distance of the vector  $(\mu_1/\sigma, \dots, \mu_k/\sigma)$  from the point  $(\bar{\mu}/\sigma) \mathbf{1}_k$ , where  $\bar{\mu}$ . and  $\mathbf{1}_k$  denotes the ordinary arithmetic mean of the  $\mu_i$  and the  $k$ -vector of ones, respectively. A useful way of extending the definition of this measure to arbitrarily unbalanced designs of the same kind consists of replacing  $\bar{\mu}$ . with the average  $\tilde{\mu}$ ., say, of the sample-size-weighted population means  $n_i \mu_i$ , and the unit weight to the

squared distance in the  $i$ th component with  $(n_i/\bar{n})$ . The resulting generalized squared Euclidean distance is then given by

$$\psi^2 = \sum_{i=1}^k (n_i/\bar{n})(\mu_i - \tilde{\mu}_.)^2 / \sigma^2, \quad (7.6)$$

with

$$\tilde{\mu}_. = \sum_{i=1}^k n_i \mu_i / \sum_{i=1}^k n_i, \quad \bar{n} = \sum_{i=1}^k n_i / k. \quad (7.7)$$

Using  $\psi^2$  as our target parameter leads to the problem of testing

$$H : \psi^2 \geq \varepsilon^2 \quad \text{versus} \quad K : \psi^2 < \varepsilon^2 \quad (7.8)$$

with suitably fixed  $\varepsilon > 0$ .

Clearly, in any balanced design, equivalence in the sense of (7.8) is satisfied if and only if the point with the standardized population means as coordinates lies in a  $k$ -dimensional sphere of radius  $\varepsilon$  around  $(\bar{\mu}_./\sigma, \dots, \bar{\mu}_./\sigma)$ . In the case of  $k = 2$  samples with common size  $n_1 = n_2 = n$ , this condition simplifies still further in that it turns out equivalent to  $|\mu_1 - \mu_2|/\sigma < \sqrt{2}\varepsilon$ . This suggests adopting the guidance given in Table 1.1(v) for specifying the tolerance  $\varepsilon$  in the two-sample case, also for arbitrary multisample designs choosing  $\varepsilon = .36/\sqrt{2} \approx .25$  and  $\varepsilon = .74/\sqrt{2} \approx .50$ , depending on whether the equivalence criterion is intended to be rather strict or comparatively weak.

The testing problem (7.8) remains invariant under a large group  $\mathcal{G}$ , say, of affine transformations which contains in particular arbitrary common rescalings and translations of all observations in the pooled sample (an exhaustive characterization of  $\mathcal{G}$  involves canonical coordinates and is not required in the present context; for full mathematical details see Lehmann and Romano, 2005, § 7.1). The usual  $F$ -statistic for the one-way fixed-effects ANOVA is a maximal invariant with respect to this group of transformations. Dropping the constant factor  $\bar{n}/(k-1)$ , it can be viewed as an estimator of the target parameter  $\psi^2$  so that we write

$$\hat{\psi}^2 = \frac{\sum_{i=1}^k (n_i/\bar{n})(\bar{X}_i - \bar{X}_{..})^2}{(N-k)^{-1} \sum_{i=1}^k \sum_{\nu=1}^{n_i} (X_{i\nu} - \bar{X}_i)^2} \quad (7.9)$$

The distribution of  $(\bar{n}/(k-1))\hat{\psi}^2$  depends on the  $\mu_i$  and  $\sigma^2$  only through  $\psi^2$  and is noncentral  $F$  with  $k-1, N-k$  degrees of freedom (where  $N$  denotes the total sample size  $\sum_{i=1}^k n_i$ ) and noncentrality parameter  $\lambda_{nc}^2 = \bar{n}\psi^2$ . It is thus an element of a family with strictly monotone likelihood ratios, and it follows that the class of all level- $\alpha$  tests for (7.8) which depend on the raw

data only through  $\hat{\psi}^2$  contains a uniformly most powerful one. The critical region of this optimal test is given by

$$\left\{ \hat{\psi}^2 < ((k-1)/\bar{n}) F_{k-1, N-k; \alpha}(\bar{n}\varepsilon^2) \right\} \quad (7.10)$$

with

$$F_{k-1, N-k; \alpha}(\bar{n}\varepsilon^2) \equiv \begin{aligned} &\text{lower } 100\alpha\text{-percentage point of a non-} \\ &\text{central } F\text{-distribution with } k-1, N-k \\ &\text{degrees of freedom and noncentrality} \\ &\text{parameter } \bar{n}\varepsilon^2. \end{aligned} \quad (7.11)$$

In view of the maximal invariance of the statistic  $\hat{\psi}^2$  with respect to the group  $\mathcal{G}$  of all transformations of the full sample space leaving the testing problem invariant, the rejection region (7.10) defines an UMPI level- $\alpha$  test for (7.8). Its size is exactly equal to  $\alpha$ , and the power against any specific alternative is strictly greater than the significance level ( $\rightarrow$  strictly unbiased testing procedure). The exact rejection probability under any parameter constellation depends only on  $\psi^2$  and can easily be computed by means of the formula:

$$\beta(\psi^2) = P[\mathcal{F}_{k-1, N-k}(\bar{n}\psi^2) < F_{k-1, N-k; \alpha}(\bar{n}\varepsilon^2)] \quad (7.12)$$

where  $\mathcal{F}_{k-1, N-k}(\tilde{\psi}^2)$  stands for a variable which has an  $F$ -distribution with  $k-1, N-k$  degrees of freedom and noncentrality parameter  $\tilde{\psi}^2$ . In particular, using SAS one will obtain the value of  $\beta(\psi^2)$  simply by submitting the statements

```
c=finv(alpha,k-1,N-k,n_*eps**2);
beta=probft(c,k-1,N-k,n_*psi**2);
```

provided, the value specified for  $\alpha$ ,  $k$ ,  $N$ ,  $\bar{n}$ ,  $\varepsilon$  and  $\psi$  has been previously assigned to the variable `alpha`, `k`, `N`, `n_`, `eps` and `psi`, respectively.

### Example 7.1

In a comparative clinical trial of 4 different antihypertensive treatments, blood pressure was recorded continuously over 24 hours in each patient enrolled. The AUC (area under the curve) for the diastolic value divided by the length of the measurement interval was defined as the primary endpoint for assessing efficacy of treatment. The raw data [mm Hg] obtained in this way are listed in the following table:

Table 7.1 Raw data of a 4-arm trial using the average of diastolic blood pressure recorded continuously over 24 hours as primary endpoint.

<i>Sample 1</i>
112.488, 103.738, 86.344, 101.708, 95.108, 105.931, 95.815, 91.864, 102.479, 102.644
<i>Sample 2</i>
100.421, 101.966, 99.636, 105.983, 88.377, 102.618, 105.486, 98.662, 94.137, 98.626, 89.367, 106.204
<i>Sample 3</i>
84.846, 100.488, 119.763, 103.736, 93.141, 108.254, 99.510, 89.005, 108.200, 82.209, 100.104, 103.706, 107.067
<i>Sample 4</i>
100.825, 100.255, 103.363, 93.230, 95.325, 100.288, 94.750, 107.129, 98.246, 96.365, 99.740, 106.049, 92.691, 93.111, 98.243

The sample means and standard deviations computed from these raw data are shown in Table 7.2. Hence, in the present example, we have:

$$\begin{aligned}
 N &= 50, \quad k = 4, \quad \bar{n} = 12.5; \quad \bar{X}_{..} = 99.3849; \\
 \sum_{i=1}^k (n_i/\bar{n})(\bar{X}_i - \bar{X}_{..})^2 &= 1.2157; \\
 (N-k)^{-1} \sum_{i=1}^k \sum_{\nu=1}^{n_i} (X_{i\nu} - \bar{X}_i)^2 &= 54.6976; \\
 \Rightarrow \hat{\psi}^2 &= 1.2157/54.6976 = .0222.
 \end{aligned}$$

Specifying  $\alpha = .05$  and  $\varepsilon = .50$ , the critical upper bound of (7.10) comes out (using SAS) as

$$\begin{aligned}
 c &= (3/12.5) \cdot F_{4-1,50-4; .05}(12.5 \cdot .25) = .24 \cdot \text{finv}(.05, 3, 46, 3.125) \\
 &= .24 \cdot .30833 = .0740.
 \end{aligned}$$

Since the observed value of the test statistic  $\hat{\psi}^2$  falls below this bound, the UMPI test rejects nonequivalence. Finally, evaluating the right-hand side of equation (7.12) yields for the power of the test against the specific alternative of identical distributions:

$$\beta(0) = P[\mathcal{F}_{3,46}(0) < (12.5/3) \cdot c] = \text{probf}(.30833, 3, 46, 0) = 18.08\%.$$

Table 7.2 Summary statistics for the raw data of Table 7.1 [with  $i$  indexing the individual samples].

$i$	$n_i$	$\bar{X}_i$	$S_i$
1	10	99.8120	7.5639
2	12	99.2903	5.9968
3	13	100.0024	10.4809
4	15	98.6407	4.5309

### 7.3 Modified studentized range test for equivalence

Under the additional assumption that the one-way ANOVA layout we are dealing with is a *balanced* one satisfying  $n_1 = \dots = n_k = n$ , an alternative approach to testing for equivalence of an arbitrary number  $k$  of normal distributions with means  $\mu_i \in \mathbb{R}$  and common variance  $\sigma^2 > 0$  can be based on the studentized range statistic. Of course, such a change in the test statistic would make little if any sense without a corresponding reformulation of hypotheses in terms of a different parametric function viewed as the basic measure of dissimilarity of the distributions under assessment. In the present context, this measure is chosen to be the maximum of all pairwise distances between standardized population means which leads to replacing (7.8) with the testing problem

$$\tilde{H}_\mu : \max_{i < j} |\mu_i - \mu_j|/\sigma \geq \delta \quad \text{versus} \quad \tilde{K}_\mu : \max_{i < j} |\mu_i - \mu_j|/\sigma < \delta, \quad (7.13)$$

with arbitrarily fixed equivalence margin  $\delta > 0$ . The studentized range statistic which will be denoted  $R_s$  in the sequel, is obtained by simply replacing both  $\max_{i < j} |\mu_i - \mu_j|$  and  $\sigma$  with its sample analogue yielding

$$R_s = \frac{\bar{X}_{(k)} - \bar{X}_{(1)}}{S} \quad (7.14)$$

where  $\bar{X}_{(i)}$  stands for the  $i$ th-smallest ( $i = 1, \dots, k$ ) among the  $k$  sample means, and  $S^2$  is the usual ANOVA estimator of the error variance as appearing in the denominator to (7.9).

In the equivalence case, determining a suitable critical constant for the statistic  $R_s$  is a rather complicated problem which has been solved by Giani and Finner (1991). As shown by these authors, the critical bound  $r_{\alpha; \delta}(n, k)$ ,

say, which the studentized range statistic has to be compared with in order to obtain a valid test for (7.13), has to be computed as the  $\alpha$ -quantile of the distribution of  $R_s$  under the specific parameter constellation  $(\mu_1/\sigma, \dots, \mu_k/\sigma) = (-\delta/2, 0, \dots, 0, \delta/2)$ . Fortunately, a program has been made available by the same group of authors (Giani et al., 2005) which computes  $r_{\alpha; \delta}(n, k)$  for arbitrary combinations of  $\alpha$ ,  $\delta$ ,  $n$  and  $k$  by numerical integration. Table 7.3 shows the critical upper bounds to  $R_s$  to be used in equivalence testing at level  $\alpha = .05$  for  $\delta \in \{1/2, 1\}$  with up to 5 samples of size 10 and 50, respectively.

Table 7.3 *Critical bounds for the studentized range test for equivalence at level  $\alpha = .05$ , for  $\delta = 1/2, 1$ ,  $n = 10, 50$  and  $k = 3, 4, 5$  [computed by means of the program SeParATE, Version 3.0B (Giani et al., 2005)].*

$\delta$	$n$	$k$	$r_{\alpha; \delta}(n, k)$
.50	10	3	.185896
"	10	4	.293307
"	10	5	.376803
"	50	3	.221237
"	50	4	.255186
"	50	5	.281114
1.00	10	3	.407372
"	10	4	.494329
"	10	5	.559521
"	50	3	.669046
"	50	4	.673074
"	50	5	.676400

### Example 7.1 (continued)

Suppose  $99.8120 \pm 7.5639$ ,  $99.2903 \pm 5.9968$ ,  $100.0024 \pm 10.4809$ ,  $98.6407 \pm 4.5309$  [cf. Table 7.2] would actually describe (in terms of  $\bar{X}_i \pm S_i$ ) the data obtained in a *balanced* trial involving  $k = 4$  groups of size  $n = 10$  each. Then, the estimator of the error variance would take on the value  $S^2 = (7.5639^2 + 5.9968^2 + 10.4809^2 + 4.5309^2)/4 = 55.8881$ , and for the studentized range statistic we would have

$$R_s = (\bar{X}_{(4)} - \bar{X}_{(1)})/S = (100.0024 - 98.6407)/\sqrt{55.8881} = .1821.$$

On the other hand, the critical bound to which  $R_s$  must be compared in a test at level  $\alpha = .05$  of  $\max_{i < j} |\mu_i - \mu_j|/\sigma \geq 1.00$  versus  $\max_{i < j} |\mu_i - \mu_j|/\sigma < 1.00$  with four groups of 10 observations each is seen from the above table to

be  $r_{.05; 1.00}(10, 4) = .494329$ . Hence, with the data and the specifications of this (synthetic) example, the modified studentized range test decides in favor of equivalence again.

Before concluding the discussion of approaches to testing for equivalence of several homoskedastic normal distributions with respect to their standardized means, it seems worthwhile to point out the *most important differences and similarities between the modified studentized range and the noncentral F-test* of § 7.2:

- Both tests are exactly valid with respect to the significance level, provided there are no deviations from the basic assumptions of the classical ANOVA model.
- The critical constant of the noncentral  $F$ -test can easily be computed without recourse to nonstandard software.
- The applicability of the studentized range statistic is restricted to balanced designs whereas the noncentral  $F$ -test allows for arbitrary imbalances between the sample sizes.
- Power comparisons between the procedures are futile since the hypotheses they are tailored for are of grossly different shape (except for  $k = 2$ ).

For the special case of testing for equivalence of just three normal distributions with respect to their standardized means, a range test which is also applicable for unequal sample sizes, has been presented by Wiens and Iglewicz (2000) (see also Wiens and Iglewicz, 1999). The procedure does even allow for heteroskedasticity. However, its usefulness is largely restricted by the fact that the critical constants for the test statistic were determined under the assumption that the population variances are known.

---

## 7.4 Testing for dispersion equivalence of more than two Gaussian distributions

If homoskedasticity of the normal distributions behind the samples  $(X_{11}, \dots, X_{1n_1}), \dots, (X_{k1}, \dots, X_{kn_k})$  cannot be taken for granted *a priori*, then it is natural that one tries to establish it by means of the actual data. In order to keep the description of a large-sample procedure for an equivalence hypothesis specifying homoskedasticity except for negligible heterogeneity of the variances as simple as possible, we restrict consideration first to balanced designs. Furthermore, it will be convenient to use the following additional

notation:

$$\zeta_i = \log \sigma_i, \quad \bar{\zeta} = k^{-1} \sum_{i=1}^k \log \sigma_i; \quad (7.15a)$$

$$Z_i = \log S_i, \quad \bar{Z} = k^{-1} \sum_{i=1}^k \log S_i. \quad (7.15b)$$

A reasonable way of translating the phrase “homoskedastic except for irrelevant differences in variability” into a formal statement about a suitable parametric function consists in requiring that all logarithmic standard deviations  $\zeta_i$  lie in a sphere of sufficiently small radius  $\varepsilon_*$  around their average  $\bar{\zeta}$ . This leads to proposing the problem of testing

$$H_\sigma : \sum_{i=1}^k (\zeta_i - \bar{\zeta})^2 \geq \varepsilon_*^2 \quad \text{versus} \quad K_\sigma : \sum_{i=1}^k (\zeta_i - \bar{\zeta})^2 < \varepsilon_*^2 \quad (7.16)$$

with some suitably fixed  $\varepsilon_* > 0$ . Some guidance for choosing the equivalence margin for  $\sum_{i=1}^k (\zeta_i - \bar{\zeta})^2$  can be provided by noting that in the two-sample case, dispersion equivalence in the sense of the above alternative hypothesis  $K_\sigma$  reduces to the condition  $(1+\varepsilon)^{-1} < \sigma_1/\sigma_2 < 1+\varepsilon$  with  $\varepsilon$  being related to  $\varepsilon_*$  by the equation  $\varepsilon_* = (1/\sqrt{2}) \log(1+\varepsilon)$ . In view of this fact, generalizing the suggestions given in Table 1.1 for choosing the equivalence margin for the case of  $k = 2$  variances leads to specifying  $\varepsilon_* = .2867$  or  $\varepsilon_* = .4901$ , depending on whether a rather strict or more liberal criterion of equivalence is intended to apply.

In order to construct a large-sample test for (7.16), we exploit the well-known fact (see, e.g., Lehmann, 1986, p. 376) that the observed log-standard deviations  $Z_1 = \log S_1, \dots, Z_k = \log S_k$  are asymptotically normal and mutually independent estimators of the corresponding population parameters  $\zeta_1 = \log \sigma_1, \dots, \zeta_k = \log \sigma_k$ . Since all of these estimators have the same asymptotic variance  $(2n)^{-1}$ , this means that for large  $n$ , the  $Z_i = \log S_i$  have distributions  $\mathcal{N}(\log \sigma_i, 1/2n)$ . In view of the mutual independence of all statistics depending on individual variables from different samples, this implies that for arbitrary values of the true  $\sigma_i$ 's, we have

$$P \left[ 2n \sum_{i=1}^k (Z_i - \bar{Z})^2 \leq v^2 \right] \approx G_{k-1, \lambda_{nc}^2}(v^2) \quad \forall v^2 > 0 \quad (7.17)$$

with  $\lambda_{nc}^2 = 2n \sum (\zeta_i - \bar{\zeta})^2$  and  $G_{k-1, \lambda_{nc}^2}(\cdot)$  as the cdf of a  $\chi^2$ -distribution with  $df = k - 1$  and noncentrality parameter  $\lambda_{nc}^2$ .

In view of (7.17), an approximately valid test at level  $\alpha$  for (7.16) is given by the rejection region

$$\left\{ Q_{k,n} < q_{k,n; \alpha}(\varepsilon_*) \right\} \quad (7.18)$$

where

$$Q_{k,n} = 2n \sum_{i=1}^k \left[ \log S_i - (1/k) \sum_{\ell=1}^k \log S_{\ell} \right]^2 \quad (7.19)$$

and

$$q_{k,n; \alpha}(\varepsilon_*) = \chi_{k-1; \alpha}^2(2n\varepsilon_*^2) \equiv \text{lower } 100\alpha \text{ percentage point}$$

of a  $\chi^2$ -distribution with  $k-1$  degrees of freedom

$$\text{and noncentrality parameter } \lambda_{nc}^2 = 2n\varepsilon_*^2. \quad (7.20)$$

The power of this test against any alternative  $(\sigma_1, \dots, \sigma_k) \in \mathbb{R}_+^k$  such that  $2n \sum_{i=1}^k \left[ \log \sigma_i - (1/k) \left( \sum_{\ell=1}^k \log \sigma_{\ell} \right) \right]^2 = \tilde{\lambda}_{nc}^2 < 2n\varepsilon_*^2$  can be approximated as

$$\beta(\tilde{\lambda}_{nc}^2) \approx G_{k-1; \tilde{\lambda}_{nc}^2}(\chi_{k-1; \alpha}^2(2n\varepsilon_*^2)). \quad (7.21)$$

### Example 7.2

The data displayed in Table 7.4 are sample variances and corresponding log-standard deviations observed in  $k = 4$  groups of common size  $n = 50$ . The Gaussian distributions behind the samples are to be tested for dispersion equivalence in the sense of (7.16) where the radius  $\varepsilon_*$  of the equivalence sphere centered at the mean of the logarithmic population standard deviations is specified  $\varepsilon_* = .2867$  corresponding to a ratio of 1.5 at most in the two-sample case. The significance level is once more fixed at  $\alpha = .05$ .

Table 7.4 Variances and logarithmic standard deviations computed from a raw data set consisting of  $k = 4$  independent samples of common size  $n = 50$ .

i	1	2	3	4
$S_i^2$	49.4	45.7	43.1	53.6
$Z_i$	1.9500	1.9110	1.8818	1.9908

For this data set, the value of the test statistic turns out as

$$2 \cdot 50 \cdot \sum_{i=1}^k (Z_i - \bar{Z})^2 = 100 \cdot .006735 = .6735.$$

On the other hand, the critical upper bound which  $2n \sum_{i=1}^k (Z_i - \bar{Z})^2$  has to be compared to, is readily computed (using once more the respective SAS function) to be

$$\chi^2_{3,0.05}(2 \cdot 50 \cdot .2867^2) = \text{cinv}(.05, 3, 8.2197) = 2.85636.$$

Since the value observed for the test statistic clearly falls short of this bound, the data set shown in the above table falls in the rejection region (7.18) so that we have to decide in favor of dispersion equivalence of the underlying normal distributions. In order to approximate the power of the test against the alternative of exact homoskedasticity, we invoke the SAS intrinsic function for the cdf rather the quantile function of the  $\chi^2$ -distribution with  $k - 1 = 3$  degrees of freedom, giving

$$\beta(0) \approx G_{3;0}(2.85636) = \text{probchi}(2.85636, 3, 0) = 58.57\%.$$

Although it is not our intention here to enter into a detailed investigation of the accuracy of the approximation (7.17) underlying the testing procedure given by (7.18–20), it should be mentioned that sufficient simulation results are available to justify the following qualitative statements: Even for large numbers  $k$  of distributions to compare, the probability of rejecting the null hypothesis of (7.16) according to (7.18) approaches the nominal significance level  $\alpha$  rather quickly with increasing common sample size  $n$ , under parameter configurations lying on the common boundary of  $H_\sigma$  and  $K_\sigma$ . Furthermore, if nonnegligible discrepancies between  $\alpha$  and exact rejection probabilities under parameter configurations  $(\sigma_1, \dots, \sigma_k)$  satisfying  $\sum_{i=1}^k [\log \sigma_i - (1/k) \sum_{\ell=1}^k \log \sigma_\ell]^2 = \varepsilon_*^2$  occur at all, the testing procedure is found to be over- rather than anticonservative throughout.

### *Generalization to the unbalanced case*

Unbalancedness of the design requires only minor modifications to the test. It suffices to replace in all sums appearing in the formula for the test statistic and the noncentrality parameter, unit weights by the weighing factors  $n_i/\bar{n}$ , with  $\bar{n} = N/k = \sum_{i=1}^k n_i/k$ .

By plugging in these weights the alternative hypothesis of dispersion equivalence takes on the form

$$K'_\sigma : \sum_{i=1}^k (n_i/\bar{n}) \left( \zeta_i - \sum_{\ell=1}^k n_\ell \zeta_\ell / N \right)^2, \quad (7.16')$$

and the test statistic has to be computed as

$$Q'_{k,n} \equiv 2 \sum_{i=1}^k n_i \left[ \log S_i - \sum_{\ell=1}^k n_\ell \log S_\ell / N \right]^2. \quad (7.19')$$

Finally, the critical upper bound to which the observed value of  $Q'_{k,n}$  has to be compared, is given by

$$q'_{k,n;\alpha}(\varepsilon_*) = \chi^2_{k-1;\alpha}(2\bar{n}\varepsilon_*^2). \quad (7.20')$$


---

## 7.5 A nonparametric $k$ -sample test for equivalence

The approach to multisample equivalence testing to be discussed in this section can be viewed as a natural extension of the Mann-Whitney test for equivalence of two continuous distributions of arbitrary shape as described in § 6.2. Although the basic idea behind this extension is almost self-explanatory, the construction is technically rather intricate involving a lengthy sequence of formulae for the various entries in the covariance matrix of a vector of  $U$ -statistics whose dimension grows with the square of the number of distributions under comparison. Instead of writing down all these details which would be rather tedious, we confine the exposition to an outline of major steps to be taken in carrying out the construction.

First of all, we must define a sensible nonparametric measure of distance between all  $k$  distributions. This can be obtained by measuring the distance between each individual pair of distributions in the same way as in the two-sample case and combining the resulting  $k(k - 1)/2$  functionals into a single overall measure by means of a suitable algebraic operation. In order to make this idea precise, let us define for any  $1 \leq i < j \leq k$

$$\pi_{ij}^+ = P[X_i > X_j] = \int F_j dF_i \quad (7.22)$$

where  $X_i$  and  $X_j$  denote “generic” observations, independent of each other, from population  $i$  and  $j$  with cdf  $F_i$  and  $F_j$ , respectively. In view of the obvious fact that  $F_i = F_j \Rightarrow \pi_{ij}^+ = 1/2$ , we proposed in § 6.2  $|\pi_{ij}^+ - 1/2|$  as a suitable nonparametric distance measure for the case of two samples. A natural generalization to the  $k$ -sample setting is to replace  $|\pi_{ij}^+ - 1/2|$  with the Euclidean distance between the vector  $(\pi_{12}^+, \dots, \pi_{1k}^+, \pi_{23}^+, \dots, \pi_{k-1,k}^+)$  of all pairwise proversion probabilities (7.22), and that point in the  $(k(k - 1)/2)$ -dimensional unit interval whose coordinates all equal  $1/2$ . Using this distance in its squared form (which turns out technically a bit more convenient), we are lead to formulate the testing problem

$$H : \sum_{i < j} (\pi_{ij}^+ - 1/2)^2 \geq \tilde{\varepsilon}^2 \quad \text{versus} \quad K : \sum_{i < j} (\pi_{ij}^+ - 1/2)^2 < \tilde{\varepsilon}^2 \quad (7.23)$$

with suitably specified equivalence margin  $\tilde{\varepsilon}^2 > 0$ .

In order to keep notation sufficiently compact, we introduce an extra symbol for representing the squared Euclidean distance between any two vectors in  $(k(k-1)/2)$ -dimensional space defining

$$d^2(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^{k(k-1)/2} (u_i - v_i)^2, \quad \mathbf{u}, \mathbf{v} \in \mathbb{R}^{k(k-1)/2}. \quad (7.24)$$

Using this convention, the condition assumed to be satisfied under the alternative hypothesis of (7.23) can be written  $d^2(\boldsymbol{\pi}^+, \mathbf{1}/\mathbf{2}) < \tilde{\varepsilon}^2$ , provided the symbols appearing as arguments to  $d^2(\cdot, \cdot)$  in that inequality are read as

$$\boldsymbol{\pi}^+ = (\pi_{12}^+, \dots, \pi_{1k}^+, \pi_{23}^+, \dots, \pi_{k-1,k}^+), \quad \mathbf{1}/\mathbf{2} = (1/2) \cdot \mathbf{1}_{k(k-1)/2}, \quad (7.25)$$

with  $\mathbf{1}_m$  denoting the vector of  $m$  1's for arbitrary  $m \in \mathbb{N}$ .

Now, a natural choice of a statistic which an asymptotic test for establishing the hypothesis  $K : d^2(\boldsymbol{\pi}^+, \mathbf{1}/\mathbf{2}) < \tilde{\varepsilon}^2$  can be based upon, is this: We estimate the squared distance between  $\boldsymbol{\pi}^+$  and  $\mathbf{1}/\mathbf{2}$  by that between the corresponding vector  $\mathbf{W}^+$ , say, of two-sample  $U$ -statistics, and the same point of reference in the parameter space  $[0, 1]^{k(k-1)/2}$ . Of course, for any  $1 \leq i < j \leq k$ , the  $(i, j)$ -component of  $\mathbf{W}^+$  has to be defined by

$$W_{ij}^+ = \frac{1}{n_i n_j} \sum_{\mu=1}^{n_i} \sum_{\nu=1}^{n_j} I_{(0, \infty)}(X_{i\mu} - X_{j\nu}) \quad (7.26)$$

[recall (6.13)].

The most tedious part of the construction consists in establishing a procedure for estimating the asymptotic standard error of the estimated distance  $d^2(\mathbf{W}^+, \mathbf{1}/\mathbf{2})$ . For that purpose, we need to know the full covariance matrix, say  $\boldsymbol{\Sigma}^{(N)} = (\sigma_{(i_1, j_1), (i_2, j_2)}^{(N)})_{1 \leq i_1 < j_1 \leq k, 1 \leq i_2 < j_2 \leq k}$ , of  $\sqrt{N}\mathbf{W}^+$ . For each element of  $\boldsymbol{\Sigma}^{(N)}$ , an exact expression can be found which is a function of the sizes of the samples involved, the appropriate components of  $\boldsymbol{\pi}^+$ , and functionals of the form  $P[X_{i_1} > X_{j_1}, X_{i_2} > X_{j_2}]$ . The resulting formulae for  $\sigma_{(i_1, j_1), (i_2, j_2)}^{(N)}$  differ markedly in structure, depending on the number of common elements in the set  $\{i_1, j_1, i_2, j_2\}$  of indices involved. For  $i_1 \neq i_2, j_1 \neq j_2$ ,  $W_{i_1 j_1}^+$  and  $W_{i_2 j_2}^+$  refer to four different samples, which trivially implies that we have

$$\sigma_{(i_1, j_1), (i_2, j_2)}^{(N)} = 0 \text{ for all } (i_1, j_1), (i_2, j_2) \text{ with } i_1 \neq i_2 \text{ and } j_1 \neq j_2. \quad (7.27)$$

The other extreme is given by the diagonal elements of  $\boldsymbol{\Sigma}^{(N)}$ : In this case,  $\{i_1, j_1, i_2, j_2\}$  reduces to a set of cardinality 2, say  $\{i, j\}$ , and  $\sigma_{(i, j), (i, j)}^{(N)}$  is readily obtained by applying formulae (6.14), (6.15) for the computation of the exact variance of a single Mann-Whitney statistic. If the pairs  $(i_1, j_1)$  and  $(i_2, j_2)$  coincide with respect to the first component but differ in the second,

we have

$$\begin{aligned}\sigma_{(i_1, j_1), (i_2, j_2)}^{(N)} &= (N/n_i) \left( P[X_i > X_{j_1}, X_i > X_{j_2}] - \pi_{ij_1}^+ \pi_{ij_2}^+ \right) \\ &= (N/n_i) \left( \int F_{j_1} F_{j_2} dF_i - \int F_{j_1} dF_i \int F_{j_2} dF_i \right) \\ \text{for } i &= i_1 = i_2,\end{aligned}\quad (7.28)$$

and so on for the other possible constellations with  $\#(\{i_1, j_1\} \cap \{i_2, j_2\}) = 1$ .

Obviously, all these expressions converge to well-defined limits, say  $\sigma_{(i_1, j_1), (i_2, j_2)}$ , as  $N \rightarrow \infty$ , provided the relative size  $n_i/N$  of each sample converges to some nondegenerate limit  $\lambda_i \in (0, 1)$  as the total sample size grows to infinity. Furthermore, it follows from well-known results of the asymptotic theory of generalized  $U$ -statistics (cf. Lee, 1990, § 3.7.1) that the existence of nondegenerate limits for all  $n_i/N$  ensures weak convergence of the normalized vector  $\sqrt{N}(\mathbf{W}^+ - \boldsymbol{\pi}^+)$  of pairwise Mann-Whitney statistics to a centered  $(k(k-1))$ -variate normal distribution with covariance matrix  $\Sigma = \lim_{N \rightarrow \infty} \Sigma^{(N)}$ . Hence, another application of the  $\delta$ -method yields

$$\begin{aligned}\sqrt{N}(d^2(\mathbf{W}^+, \mathbf{1}/\mathbf{2}) - d^2(\boldsymbol{\pi}^+, \mathbf{1}/\mathbf{2})) \\ \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2[d^2(\mathbf{W}^+, \mathbf{1}/\mathbf{2})]) \text{ as } N \rightarrow \infty\end{aligned}\quad (7.29)$$

with

$$\sigma^2[d^2(\mathbf{W}^+, \mathbf{1}/\mathbf{2})] = 4(\boldsymbol{\pi}^+ - \mathbf{1}/\mathbf{2})\Sigma(\boldsymbol{\pi}^+ - \mathbf{1}/\mathbf{2})'. \quad (7.30)$$

Finally, under some mild additional regularity conditions, each entry  $\sigma_{(i_1, j_1), (i_2, j_2)}^{(N)}$  in the exact covariance matrix  $\Sigma^{(N)}$  can be consistently estimated by means of a suitable function of two-sample ( $\rightarrow$  diagonal elements) or two- and three-sample  $U$ -statistics ( $\rightarrow$  off-diagonal elements). In view of  $\sigma_{(i_1, j_1), (i_2, j_2)}^{(N)} \rightarrow \sigma_{(i_1, j_1), (i_2, j_2)} \forall ((i_1, j_1), (i_2, j_2))$ , these estimators are *a fortiori* consistent for the respective elements of the limiting covariance matrix  $\Sigma$ . Thus, if we denote by  $\widehat{\Sigma}^{(N)}$  the matrix obtained by replacing each  $\sigma_{(i_1, j_1), (i_2, j_2)}^{(N)}$  with its consistent estimator, the asymptotic variance (7.30) of  $\sqrt{N}d^2(\mathbf{W}^+, \mathbf{1}/\mathbf{2})$  can be consistently estimated by

$$\hat{\sigma}_N^2[d^2(\mathbf{W}^+, \mathbf{1}/\mathbf{2})] = 4(\mathbf{W}^+ - \mathbf{1}/\mathbf{2})\widehat{\Sigma}^{(N)}(\mathbf{W}^+ - \mathbf{1}/\mathbf{2})'. \quad (7.31)$$

Hence, we have

$$\frac{\sqrt{N}(d^2(\mathbf{W}^+, \mathbf{1}/\mathbf{2}) - d^2(\boldsymbol{\pi}^+, \mathbf{1}/\mathbf{2}))}{\hat{\sigma}_N[d^2(\mathbf{W}^+, \mathbf{1}/\mathbf{2})]} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) \text{ as } N \rightarrow \infty \quad (7.32)$$

under any  $k$ -tuple  $(F_1, \dots, F_k)$  of continuous cdf's such that  $\Sigma$  is positive definite and  $\boldsymbol{\pi}^+ \neq \mathbf{1}/\mathbf{2}$ . Accordingly, an approximate level- $\alpha$  test for (7.23) can be based on the decision rule:

Reject nonequivalence in the sense of  $d^2(\boldsymbol{\pi}^+, \mathbf{1}/\mathbf{2}) \geq \tilde{\varepsilon}^2$  iff

$$d^2(\mathbf{W}^+, \mathbf{1}/\mathbf{2}) < \tilde{\varepsilon}^2 - u_{1-\alpha} \cdot \hat{\sigma}_N[d^2(\mathbf{W}^+, \mathbf{1}/\mathbf{2})]/\sqrt{N}. \quad (7.33)$$

*Example 7.1 (continued)*

To illustrate the nonparametric approach to multisample equivalence testing described above, we reanalyze the data set introduced in Example 7.1. With the entries in Table 7.1, the  $k(k - 1)/2 = 6$  pairwise Mann-Whitney statistics are computed to be:

$$W_{12}^+ = .52500, \quad W_{13}^+ = .49231, \quad W_{14}^+ = .57333;$$

$$W_{23}^+ = .44231, \quad W_{24}^+ = .56667, \quad W_{34}^+ = .58974.$$

Thus, the result of estimating the squared distance between  $\boldsymbol{\pi}^+$  and  $\mathbf{1}/\mathbf{2}$  is

$$\begin{aligned} d^2(\mathbf{W}^+, \mathbf{1}/\mathbf{2}) &= (.52500 - .5)^2 + (.49231 - .5)^2 + (.57333 - .5)^2 \\ &\quad + (.44231 - .5)^2 + (.56667 - .5)^2 + (.58974 - .5)^2 \\ &= .021888. \end{aligned}$$

For the estimated exact covariance matrix  $\hat{\Sigma}^{(N)}$  of  $\sqrt{N} \mathbf{W}^+$ , we obtain

$$N^{-1} \hat{\Sigma}^{(N)} = \begin{pmatrix} .01392 & .00704 & .01112 & -.00386 & -.00650 & .00000 \\ .00704 & .01271 & .00777 & .01012 & .00000 & -.01026 \\ .01112 & .00777 & .01487 & .00000 & .00371 & .00214 \\ -.00386 & .01012 & .00000 & .01227 & .00440 & -.01142 \\ -.00650 & .00000 & .00371 & .00440 & .01195 & .00276 \\ .00000 & -.01026 & .00214 & -.01142 & .00276 & .01257 \end{pmatrix}$$

Plugging in the observed values of  $\mathbf{W}^+$  and  $N^{-1} \hat{\Sigma}^{(N)}$  on the right-hand side of (7.31), the empirical standard error of  $d^2(\mathbf{W}^+, \mathbf{1}/\mathbf{2})$  is found to be

$$\begin{aligned} \hat{\sigma}_N [d^2(\mathbf{W}^+, \mathbf{1}/\mathbf{2})]/\sqrt{N} &= \\ 2 \cdot [( .02500, -.00769, .07333, -.05769, .06667, .08974) (N^{-1} \hat{\Sigma}^{(N)}) \\ ( .02500, -.00769, .07333, -.05769, .06667, .08974)' ]^{1/2} &= .045094. \end{aligned}$$

Fixing the significance level at the conventional value  $\alpha = .05$ , there remains to specify the equivalence margin  $\tilde{\varepsilon}^2$  for  $d^2(\boldsymbol{\pi}^+, \mathbf{1}/\mathbf{2})$ . For the two-sample case,  $|\pi_+ - 1/2| < .20$  has been proposed in § 1.7 as a reasonable criterion of equivalence. This motivates to choose  $\tilde{\varepsilon}$  as the radius of the smallest sphere around  $\mathbf{1}/\mathbf{2}$  in  $(k(k - 1))$ -dimensional space which contains all points with every coordinate equal to  $1/2 \pm .20$ . This leads to the specification  $\tilde{\varepsilon}^2 = (k(k - 1))/2 \cdot .20^2$  so that we use  $\tilde{\varepsilon}^2 = 6 \cdot .04 = .24$  in the present example.

With  $\alpha = .05$ ,  $\tilde{\varepsilon}^2 = .24$  and  $\hat{\sigma}_N [d^2(\mathbf{W}^+, \mathbf{1}/\mathbf{2})]/\sqrt{N} = .045094$ , the critical upper bound which the observed value of  $d^2(\mathbf{W}^+, \mathbf{1}/\mathbf{2})$  has to be compared to according to (7.33), is  $.24 - 1.644854 \cdot .045094 = .165827$ . Hence, the raw data shown in Table 7.1 fall in the rejection region of our nonparametric test at level  $\alpha = .05$  for equivalence of the distributions from which the  $k = 4$  samples under analysis are taken.

---

## *Equivalence tests for multivariate data*

All testing procedures considered in this chapter are primarily tailored for data taken from a  $k$ -variate normal distribution, with  $k$  denoting a fixed natural number  $\geq 2$ . The methods discussed in § 8.1 can be viewed as repeated measurement analogues of the equivalence tests obtained in §§ 7.2–3 for the ANOVA one-way layout. The remaining sections are devoted to multivariate analogues of the two-sample  $t$ -tests for equivalence of § 6.1 and the double one-sided testing procedure, respectively.

---

### 8.1 Equivalence tests for several dependent samples from normal distributions

#### 8.1.1 Generalizing the paired $t$ -test for equivalence by means of the $T^2$ -statistic

Regarding the basic structure of the data under analysis, the setting considered in this subsection differs from that of the paired  $t$ -test [ $\rightarrow$  § 5.3] only by the fact that the number  $k$  of conditions [treatments, timepoints at which some quantity is repeatedly measured, etc.] under which the observations are taken from each subject in a sample of size  $n$ , is no longer restricted to 2. The pattern of intraindividual correlations is left completely unspecified, but all distributions involved have to exhibit Gaussian form. More precisely speaking, we start from the assumption that the data set consists of  $n$  mutually independent vectors  $(X_{11}, \dots, X_{k1}), \dots, (X_{1n}, \dots, X_{kn})$  such that

$$(X_{1i}, \dots, X_{ki}) \sim \mathcal{N}((\mu_1, \dots, \mu_k), \Sigma), \quad \forall i = 1, \dots, n, \quad (8.1)$$

where

$$\Sigma \equiv \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1k} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \dots & \sigma_k^2 \end{pmatrix} \quad (8.2)$$

denotes some positive definite (symmetric) matrix of order  $k \times k$ .

A natural way of generalizing the distance measure  $\delta^2/\sigma_D^2 = (\mu_1 - \mu_2)^2/(\sigma_1^2 + \sigma_2^2 - 2\sigma_{12})$  underlying the paired  $t$ -test with symmetric specification of the hypothetical equivalence range, to the  $k$ -variate setting (8.1) is to use the Mahalanobis distance of a set of  $k - 1$  contrasts  $(\delta_1, \dots, \delta_{k-1})$  in the  $\mu$ 's from the origin as the parametric function of interest. Of course, the covariance matrix  $\Sigma_D$ , say, with respect to which this distance has to be taken, is that of the corresponding contrasts in the components of the random vectors primarily observed. For definiteness, we choose these contrasts as pairwise differences between successive components of the respective vectors, defining

$$\delta_j = \mu_{j+1} - \mu_j, \quad j = 1, \dots, k - 1. \quad (8.3)$$

Then, the entries in  $\Sigma_D$ , say  $\sigma_{jl}^D$ , are given by

$$\sigma_{jl}^D = \sigma_{jl} + \sigma_{j+1,l+1} - \sigma_{j,l+1} - \sigma_{j+1,l}, \quad 1 \leq j, l \leq k - 1. \quad (8.4)$$

The equivalence hypothesis we want to establish specifies that the true vector  $\boldsymbol{\delta} = (\delta_1, \dots, \delta_{k-1})$  of mean differences between “adjacent” components of the  $\mathbf{X}_1, \dots, \mathbf{X}_n$  is sufficiently close to  $\mathbf{0}$ , with distances between pairs  $(\mathbf{d}_1, \mathbf{d}_2)$  of points in  $\mathbb{R}^{k-1}$  measured in terms of  $(\mathbf{d}_1 - \mathbf{d}_2)\Sigma_D^{-1}(\mathbf{d}_1 - \mathbf{d}_2)'$ . Accordingly, the equivalence testing problem we are now interested in reads

$$H : \boldsymbol{\delta}\Sigma_D^{-1}\boldsymbol{\delta}' \geq \varepsilon^2 \text{ versus } K : \boldsymbol{\delta}\Sigma_D^{-1}\boldsymbol{\delta}' < \varepsilon^2. \quad (8.5)$$

It is important to note that for any fixed  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_k) \in \mathbb{R}^k$ , replacing the  $\delta_j$  as defined in (8.3) by any other vector  $\tilde{\boldsymbol{\delta}}$  of contrasts does not change the value of  $\boldsymbol{\delta}\Sigma_D^{-1}\boldsymbol{\delta}'$ . Actually, if we denote by  $\mathbf{C}$  the matrix by which  $\boldsymbol{\mu}'$  is premultiplied in generating the vector  $\boldsymbol{\delta}'$  with components (8.3), and let  $\tilde{\mathbf{C}}$  be any other  $(k - 1) \times k$ -matrix of rank  $k - 1$  with  $\tilde{\mathbf{C}}(1, 1, \dots, 1)' = \mathbf{0}$ , elementary facts from linear algebra imply that there holds  $\tilde{\mathbf{C}}' = \mathbf{C}'\mathbf{A}$  for some nonsingular  $(k - 1) \times (k - 1)$ -matrix  $\mathbf{A}$ . In view of this, both  $\tilde{\boldsymbol{\delta}}$  and  $\boldsymbol{\delta}$ , and the random vectors  $\tilde{\mathbf{D}} = (\tilde{D}_1, \dots, \tilde{D}_{k-1})$  and  $\mathbf{D} = (D_1, \dots, D_{k-1})$  whose expectations are given by the former, are related to each other through the transformation  $\mathbf{V} = \mathbf{U}\mathbf{A}$  which induces the relationship  $\Sigma_{\tilde{D}} = \mathbf{A}'\Sigma_D\mathbf{A}$  between the covariance matrices of  $\tilde{\mathbf{D}}$  and  $\mathbf{D}$ . Making use of the identities  $\tilde{\boldsymbol{\delta}} = \boldsymbol{\delta}\mathbf{A}$ ,  $\Sigma_{\tilde{D}} = \mathbf{A}'\Sigma_D\mathbf{A}$ , only a little bit more algebra is required to show that there holds  $\tilde{\boldsymbol{\delta}}\Sigma_{\tilde{D}}^{-1}\tilde{\boldsymbol{\delta}}' = \boldsymbol{\delta}\Sigma_D^{-1}\boldsymbol{\delta}'$  indeed.

An optimal test for the equivalence problem (8.5) can be based on the empirical counterpart of the theoretical Mahalanobis distance  $\boldsymbol{\delta}\Sigma_D^{-1}\boldsymbol{\delta}'$  as obtained by replacing  $\boldsymbol{\delta}$  and  $\Sigma_D$  with

$$\bar{\mathbf{D}} \equiv n^{-1} \sum_{i=1}^n (D_{1i}, \dots, D_{k-1,i}) \quad (8.6)$$

and

$$\begin{aligned} \mathbf{S}_D \equiv (n-1)^{-1} \sum_{i=1}^n (D_{1i} - \bar{D}_1, \dots, D_{k-1,i} - \bar{D}_{k-1})' \\ (D_{1i} - \bar{D}_1, \dots, D_{k-1,i} - \bar{D}_{k-1}), \end{aligned} \quad (8.7)$$

respectively. A well-known result from the classical theory of parametric multivariate inference (see, e.g., Anderson, 1984, § 5.4) states that the distribution of

$$T^2 \equiv n \bar{\mathbf{D}} \mathbf{S}_D^{-1} \tilde{\mathbf{D}}' \quad (8.8)$$

depends on  $\boldsymbol{\delta}$  and  $\Sigma_D$  only through the actual value, say  $\tau^2$ , of  $\boldsymbol{\delta} \Sigma_D^{-1} \boldsymbol{\delta}'$ , in that we have

$$T^2 (n-k+1)/((n-1)(k-1)) \stackrel{d}{=} \mathcal{F}_{k-1, n-k+1}(n\tau^2), \quad (8.9)$$

where  $\mathcal{F}_{\nu_1, \nu_2}(\psi^2)$  stands for a random variable following an  $F$ -distribution with  $\nu_1, \nu_2$  degrees of freedom and noncentrality-parameter  $\psi^2$ . Consequently, an exact level- $\alpha$  test for (8.5) consists of checking the observed data for inclusion in the critical region

$$\{T^2 < ((n-1)(k-1)/(n-k+1)) F_{k-1, n-k+1; \alpha}(n\varepsilon^2)\} \quad (8.10)$$

with  $F_{\nu_1, \nu_2; \alpha}(\psi^2)$  denoting the lower  $100\alpha$  percentage point of the distribution of  $\mathcal{F}_{\nu_1, \nu_2}(\psi^2)$  for any  $(\nu_1, \nu_2) \in \mathbb{N}^2$  and  $\psi^2 \geq 0$ .

Table 8.1 shows the result of computing the critical upper bounds to the statistic  $T^2$  of (8.8) for three choices of the equivalence margin, five different values of the number  $k$  of conditions under comparison, and the same range of sample sizes  $n$  which is covered by Table 5.11 for the paired-sample setting. Omission of the bivariate case from the current table has been done with a view to the fact that for  $k = 2$ , we simply would have to tabulate the squares of the respective entries in the former.

The power function of the  $T^2$ -test for equivalence of  $k$  normal distributions to be analyzed on the basis of dependent samples can likewise be evaluated exactly by means of easily accessible computational tools. Actually, denoting the rejection probability of the test based on (8.10) under any parameter configuration  $(\boldsymbol{\delta}, \Sigma_D)$  with  $\boldsymbol{\delta} \Sigma_D^{-1} \boldsymbol{\delta}' = \tau^2 \in [0, \infty)$  by  $\beta(\tau^2)$ , we can write

$$\beta(\tau^2) = P[\mathcal{F}_{k-1, n-k+1}(n\tau^2) \leq F_{k-1, n-k+1; \alpha}(n\varepsilon^2)]. \quad (8.11)$$

Against null alternatives specifying that the true vector  $\boldsymbol{\delta}$  of mean differences between the  $k$  treatments is zero, the power of the test turns out to be as shown in Table 8.2 which is otherwise organized in the same way as the preceding table. According to intuition, one would expect that, given everything else, the chance of establishing equivalence of the  $k$  marginal distributions when they are in fact identical in terms of their means, decreases with the number of treatment conditions under comparison. The exact numerical results obtained by means of formula (8.11) clearly confirm the correctness of this assumption.

Table 8.1 Critical upper bound  $T_{.05; n-1}^{2;k-1}(\varepsilon)$  to the  $T^2$ -statistic to be used when testing for equivalence of  $k$  normal distributions with dependent samples of size  $n \in \{10, 20, \dots, 100\}$ , for  $\alpha = 5\%$  and three different specifications of the equivalence margin  $\varepsilon$  to the root of the Mahalanobis distance of the vector of mean differences from the origin.

$n$	$\varepsilon$	$k =$				
		3	4	5	6	8
10	0.25	0.15846	0.53337	1.13575	2.01545	5.27403
10	0.50	0.38473	0.95198	1.74680	2.83976	6.73052
10	1.00	3.05419	4.17579	5.59033	7.44170	13.80923
20	0.25	0.20122	0.58067	1.10236	1.74900	3.42816
20	0.50	0.96156	1.63675	2.40659	3.28112	5.40902
20	1.00	8.43408	9.63571	10.96366	12.43980	15.95362
30	0.25	0.26599	0.68338	1.21517	1.83910	3.33994
30	0.50	1.87865	2.62803	3.43654	4.30964	6.27486
30	1.00	14.74092	16.01615	17.37638	18.83065	22.06458
40	0.25	0.35166	0.81262	1.36781	1.99705	3.45082
40	0.50	3.01525	3.81160	4.65233	5.54054	7.47309
40	1.00	21.56257	22.89350	24.28965	25.75604	28.92250
50	0.25	0.46040	0.96453	1.54568	2.18794	3.63237
50	0.50	4.29730	5.12787	5.99437	6.89877	8.82984
50	1.00	28.72676	30.10126	31.52912	33.01357	36.16660
60	0.25	0.59379	1.13775	1.74411	2.40155	3.85127
60	0.50	5.68531	6.54262	7.43024	8.34955	10.28915
60	1.00	36.14076	37.55055	39.00570	40.50849	43.66685
70	0.25	0.75213	1.33113	1.96069	2.63342	4.09472
70	0.50	7.15530	8.03435	8.93969	9.87233	11.82388
70	1.00	43.74764	45.18680	46.66554	48.18555	51.35659
80	0.25	0.93439	1.54332	2.19369	2.88098	4.35653
80	0.50	8.69127	9.58850	10.50898	11.45348	13.41794
80	1.00	51.50940	52.97355	54.47287	56.00867	59.19524
90	0.25	1.13865	1.77283	2.44165	3.14247	4.63317
90	0.50	10.28185	11.19460	12.12821	13.08332	15.06066
90	1.00	59.39914	60.88491	62.40237	63.95258	67.15558
100	0.25	1.36262	2.01813	2.70326	3.41649	4.92238
100	0.50	11.91862	12.84485	13.79002	14.75464	16.74442
100	1.00	67.39698	68.90172	70.43534	71.99868	75.21811

Table 8.2 *Power against the alternative  $\mu_1 = \dots = \mu_k$  attained in the  $T^2$ -test (8.10) when using the critical constants shown in Table 8.1 for the  $k$ -sample setting with dependent observations.*

$n$	$\varepsilon$	$k =$				
		3	4	5	6	8
10	0.25	.06743	.06604	.06455	.06310	.06020
10	0.50	.15417	.13883	.12626	.11562	.09729
10	1.00	.68924	.58339	.49550	.41991	.29101
20	0.25	.09046	.08695	.08373	.08093	.07634
20	0.50	.35874	.30494	.26842	.24119	.20192
20	1.00	.96334	.93324	.89735	.85710	.76799
30	0.25	.11999	.11286	.10690	.10202	.09450
30	0.50	.58471	.50349	.44562	.40136	.33666
30	1.00	.99683	.99289	.98688	.97853	.95422
40	0.25	.15680	.14413	.13437	.12672	.11536
40	0.50	.75706	.67871	.61614	.56469	.48430
40	1.00	.99977	.99939	.99870	.99758	.99350
50	0.25	.20104	.18081	.16617	.15505	.13902
50	0.50	.86702	.80707	.75332	.70532	.62377
50	1.00	.99998	.99995	.99989	.99977	.99926
60	0.25	.25204	.22249	.20207	.18691	.16546
60	0.50	.93060	.89061	.85081	.81230	.74088
60	1.00	1.0000	1.0000	.99999	.99998	.99993
70	0.25	.30831	.26842	.24164	.22201	.19456
70	0.50	.96508	.94072	.91412	.88641	.83042
70	1.00	1.0000	1.0000	1.0000	1.0000	.99999
80	0.25	.36782	.31758	.28424	.25994	.22610
80	0.50	.98293	.96903	.95255	.93418	.89388
80	1.00	1.0000	1.0000	1.0000	1.0000	1.0000
90	0.25	.42842	.36878	.32912	.30017	.25979
90	0.50	.99186	.98430	.97467	.96324	.93616
90	1.00	1.0000	1.0000	1.0000	1.0000	1.0000
100	0.25	.48820	.42082	.37547	.34211	.29531
100	0.50	.99619	.99224	.98687	.98011	.96290
100	1.00	1.0000	1.0000	1.0000	1.0000	1.0000

Obviously, the argument behind the fact that  $\tilde{\boldsymbol{\delta}} = \boldsymbol{\delta} \mathbf{A} \Rightarrow \tilde{\boldsymbol{\delta}} \boldsymbol{\Sigma}_{\tilde{\mathbf{D}}}^{-1} \tilde{\boldsymbol{\delta}}' = \boldsymbol{\delta} \boldsymbol{\Sigma}_D^{-1} \boldsymbol{\delta}'$  for any nonsingular  $(k - 1) \times (k - 1)$ -matrix  $\mathbf{A}$ , likewise applies when  $\boldsymbol{\delta}$  and  $\boldsymbol{\Sigma}_D^{-1}$  are replaced with their empirical counterparts  $\tilde{\mathbf{D}}$  and  $\mathbf{S}_D^{-1}$ . Thus, the statistic  $T^2$  of (8.8) is left invariant under arbitrary linear one-to-one transformations of the sample contrast-vectors  $\mathbf{D}_i$ , and can even be shown (see, e.g., Lehmann, 1986, § 8.2) to be maximal invariant with respect to this group

of transformations. Furthermore, in view of (8.10) and the maximal invariance of  $T^2$ , the same arguments which lead in § 7.2 to the conclusion that the ANOVA  $F$ -test for equivalence of  $k$  normal distributions from which independent samples are given, can be used to infer that the test given by (8.10) maximizes the power uniformly among all invariant level- $\alpha$  tests for (8.5).

*Example 8.1*

In a study of the effects of deprivation on voluntary alcohol intake in alcohol-preferring rat lines, one of the objectives was to show that in some of the lines included in the experiments, the total intake of ethanol [g/kg/day] after the deprivation phase remained roughly constant over time. Measurements were taken on  $k = 4$  consecutive days after the animals had been relocated to a setting providing free access to alcohol drinking solutions. Table 8.3 shows the arithmetic means and the empirical covariance matrix calculated from  $n = 20$  data vectors.

Table 8.3 *Summary statistics for a sample of  $n = 20$  repeated measurements of total ethanol intake [g/kg/day] taken at  $k = 4$  consecutive days in rats of an alcohol-preferring line after deprivation.*

	Day $j =$			
	1	2	3	4
$\bar{X}_j$	5.0967989	5.0942015	5.2835529	5.1851779
$S_{1j}^x$	2.095013962	1.521976271	0.549312992	-0.260377815
$S_{2j}^x$	1.521976271	1.710873546	0.881500432	0.541358142
$S_{3j}^x$	0.549312992	0.881500432	1.106519118	1.230386226
$S_{4j}^x$	-0.260377815	0.541358142	1.230386226	2.313748226

The elements of the estimated contrast vector  $\bar{\mathbf{D}}$  and the empirical covariance matrix  $\mathbf{S}_D$  of  $\mathbf{D}$  are easily obtained by replacing the population parameters appearing in Equations (8.3) and (8.4) with the respective entries in Table 8.3, leading to the values displayed in Table 8.4. Finally, as soon as  $\bar{\mathbf{D}}$  and  $\mathbf{S}_D$  have been calculated, evaluating the quadratic form  $\bar{\mathbf{D}}\mathbf{S}_D^{-1}\tilde{\mathbf{D}}'$  requires to write down just a single assignment statement in any matrix-oriented programming environment like SAS/IML or R. For the test statistic itself, we obtain  $T^2 = 1.7386013$  which is slightly larger than the critical upper bound we read from Table 8.1 for  $n = 20$ ,  $\varepsilon = .50$  and  $k = 4$ . Thus, specifying the equivalence margin to the squared Mahalanobis distance of the true value of  $\delta$  from the origin to be  $.50^2$ , we have to accept the null hypothesis of nonequivalence in the present case. Keeping in mind, that, according to the numerical

Table 8.4 Estimated mean differences and variances-covariances of intraindividual differences between successive time points obtained with the data shown in Table 8.3.

	$j =$		
	1	2	3
$\bar{D}_j$	-0.0025970	0.1893514	-0.0983750
$S_{1j}^D$	0.7619350	0.1432902	0.4695485
$S_{2j}^D$	0.1432902	1.0543918	0.4640094
$S_{3j}^D$	0.4695485	0.4640094	0.9594949

material shown in Table 8.2, the maximum power of the test is only slightly larger than 30% under the conditions of this example, it is not surprising that the decision eventually to be taken turns out negative.

### 8.1.2 A many-one equivalence test for dependent samples based on Euclidean distances

The distance measure used in the preceding subsection as a basis for constructing an equivalence test for several dependent samples from distributions of Gaussian form might be criticized for leading to equivalence regions whose geometrical shape strongly depends on the underlying covariance structure. The potential difficulties entailed with this fact are best clarified by visualizing the theoretical equivalence region corresponding to the alternative hypothesis  $K$  of (8.5) in the bivariate case with identical variances  $\sigma_{11}^D = \sigma_{22}^D$  of both components of the vector of intraindividual differences, for various values of the theoretical correlation coefficient  $\varrho^D$ . As can be seen from Figure 8.1, it is a circle around zero for  $\varrho^D = 0$ , but an ellipse whose major diameter increases to infinity as  $\varrho^D \rightarrow 1$  whereas in the direction of the minor axis, the diameter remains bounded above by  $2\varepsilon$  for any value of  $\varrho^D$ .

In this subsection, we keep the shape of the theoretical equivalence region independent of the covariance structure of  $\mathbf{D} = (D_1, \dots, D_{k-1})$  specifying under the alternative hypothesis that the vector  $(\delta_1/\sigma_1^D, \dots, \delta_{k-1}/\sigma_{k-1}^D)$  of standardized expected values has Euclidean distance  $< \varepsilon$  from the origin for some fixed  $\varepsilon > 0$ . Of course, this means, that the hypothetical equivalence region is a sphere of radius  $\varepsilon$  around  $\mathbf{0}$  in  $(k-1)$ -dimensional space. Given the expected value  $\boldsymbol{\mu} = (\mu_1, \dots, \mu_k)$  of the primarily observed vector  $\mathbf{X} = (X_1, \dots, X_k)$ , the value of  $\sum_{j=1}^{k-1} (\delta_j/\sigma_j^D)^2$  clearly depends on the choice of the contrast matrix used for reducing  $\mathbf{X}$  to the  $(k-1)$ -dimensional vector  $\mathbf{D}$  of intraindividual differences. Hence, the testing procedure to be derived in the sequel is tailored for settings where the contrast of interest is uniquely determined from the context. The prototype of such a setting occurs in any

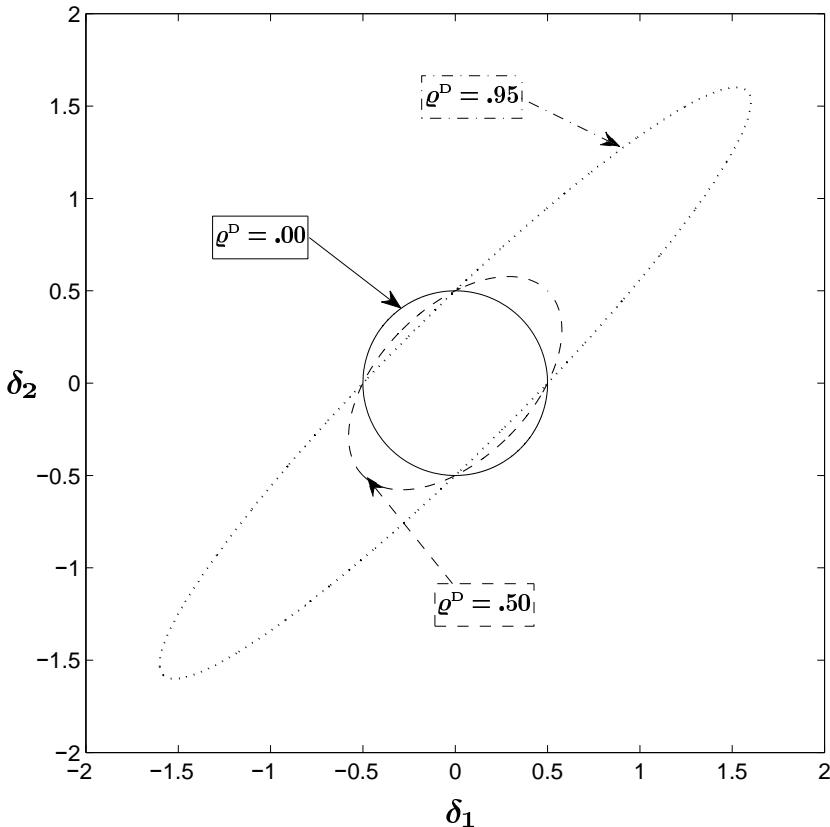


Figure 8.1 *Shape of the equivalence region in terms of Mahalanobis distance in the bivariate case with equal marginal variances and  $\varepsilon^2 = .25$ , for three different values of the correlation coefficient  $\rho^D$  between the intraindividual differences.*

trial involving  $k \geq 2$  treatments of which one (say that indexed by  $j = 1$ ) serves as a control for all others so that the only set of contrasts to be taken in consideration is given by

$$\delta_j = \mu_{j+1} - \mu_1, \quad j = 1, \dots, k-1. \quad (8.12)$$

For the corresponding transforms of observed variables, we likewise use the same symbols as in the preceding subsection redefining  $D_j$  and  $D_{ji}$  as

$$D_j = X_{j+1} - X_1, \quad j = 1, \dots, k-1, \quad (8.13)$$

and

$$D_{ji} = X_{j+1,i} - X_{1i}, \quad j = 1, \dots, k-1, \quad i = 1, \dots, n, \quad (8.14)$$

respectively.

Another price (additionally to having to dispense with invariance against relabeling the treatments under comparison) we will have to pay for replacing Mahalanobis with ordinary Euclidean distance is that the problem of testing

$$H : \sum_{j=1}^{k-1} (\delta_j / \sigma_j^D)^2 \geq \varepsilon^2 \quad \text{versus} \quad K : \sum_{j=1}^{k-1} (\delta_j / \sigma_j^D)^2 < \varepsilon^2 \quad (8.15)$$

can only be solved by means of asymptotic methods. A suitable starting point for the asymptotic construction is obtained by introducing the plug-in estimator, say  $\hat{Q}_D$ , of the squared length of the vector  $(\delta_1 / \sigma_1^D, \dots, \delta_{k-1} / \sigma_{k-1}^D)$  of standardized intraindividual mean differences. Defining

$$g(\delta_1, \sigma_1^D, \dots, \delta_{k-1}, \sigma_{k-1}^D) \equiv \sum_{j=1}^{k-1} (\delta_j / \sigma_j^D)^2, \quad (8.16)$$

$\hat{Q}_D$  admits the representation

$$\hat{Q}_D = g(\hat{D}_1, S_1^D, \dots, \hat{D}_{k-1}, S_{k-1}^D), \quad (8.17)$$

where

$$\hat{D}_j = n^{-1} \sum_{i=1}^n D_{ji}, \quad S_j^D = \left[ (n-1)^{-1} \sum_{i=1}^n (D_{ji} - \hat{D}_j)^2 \right]^{1/2} \quad (8.18)$$

for each  $j = 1, \dots, k-1$ .

Now the major task is to derive the asymptotic distribution of the statistic  $\hat{Q}_D$  by means of the multivariate  $\delta$ -method as described, e.g., in Bishop et al. (1975, §14.6.3). In order to make the details precise, it is convenient to simplify notation a bit by setting  $k-1 = p$  and  $\sigma_{jl}^D = \tau_{jl}$  for all  $1 \leq j, l \leq p$ . Then, well-known results from the distribution theory for sample means and variances/covariances under the multivariate normal model (see in particular Anderson, 1984, Theorem 3.4.4) imply that

$$\sqrt{n}((\hat{D}_1, S_1^D, \dots, \hat{D}_{k-1}, S_{k-1}^D) - (\delta_1, \sigma_1^D, \dots, \delta_{k-1}, \sigma_{k-1}^D)) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \Sigma^*) \quad (8.19)$$

as  $n \rightarrow \infty$ , with

$$\Sigma^* = \begin{pmatrix} \tau_1^2 & 0 & \tau_{12} & 0 & \dots & \tau_{1p} & 0 \\ 0 & \tau_1^2/2 & 0 & \tau_{12}^2/(2\tau_1\tau_2) & \dots & 0 & \tau_{1p}^2/(2\tau_1\tau_p) \\ \tau_{12} & 0 & \tau_2^2 & 0 & \dots & \tau_{2p} & 0 \\ 0 & \tau_{12}^2/(2\tau_1\tau_2) & 0 & \tau_2^2/2 & \dots & 0 & \tau_{2p}^2/(2\tau_2\tau_p) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \tau_{1p} & 0 & \tau_{2p} & 0 & \dots & \tau_p^2 & 0 \\ 0 & \tau_{1p}^2/(2\tau_1\tau_p) & 0 & \tau_{2p}^2/(2\tau_2\tau_p) & \dots & 0 & \tau_p^2/2 \end{pmatrix}. \quad (8.20)$$

Furthermore, the function  $g(\cdot)$  obviously has a differential anywhere in  $\mathbb{R} \times \mathbb{R}_+ \times \dots \times \mathbb{R} \times \mathbb{R}_+$ , and the components of the gradient vector at any point  $(\delta_1, \tau_1, \dots, \delta_p, \tau_p)$  in this space are easily calculated to be

$$\frac{\partial g}{\partial \delta_j} = \frac{2\delta_j}{\tau_j^2}, \quad \frac{\partial g}{\partial \tau_j} = -\frac{2\delta_j^2}{\tau_j^3}, \quad j = 1, \dots, p. \quad (8.21)$$

In order to determine the asymptotic variance, say  $\sigma_a^2[\hat{Q}_D]$ , of the test statistic  $\hat{Q}_D$  multiplied by  $\sqrt{n}$  as a normalizing factor under an arbitrarily selected parameter configuration, we have to evaluate the quadratic form  $\mathbf{x} \mapsto \mathbf{x}\Sigma^*\mathbf{x}'$  at  $\mathbf{x} = \left( \frac{\partial g}{\partial \delta_1}, \frac{\partial g}{\partial \tau_1}, \dots, \frac{\partial g}{\partial \delta_p}, \frac{\partial g}{\partial \tau_p} \right)$ . With the explicit expressions appearing in (8.20) and (8.21), this gives after suitable rearrangements of terms

$$\begin{aligned} \sigma_a^2[\hat{Q}_D] &= \sum_{j=1}^p \sum_{l=1}^p 2\delta_j \delta_l (\tau_{jl}/(\tau_j^2 \tau_l^2)) [2 + \delta_j \delta_l \tau_{jl}/(\tau_j^2 \tau_l^2)] \\ &\equiv \sum_{j=1}^{k-1} \sum_{l=1}^{k-1} 2\delta_j \delta_l (\sigma_{jl}^D / (\sigma_{jj}^D \sigma_{ll}^D)) [2 + \delta_j \delta_l \sigma_{jl}^D / (\sigma_{jj}^D \sigma_{ll}^D)]. \end{aligned} \quad (8.22)$$

The reason why this formula is worth being presented in the present context is that the theorem underlying the  $\delta$ -method implies that there holds

$$\sqrt{n}(\hat{Q}_D - Q_D)/\sigma_a[\hat{Q}_D] \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1), \quad (8.23)$$

with  $Q_D$  standing for the parametric function (8.16) of interest. Moreover, the asymptotic variance  $\sigma_a^2[\hat{Q}_D]$  of the normalized test statistic can be consistently estimated by replacing  $\delta_j$  and  $\sigma_{jl}^D$  with  $\bar{D}_j$  and  $S_{jl}^D \equiv$  and  $(n-1)^{-1} \sum_{i=1}^n (D_{ji} - \bar{D}_j)(D_{li} - \bar{D}_l)$ , respectively, in (8.22) for all  $1 \leq j, l \leq k-1$ . The explicit formula for this plug-in estimator reads

$$\hat{\sigma}_a^2[\hat{Q}_D] = \sum_{j=1}^{k-1} \sum_{l=1}^{k-1} 2\bar{D}_j \bar{D}_l (S_{jl}^D / (S_{jj}^D S_{ll}^D)) [2 + \bar{D}_j \bar{D}_l S_{jl}^D / (S_{jj}^D S_{ll}^D)], \quad (8.24)$$

analogously to (8.22). Since the alternative hypothesis under consideration specifies that the true value of  $Q_D$  is sufficiently small, the one-sided rejection region to be used according to (2.6) in an asymptotically valid test for noninferiority has to be replaced with a region containing all points in the sample space for which the normalized test statistic centered about the equivalence margin  $\varepsilon^2$  falls below the lower  $\alpha$  standard normal quantile. Thus, a large-sample solution to the testing problem (8.15) rejects if and only if it turns out that

$$\sqrt{n}(\hat{Q}_D - \varepsilon^2)/\hat{\sigma}_a[\hat{Q}_D] < u_\alpha. \quad (8.25)$$

*Example 8.1 (continued)*

For a reanalysis of the data set underlying Table 8.3 by means of the test given by (8.25), we reduce all vectors primarily observed to intraindividual changes against the value measured on Day 1 [cf. (8.14)]. The means and empirical variances/covariances obtained with these recalculated difference vectors are shown in Table 8.5. With these values, the test statistic  $\hat{Q}_D$  is easily computed

Table 8.5 *Estimated mean differences and variances-covariances of intraindividual changes against the value measured on Day 1 as baseline for the data shown in Table 8.3.*

	$j =$		
	1	2	3
$\bar{D}_j$	-0.002597	0.186754	0.088379
$S_{1j}^D$	0.7619350	0.9052251	1.3747736
$S_{2j}^D$	0.9052251	2.1029071	3.0364650
$S_{3j}^D$	1.3747736	3.0364650	4.9295178

to be  $\hat{Q}_D = 0.0181785$ . For the estimated standard error of  $\sqrt{n}\hat{Q}_D$ , we find through plugging-in into formula (8.24) that  $\hat{\sigma}_a[\hat{Q}_D] = 0.3304173$ . Using the same specification of the equivalence margin as before, the left-hand side of (8.25) turns out to be  $-3.13766$  which is distinctly smaller than  $u_{.05} = -1.64485$ . So, this time the decision is in favor of equivalence.

As is the case for virtually all asymptotic testing procedures, a question of considerable importance is about its (maximum) rejection probability under the null hypothesis when used with samples of small or moderate size. In order to provide sufficient insight into this issue, another set of simulation experiments has been performed covering fairly different correlation structures and distributions of the total squared distance  $\varepsilon^2$  of  $(\delta_1/\sigma_1^D, \dots, \delta_{k-1}/\sigma_{k-1}^D)$  from the origin over the components of this vector. (Varying the parameter configuration by looking at different specifications of the  $\delta_j$  and  $\sigma_j^D$  leading to the same pattern of standardized expected differences is unnecessary since both the hypotheses and the testing procedure are obviously scale invariant). Table 8.7 shows the results of such a simulation study for dimension  $p = 4$  of the vector of intraindividual observations to be assessed and the same specifications of the equivalence margin and the nominal significance level made in the example, and the four correlation matrices specified in Table 8.6.

A glance over the rejection probabilities obtained with samples generated under the null hypothesis clearly reveals that the accuracy of the large-sample approximation upon which the test (8.25) relies, is acceptable, if at all, only for an equi-correlation structure of  $(D_1, \dots, D_{k-1})$  with low common value of the off-diagonal entries. Again [recall §§ 5.2.2, 5.2.3, 6.6.3, 6.6.6], a straightforward way of protecting oneself from grossly exceeding the target significance level is to reduce the nominal level as far as necessary. Tentatively, such a reduced nominal level  $\alpha^*$  can be determined through simulation under the parameter configuration which, according to the results presented in Table 8.7, came out least favorable (in the sense of yielding the largest rejection probability at the boundary of the null hypothesis) for the respective sample size. In terms of  $\alpha - \alpha^*$ , the extent of the required corrections is huge: For the sample sizes covered by Table 8.7, we obtained the reduced levels  $2.5 \cdot 10^{-7}$  [ $\rightarrow n = 25$ ],  $10^{-6}$  [ $\rightarrow n = 50$ ],  $10^{-3}$  [ $\rightarrow n = 100$ ], and  $.014$  [ $\rightarrow n = 400$ ], respectively. The power results shown in Table 8.8 hold for the test when performed at these corrected nominal levels and have also been computed by simulation. As can be expected in view of the behavior of the power function on the common boundary of  $H$  and  $K$ , the rejection probability strongly depends on the correlation structure also under null alternatives. Reducing the nominal significance level as much as necessary for keeping the maximum rejection probability under the null hypothesis below 5% also affects the qualitative result of the test in the above example: For  $n = 20$ , a slightly less extreme choice of  $\alpha^*$  (as compared to  $n = 25$ ) turns out to work, namely  $\alpha^* = 2.5 \cdot 10^{-6}$ . Accordingly, the observed value of the standardized test statistic has actually to be compared to  $-4.56479$  rather than  $-1.64485$ . Since we obtained  $\sqrt{n}(\hat{Q}_D - \varepsilon^2)/\hat{\sigma}_a[\hat{Q}_D] = -3.13766$ , the level-corrected test for equivalence in the sense of  $\sum_{j=1}^{k-1} (\delta_j/\sigma_j^D)^2 < \varepsilon^2$  likewise fails to allow rejection of nonequivalence, as did the exact test for equivalence in terms of Mahalanobis distance.

Table 8.6 Correlation matrices used for generating the data in the simulation experiments reported in Table 8.7.

Correlation Matrix #		$\varrho_{12}^{(r)}$	$\varrho_{13}^{(r)}$	$\varrho_{14}^{(r)}$	$\varrho_{23}^{(r)}$	$\varrho_{24}^{(r)}$	$\varrho_{34}^{(r)}$
$(r)$							
1		.25	.25	.25	.25	.25	.25
2		.90	.90	.90	.90	.90	.90
3		.25	.50	.95	.05	.25	.65
4		-.65	.35	.45	.35	.35	.95

Table 8.7 Simulated rejection probabilities of the test (8.25) at the boundary of the hypotheses (8.15) for the correlation structures (1)–(4) of Table 8.6 and the following patterns of contributions of the components of  $(\delta_1/\tau_1, \dots, \delta_p/\tau_p)$  to the assumed value  $\varepsilon^2 = .25$  of  $Q_D$ : (a)  $(\delta_j/\tau_j)^2 = \varepsilon^2/p \forall j = 1, \dots, p$ ; (b)  $(\delta_1/\tau_1)^2 = \varepsilon^2, (\delta_j/\tau_j)^2 = 0 \forall j = 2, \dots, p$ ; (c)  $(\delta_j/\tau_j)^2 = j\varepsilon^2/(p(p+1)/2) \forall j = 1, \dots, p$  [ $p = 4, \varepsilon = .50, \alpha = .05$ ; 100,000 replications per Monte Carlo experiment].

$n$	Correlation Matrix # (r)	Pattern of Standardized $\delta$ 's		
		(a)	(b)	(c)
25	1	.04977	.02211	.04759
"	2	.14779	.02957	.14378
"	3	.07452	.02133	.07411
"	4	.06394	.03295	.06983
50	1	.06802	.03490	.06496
"	2	.14659	.05104	.14670
"	3	.09218	.04576	.09385
"	4	.08176	.05369	.08441
100	1	.06667	.04424	.06346
"	2	.11849	.07822	.11954
"	3	.08691	.06441	.08709
"	4	.07789	.06286	.07891
400	1	.05873	.05071	.05731
"	2	.07982	.09123	.08087
"	3	.06778	.06758	.06879
"	4	.06663	.05921	.06569

Table 8.8 Simulated power of the level-corrected test (8.25) against the alternatives obtained by setting  $\delta_j = 0 \forall j = 1, \dots, p$  in the constellations covered by Table 8.7.

$n$	Correlation Matrix # (r)			
	1	2	3	4
25	0.01745	0.11521	0.03730	0.04463
50	0.16601	0.25811	0.19388	0.17744
100	0.84795	0.64818	0.75595	0.72903
400	1.00000	0.99954	1.00000	1.00000

### 8.1.3 Testing for equivalence with dependent samples and an indifference zone of rectangular shape

Another reasonable way of specifying under the alternative hypothesis to be established, an equivalence region whose shape does not depend on the pattern of associations between the components of the primary data vectors, is to require that for all pairs  $(j, l)$  with  $1 \leq j < l \leq k$ , there holds  $|\mu_j - \mu_l| < \varepsilon_{jl}$  for suitable equivalence margins  $\varepsilon_{jl} > 0$ . The latter may be chosen as fixed constants involving no other unknown population parameters, or as multiples of the theoretical standard deviations, say  $\sigma_{(j,l)}$ , of the differences  $X_j - X_l$  between the corresponding components of a random vector from the  $k$ -dimensional distribution under assessment. In both cases, the indifference zone making up the theoretical equivalence region, corresponds to a rectangle in the  $(k(k-1)/2)$ -dimensional parameter space of all pairwise differences of expected values. We start with a consideration of the case that interest is on establishing equivalence in terms of standardized differences between the marginal means.

The testing problem to be treated then, reads

$$H : |\mu_j - \mu_l|/\sigma_{(j,l)} \geq \varepsilon \text{ for at least one } (j, l) \in \{1, \dots, k\}^2 \text{ with } j < l \\ \text{versus } K : |\mu_j - \mu_l|/\sigma_{(j,l)} < \varepsilon \forall 1 \leq j < l \leq k, \quad (8.26)$$

with arbitrarily fixed  $\varepsilon > 0$ . Of course, the variances  $\sigma_{(j,l)}^2$  are related to the entries in the covariance matrix  $\Sigma$  of the primarily observed vectors  $(X_{11}, \dots, X_{k1}), \dots, (X_{1n}, \dots, X_{kn})$  [recall (8.2)] by

$$\sigma_{(j,l)}^2 = \sigma_j^2 + \sigma_l^2 - 2\sigma_{jl}. \quad (8.27)$$

A straightforward solution to the problem (8.26) is obtained by applying the intersection union principle of § 7.1 with  $q = k(k-1)/2$  and paired  $t$ -tests for equivalence as building blocks. This yields the decision rule:

Reject  $H$  in favor of  $K$  if and only if

$$\sqrt{n}|\bar{X}_j - \bar{X}_l|/S_{(j,l)} < [F_{1,n-1;\alpha}(n\varepsilon^2)]^{1/2} \quad \forall 1 \leq j < l \leq k, \quad (8.28)$$

with

$$S_{(j,l)}^2 = S_j^2 + S_l^2 - 2S_{jl} = (n-1)^{-1} \left[ \sum_{i=1}^n ((X_{ji} - X_{li}) - (\bar{X}_j - \bar{X}_l))^2 \right]. \quad (8.29)$$

From general facts stated in § 7.1 about intersection-union tests for problems in which the alternative hypothesis specifies that several elementary hypotheses hold true simultaneously, it follows that the procedure defined by (8.28) is exactly valid, in the sense that the type-I error risk is less than or equal to  $\alpha$  under any parameter constellation belonging to  $H$ . Except for  $k = 2$ , exact power computations are largely impracticable even under null

alternatives assuming  $\mu_j - \mu_l = 0 \forall (j, l)$ . Table 8.10 shows, for the case of observations of dimension  $k = 5$ , the simulated power attained against special alternatives of that kind distinguished by 5 different correlation structures (including pairwise independence) as specified in Table 8.9, and equal marginal variances. The fact that for the first three correlation structures (labeled 0 through 2), the results are identical except for some slight variation due to the simulation error, may seem surprising at first sight. Mathematically, it is a consequence of a classical result (stated, e.g., in § 5.3 of Kotz et al., 2000)

Table 8.9 Correlation matrices used for generating 5-dimensional data vectors in studying the power of the union-intersection test given by (8.28).

Correlation Matrix #											
(r)	$\varrho_{12}^{(r)}$	$\varrho_{13}^{(r)}$	$\varrho_{14}^{(r)}$	$\varrho_{15}^{(r)}$	$\varrho_{23}^{(r)}$	$\varrho_{24}^{(r)}$	$\varrho_{25}^{(r)}$	$\varrho_{34}^{(r)}$	$\varrho_{35}^{(r)}$	$\varrho_{45}^{(r)}$	
0	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	
1	.25	.25	.25	.25	.25	.25	.25	.25	.25	.25	
2	.90	.90	.90	.90	.90	.90	.90	.90	.90	.90	
3	.15	.25	.50	.95	.55	.25	.25	.45	.25	.70	
4	-.65	.35	.45	.10	.35	.35	-.20	.95	.20	.10	

Table 8.10 Simulated power of the intersection-union test (8.28) at nominal level  $\alpha = .05$  against null alternatives specifying  $\mu_j - \mu_l = 0 \forall (j, l)$  for 5-dimensional observations, equivalence margin  $\varepsilon = .50$ , homogeneous marginal variances, and the correlation structures listed in Table 8.9 [100,000 replications per Monte Carlo experiment].

$n$	Correlation Matrix # (r)				
	0	1	2	3	4
10	.00048	.00048	.00051	.00132	.00081
20	.02377	.02400	.02399	.05045	.03756
30	.17186	.17000	.16985	.25610	.23052
40	.41566	.41294	.41413	.50967	.49344
50	.63755	.63892	.63935	.71179	.70018
60	.79569	.79808	.79959	.84440	.84135
70	.89525	.89424	.89521	.92058	.91893
80	.94899	.94868	.94795	.96149	.96157
90	.97572	.97550	.97541	.98122	.98113
100	.98874	.98930	.98885	.99120	.99150

on multivariate normal random variables with equi-correlated components: If, for each  $i = 1, \dots, n$  and  $1 \leq j < l \leq k$ , there holds  $\mu_j = \mu_l = \mu$ ,  $\sigma_j = \sigma_l = \sigma_0$  and  $\varrho_{jl} = \varrho$  for some  $\mu \in \mathbb{R}$ ,  $\sigma_0 > 0$  and  $0 \leq \varrho < 1$ , then we have  $X_{ji} - X_{li} \stackrel{d}{=} \sigma_0\sqrt{1-\varrho}(Z_{ji} - Z_{li})$  where all  $Z$ 's are standard normally distributed and mutually independent. Clearly, this implies, that the joint distribution of the  $(k(k-1)/2)$ -vector of pairwise  $t$ -statistics obtained from the sample  $(X_{11}, \dots, X_{k1}), \dots, (X_{1n}, \dots, X_{kn})$  is the same as that computed from the unit normal vectors  $(Z_{1i}, \dots, Z_{ki})$ ,  $i = 1, \dots, n$ , and thus does not depend on  $\varrho$ .

Measuring the distance between pairs of marginal distributions in terms of *nonscaled* differences of their expected values leads to reformulating the testing problem (8.26) through dropping the standard deviations  $\sigma_{(j,l)}$ . The resulting pair of hypotheses can be written in a more compact way as a couple of statements about the range among the population means  $\mu_1, \dots, \mu_k$ . Actually, when we define  $\mu_{(1)} \equiv \min_{1 \leq j \leq k} \mu_j$ ,  $\mu_{(k)} \equiv \max_{1 \leq j \leq k} \mu_j$  and replace  $\sigma_{(j,l)}$  with unity for all  $(j, l)$ , (8.26) simplifies to

$$H : \mu_{(k)} - \mu_{(1)} \geq \varepsilon \quad \text{versus} \quad K : \mu_{(k)} - \mu_{(1)} < \varepsilon. \quad (8.30)$$

In this nonscaled version, the testing problem is likewise easy to solve through applying the intersection-union principle. The natural choice of the elementary tests for equivalence of pairs of marginal distributions is now the paired-samples version of the two one-sided  $t$ -tests procedure, or, equivalently, the interval inclusion test with central  $t$ -based confidence bounds to the respective mean within-subject differences. After suitable rearrangements of the double inequalities defining each of these  $k(k-1)/2$  elementary tests, the compound decision rule of the intersection-union procedure can be written:

Reject  $H$  in favor of  $K$  if and only if

$$|\bar{X}_j - \bar{X}_l| < \varepsilon - (S_{(j,l)} / \sqrt{n})t_{n-1; 1-\alpha} \quad \forall 1 \leq j < l \leq k, \quad (8.31)$$

where  $t_{n-1; 1-\alpha}$  stands for the  $(1 - \alpha)$ -quantile of the central  $t$ -distribution with  $n - 1$  degrees of freedom, and all other symbols have the same meaning as in (8.28–9).

In designing a simulation study providing useful insights into the behavior of the power function of this alternative testing procedure following the intersection-union principle, it must be kept in mind that, in contrast to the procedure using paired  $t$ -tests as building blocks, (8.31) is not invariant under rescaling the primary observations and their components. This is the reason why Table 8.11 shows the result of determining by simulation the power of the procedure against null alternatives, which differ not only with respect to the correlation structure but also the assumed common value of the marginal variances.

Table 8.11 Simulated power of the intersection-union test (8.31) at nominal level  $\alpha = .05$  against null alternatives specifying  $\mu_j - \mu_l = 0 \forall (j, l)$  for 5-dimensional observations, equivalence margin  $\varepsilon = .50$  to the nonscaled range among the  $\mu_j$ , different values of the standard deviation  $\sigma_0$  assumed to be common to all components of the observed random vector, and the correlation structures listed in Table 8.9 [100,000 replications per Monte Carlo experiment].

$n$	Correlation Matrix # (r)	$\sigma_0 =$				
		0.70	0.85	1.00	1.15	1.30
25	0	0.06078	0.00279	0.00002	0.00000	0.00000
"	1	0.22542	0.03192	0.00186	0.00002	0.00000
"	2	1.00000	0.99997	0.99852	0.98216	0.91410
"	3	0.46092	0.14465	0.02476	0.00229	0.00014
"	4	0.10677	0.00990	0.00036	0.00001	0.00000
50	0	0.66422	0.27602	0.06867	0.00969	0.00054
"	1	0.88827	0.56387	0.24125	0.07193	0.01405
"	2	1.00000	1.00000	1.00000	0.99999	0.99991
"	3	0.94622	0.75659	0.47051	0.22677	0.08033
"	4	0.66700	0.33531	0.10851	0.01715	0.00107
75	0	0.94409	0.69393	0.36030	0.13813	0.03739
"	1	0.99317	0.90122	0.65685	0.36769	0.16435
"	2	1.00000	1.00000	1.00000	1.00000	1.00000
"	3	0.99603	0.95208	0.81519	0.59470	0.37301
"	4	0.90407	0.68614	0.41757	0.19045	0.05799
100	0	0.99285	0.90648	0.66334	0.37274	0.16575
"	1	0.99961	0.98548	0.88590	0.67308	0.41991
"	2	1.00000	1.00000	1.00000	1.00000	1.00000
"	3	0.99981	0.99212	0.94558	0.82536	0.63899
"	4	0.97371	0.86438	0.66086	0.42964	0.22348

Even when the common standard deviation  $\sigma_0$  of the components of the observed random vectors is set equal to unity, the power attained in the test (8.31) with some given sample size  $n$  turns out to be very sensitive to changes in the correlation structure. Through comparison with the entries in the previous table, it becomes evident that this property distinguishes the nonscaled version of the intersection-union test quite sharply from its counterpart (8.28) tailored for establishing equivalence with respect to the standardized differences of means. The most marked changes occur under transition from cor-

relation structure 0 [independence] to 2 [equal correlations of common value  $\varrho = .90$ ]. Taking an asymptotic point of view, that part of the phenomenon is fairly easy to understand: For  $\varrho_{jl} = \varrho \forall (j, l)$  and  $\sigma_1 = \dots = \sigma_k = 1$ , each of the elementary tests making up the procedure (8.31) is asymptotically equivalent to the test with rejection region  $\{\sqrt{n}|\bar{X}_j - \bar{X}_l|/\sqrt{2(1-\varrho)} < \sqrt{n}\varepsilon/\sqrt{2(1-\varrho)} - u_{1-\alpha}\}$  [cf. p. 64]. Consequently, increasing the common correlation coefficient from zero to some nonzero value  $\varrho < 1$  has on its asymptotic power against  $\mu_j - \mu_l = E(\bar{X}_j - \bar{X}_l) = 0$  the same effect as increasing the equivalence margin by the factor  $(1-\varrho)^{-1/2}$ . Specifically, for  $\varepsilon = 0.50$  and  $\varrho = .90$ , this would amount to increasing the equivalence margin to  $\approx 1.58$ . The difference to .50 intuitively explains the huge increase in power of the compound testing procedure as a whole.

### 8.1.4 Discussion

None of the testing procedures made available in this subsection for establishing equivalence of several normal distributions from which dependent samples are taken, seems fully satisfactory. The test based on Hotelling's  $T^2$  has the advantage of guaranteeing exact coincidence of the size of its critical region with the prespecified nominal significance level  $\alpha$ . Furthermore, it satisfies a fairly strong optimality criterion (UMPI), and its power against null alternatives does not depend in any way on the covariance matrix. Unfortunately, the importance of these desirable properties has to be considerably relativized taking into account the fact [illustrated by [Figure 8.1](#)] that the theoretical equivalence region which the true parameter configuration can be said to belong to in the case of a positive decision, highly depends both in shape and geometric content on the correlation structure. Not a few users of the methods under discussion might feel unable to see why this is reasonable.

The pros and cons of the procedure derived in § 8.1.2 are largely dual to those of the exact test for equivalence as defined in terms of Mahalanobis distance of the vector of expected pairwise intrasubject differences from zero. Measuring the distance of a vector of real numbers from the origin through Euclidean metric has a long tradition in many scientific areas, and the shape of the resulting region in the space of standardized expected values is spherical for any underlying correlation structure. However, the testing procedure obtained for the equivalence problem formulated in terms of Euclidean distances relies on first-order asymptotics, and convergence of the rejection probability under various parameter configurations falling on the boundary of both hypotheses to the prespecified significance level turned out to be extremely slow. Even apart from the problem of performing a test of maybe considerably increased size which can be addressed by reducing the nominal level, the power against null alternatives is fairly sensitive against changes in the correlation structure so that sample size determination requires much more

detailed knowledge about the true parameter configuration as compared to the exact test based on Hotelling's one-sample  $T^2$ -statistic.

Regarding the geometric shape of the indifference zone specified under the alternative hypothesis one aims to establish, an option comparable in intuitive appeal to the approach based on Euclidean metric, is to require pairwise equivalence of the marginal means for all possible pairs leading to a rectangular equivalence region in  $(k(k - 1)/2)$ -dimensional space. Both testing procedures proposed in § 8.1.3 for that type of hypotheses formulation make use of the intersection-union principle being thus exactly valid in terms of the significance level though prone to conservatism. Remarkably, the extent of that conservatism is by no means dramatic as turns out when comparing the first of these procedures (tailored for establishing equivalence in terms of standardized expected values) to a competitor ensuring coincidence between nominal significance level and effective size of the critical region. Although having to rely on much more restrictive model assumptions (homoskedasticity and equality of all pairwise correlations), a procedure of this latter kind is that studied by Giani and Finner (1991). For  $k = 5$ ,  $\varrho = 0.00$ ,  $\sigma_0 = 1.00$  and  $\delta_0 \equiv \varepsilon / (\sigma_0 \sqrt{1 - \varrho}) = .70$ , these authors report their test to reach a power of 95% against  $\mu_1 = \dots = \mu_k$  when performed with  $n = 76$  observations at (nominal) level  $\alpha = .05$ . Since, in the homoskedastic case with unit variances and vanishing pairwise correlations, we have  $\sigma_{(j,l)} = \sqrt{2} \forall (j, l)$  and  $\varepsilon$  was set, according to (8.26), as the equivalence margin to  $(\mu_j - \mu_l) / \sigma_{(j,l)}$ , this value has to be compared with the power of our test for  $\varepsilon = .70 / \sqrt{2} = .4945 \approx .5$  and  $n = 76$  under correlation structure (0) [recall Table 8.9], which, from the second column of Table 8.10, is estimated to fall in the interval from  $\approx .895$  through  $\approx .95$  (an extra simulation experiment run with the exact values of  $\varepsilon$  and  $n$  gave .92244). Taking into consideration that the intersection-union test based on pairwise  $t$ -tests for equivalence is valid under any covariance structure, the loss in power compared to Giani and Finner's procedure seems far from being serious for practical purposes.

## 8.2 Multivariate two-sample tests for equivalence

### 8.2.1 A two-sample test for equivalence based on Hotelling's $T^2$

The test to be studied in this subsection can be viewed as the natural multivariate generalization of the two-sample  $t$ -test for equivalence of two homoskedastic univariate Gaussian distributions introduced in § 6.1 since it reduces to the latter in the case that  $k = 1$ . The changes which have to be made to its one-sample analogue (8.10) in order to obtain the decision rule

to be used in the case of two samples from  $k$ -variate normal distributions are fairly similar to those leading from the classical one-sample  $T^2$ -test to a test of the null hypothesis that the vectors of expected values of two  $k$ -variate Gaussian distributions with common though unknown covariance matrix coincide. Analogously to (8.1), the basic assumption is that we are given two independent samples  $(X_{11}, \dots, X_{k1}), \dots, (X_{1m}, \dots, X_{km}), (Y_{11}, \dots, Y_{k1}), \dots, (Y_{1n}, \dots, Y_{kn})$  of  $k$ -variate observations of possibly different sizes  $m$  and  $n$  such that

$$\begin{aligned} (X_{1u}, \dots, X_{ku}) &\sim \mathcal{N}((\mu_1, \dots, \mu_k), \Sigma), \quad \forall u = 1, \dots, m, \\ (Y_{1v}, \dots, Y_{kv}) &\sim \mathcal{N}((\nu_1, \dots, \nu_k), \Sigma), \quad \forall v = 1, \dots, n. \end{aligned} \quad (8.32)$$

For the elements of the common (positive definite) covariance matrix  $\Sigma$ , the notation introduced in (8.2) will be adopted in the two-sample case as well.

The equivalence hypothesis we are now interested in specifies that the (squared) Mahalanobis distance between the two population mean vectors  $\mu$  and  $\nu$  is smaller than some equivalence margin  $\varepsilon^2$ . Thus, the complete formulation of the testing problem is

$$\begin{aligned} H : (\mu - \nu)\Sigma^{-1}(\mu - \nu)' &\geq \varepsilon^2 \\ \text{versus } K : (\mu - \nu)\Sigma^{-1}(\mu - \nu)' &< \varepsilon^2. \end{aligned} \quad (8.33)$$

An optimal test for (8.33) can be based on the two-sample counterpart of the  $T^2$ -statistic (8.8) which is traditionally likewise named after H. Hotelling. Again, it can be represented as a normalized plug-in estimator of the population Mahalanobis distance. More precisely speaking,  $\mu$ ,  $\nu$  and  $\Sigma$  have, respectively, to be replaced with the empirical mean vectors  $\bar{\mathbf{X}} = m^{-1} \sum_{u=1}^m (X_{1u}, \dots, X_{ku})$ ,  $\bar{\mathbf{Y}} = n^{-1} \sum_{v=1}^n (Y_{1v}, \dots, Y_{kv})$ , and the pooled covariance matrix

$$\begin{aligned} \mathbf{S} = \frac{1}{N-2} \left( \sum_{u=1}^m (X_{1u} - \bar{X}_1, \dots, X_{ku} - \bar{X}_k)' (X_{1u} - \bar{X}_1, \dots, X_{ku} - \bar{X}_k) \right. \\ \left. + \sum_{v=1}^n (Y_{1v} - \bar{Y}_1, \dots, Y_{kv} - \bar{Y}_k)' (Y_{1v} - \bar{Y}_1, \dots, Y_{kv} - \bar{Y}_k) \right), \end{aligned} \quad (8.34)$$

with  $N$  denoting the total sample size  $m+n$ . Using  $mn/N$  as a normalizing factor, the test statistic can now be written

$$T^2 = \frac{mn}{N} (\bar{\mathbf{X}} - \bar{\mathbf{Y}}) \mathbf{S}^{-1} (\bar{\mathbf{X}} - \bar{\mathbf{Y}})' . \quad (8.35)$$

The distribution of  $T^2$  under any  $(\mu, \nu, \Sigma)$  such that  $(\mu - \nu)\Sigma^{-1}(\mu - \nu)' = \delta^2 > 0$ , is well-known (see, eg., Anderson, 1984, § 5.2.2) to be given by

$$T^2 (N - k - 1) / (k(N - 2)) \stackrel{d}{=} \mathcal{F}_{k, N-k-1}(mn\delta^2/N), \quad (8.36)$$

where, as before,  $\mathcal{F}_{k,N-k-1}(\psi^2)$  denotes a random variable following an  $F$ -distribution with  $k, N - k - 1$  degrees of freedom and noncentrality parameter  $\psi^2 \geq 0$ . Hence, the critical region of an exact level- $\alpha$  test for (8.33) can be written

$$\left\{ T^2 < (k(N-2)/(N-k-1)) F_{k,N-k-1;\alpha}(mn\varepsilon^2/N) \right\}. \quad (8.37)$$

Table 8.12 is the analogue of Table 8.1 showing the critical upper bounds to the two-sample  $T^2$ -statistic to be used in testing (8.33) for the same three choices of the equivalence margin and the same range for the dimension  $k$  of the data vectors. The underlying design is assumed to be balanced so that  $n$  stands for the number of observations contained in each group. The sample size is let increase in steps of 5 from 10 through 50; beyond 50, the increment is 10.

The power of the equivalence test with critical region (8.37) against any specific alternative  $(\boldsymbol{\mu}, \boldsymbol{\nu}, \boldsymbol{\Sigma})$  with  $(\boldsymbol{\mu} - \boldsymbol{\nu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\nu})' = \delta_a^2 \in [0, \varepsilon^2]$  admits the explicit representation

$$\beta(\delta_a^2) = P[\mathcal{F}_{k,N-k-1}(mn\delta_a^2/N) \leq F_{k,N-k-1;\alpha}(mn\varepsilon^2/N)] \quad (8.38)$$

where  $\mathcal{F}_{k,N-k-1}(\psi^2)$  and  $F_{k,N-k-1;\alpha}(\psi^2)$  has the same meaning as in (8.36) and (8.37), respectively. The tabulation of power values given in Table 8.13 relates to null alternatives under which both mean vectors  $\boldsymbol{\mu}$  and  $\boldsymbol{\nu}$  coincide and covers the same choices of  $n, k$  and  $\varepsilon$  as were made when compiling Table 8.12.

Like its univariate counterpart [cf. § 6.1], the  $T^2$ -test for equivalence of two multivariate Gaussian distributions with common covariance structure exhibits a number of noticeable mathematical properties. First of all, it is unbiased at the prespecified nominal significance level  $\alpha \in (0, 1)$ . Moreover, it is uniformly most powerful among all level- $\alpha$  tests for (8.33) which remain invariant under a large group of linear transformations of the sample space generated by several subgroups. The most important of these transformations are of the form

$$(X_{1u}, \dots, X_{ku}) \mapsto (a_1, \dots, a_k) + (X_{1u}, \dots, X_{ku}) \mathbf{B}, u = 1, \dots, m, \\ (Y_{1v}, \dots, Y_{kv}) \mapsto (a_1, \dots, a_k) + (Y_{1v}, \dots, Y_{kv}) \mathbf{B}, v = 1, \dots, n,$$

with  $(a_1, \dots, a_k) \in \mathbb{R}^k$  and  $\mathbf{B}$  as an arbitrary nonsingular  $k \times k$  matrix.

Table 8.12 Critical constant  $T_{0.05; n, n}^{2;k}(\varepsilon)$  for the two-sample  $T^2$ -test for equivalence of two homoskedastic  $k$ -variate normal distributions with samples of common size  $n = 10(5)50(10)80$ , significance level  $\alpha = 5\%$ , and equivalence margin  $\varepsilon \in \{.25, .50, 1.00\}$  to the root of the Mahalanobis distance of both mean vectors.

$n$	$\varepsilon$	$k =$				
		2	3	4	5	10
10	0.25	0.12733	0.43070	0.88562	1.47559	6.83045
10	0.50	0.20187	0.58383	1.11129	1.76869	7.47811
10	1.00	0.96363	1.64415	2.42416	3.31564	10.34322
15	0.25	0.13462	0.43722	0.87233	1.41398	5.54220
15	0.50	0.26634	0.68485	1.21902	1.84702	6.33757
15	1.00	1.87953	2.63189	3.44537	4.32590	10.01409
20	0.25	0.14398	0.45258	0.88469	1.41115	5.15320
20	0.50	0.35190	0.81354	1.37009	2.00154	6.14646
20	1.00	3.01540	3.81391	4.65792	5.55069	10.88636
25	0.25	0.15460	0.47170	0.90643	1.42816	4.99010
25	0.50	0.46058	0.96519	1.54725	2.19094	6.19976
25	1.00	4.29694	5.12926	5.99823	6.90591	12.10975
30	0.25	0.16630	0.49314	0.93303	1.45464	4.91821
30	0.50	0.59394	1.13827	1.74529	2.40375	6.35618
30	1.00	5.68458	6.54338	7.43302	8.35489	13.50519
35	0.25	0.17903	0.51636	0.96275	1.48663	4.89240
35	0.50	0.75226	1.33155	1.96163	2.63514	6.56813
35	1.00	7.15429	8.03467	8.94171	9.87647	15.00741
40	0.25	0.19279	0.54112	0.99478	1.52231	4.89325
40	0.50	0.93449	1.54367	2.19447	2.88238	6.81493
40	1.00	8.69003	9.58846	10.51043	11.45676	16.58506
45	0.25	0.20762	0.56727	1.02869	1.56073	4.91104
45	0.50	1.13874	1.77313	2.44231	3.14364	7.08603
45	1.00	10.28043	11.19428	12.12924	13.08594	18.22036
50	0.25	0.22355	0.59478	1.06423	1.60131	4.94032
50	0.50	1.36270	2.01839	2.70383	3.41750	7.37538
50	1.00	11.91706	12.84431	13.79070	14.75675	19.90197
60	0.25	0.25891	0.65371	1.13965	1.68775	5.02151
60	0.50	1.86097	2.55052	3.26322	3.99852	7.99502
60	1.00	15.30460	16.25412	17.21980	18.20203	23.37521
70	0.25	0.29926	0.71782	1.22036	1.78003	5.12215
70	0.50	2.41433	3.12986	3.86449	4.61795	8.65598
70	1.00	18.81542	19.78287	20.76433	21.76010	26.96377
80	0.25	0.34498	0.78711	1.30601	1.87735	5.23538
80	0.50	3.01198	3.74842	4.50119	5.27009	9.34877
80	1.00	22.42484	23.40714	24.40185	25.40916	30.64290

Table 8.13 *Power of the two-sample  $T^2$ -test for equivalence at level  $\alpha = .05$  against the alternative that both mean vectors coincide, for the values of  $k$ ,  $n$  and  $\varepsilon$  covered by Table 8.12.*

$n$	$\varepsilon$	$k =$				
		2	3	4	5	10
10	0.25	.05815	.05767	.05715	.05667	.05473
10	0.50	.09044	.08691	.08366	.08083	.07069
10	1.00	.35808	.30400	.26723	.23976	.15938
15	0.25	.06270	.06193	.06115	.06043	.05781
15	0.50	.11998	.11282	.10684	.10194	.08629
15	1.00	.58400	.50257	.44449	.40006	.26961
20	0.25	.06757	.06646	.06537	.06439	.06093
20	0.50	.15678	.14410	.13432	.12664	.10366
20	1.00	.75651	.67794	.61519	.56357	.39667
25	0.25	.07278	.07127	.06982	.06854	.06415
25	0.50	.20102	.18077	.16611	.15498	.12309
25	1.00	.86665	.80652	.75261	.70446	.52700
30	0.25	.07836	.07637	.07451	.07290	.06748
30	0.50	.25201	.22244	.20201	.18684	.14462
30	1.00	.93038	.89026	.85032	.81168	.64744
35	0.25	.08432	.08178	.07946	.07747	.07094
35	0.50	.30827	.26838	.24158	.22194	.16824
35	1.00	.96496	.94051	.91382	.88600	.74932
40	0.25	.09067	.08749	.08465	.08225	.07452
40	0.50	.36778	.31753	.28418	.25987	.19387
40	1.00	.98287	.96891	.95237	.93392	.82931
45	0.25	.09743	.09353	.09011	.08726	.07823
45	0.50	.42838	.36872	.32905	.30010	.22137
45	1.00	.99182	.98424	.97457	.96309	.88827
50	0.25	.10462	.09989	.09584	.09250	.08206
50	0.50	.48816	.42077	.37540	.34204	.25057
50	1.00	.99617	.99221	.98681	.98003	.92942
60	0.25	.12034	.11362	.10810	.10365	.09014
60	0.50	.59964	.52312	.46926	.42858	.31319
60	1.00	.99920	.99821	.99670	.99460	.97436
70	0.25	.13790	.12872	.12147	.11573	.09875
70	0.50	.69519	.61771	.55996	.51471	.37977
70	1.00	.99984	.99962	.99924	.99866	.99162
80	0.25	.15736	.14521	.13594	.12875	.10791
80	0.50	.77290	.70056	.64322	.59641	.44811
80	1.00	.99997	.99992	.99983	.99969	.99749

*Example 8.2*

The data set we use for illustrating the two-sample  $T^2$ -test for equivalence is taken from the same context as described in the brief introduction to Example 8.1. This time, two different lines of alcohol-preferring rats are to be assessed for similarity of the profiles obtained by measuring an animal's total intake of ethanol [g/kg/day] on  $k = 4$  consecutive days after alcohol deprivation. A sample of size  $n = 15$  was available for the experiment from each line. Table 8.14 shows the arithmetic means and the empirical covariance matrix calculated from both samples of data vectors.

Table 8.14 *Summary statistics for two independent samples of repeated measurements of total ethanol intake [g/kg/day] taken in rats of two different alcohol-preferring lines at  $k = 4$  consecutive days after deprivation.*

	Day $j =$			
	1	2	3	4
$\bar{X}_j$	5.266324	4.7816957	4.7213656	5.1046539
$S_{1j}^x$	6.8492111	-1.398337	-0.848002	-0.758188
$S_{2j}^x$	-1.398337	1.0855432	0.4738331	0.5639499
$S_{3j}^x$	-0.848002	0.4738331	1.0993696	0.6963447
$S_{4j}^x$	-0.758188	0.5639499	0.6963447	0.6525258
$\bar{Y}_j$	5.3756573	4.4723527	4.8108581	5.0109942
$S_{1j}^y$	5.7589209	0.2711184	-0.284273	-0.254542
$S_{2j}^y$	0.2711184	0.8766867	0.183452	0.1820176
$S_{3j}^y$	-0.284273	0.183452	0.7138245	0.4123788
$S_{4j}^y$	-0.254542	0.1820176	0.4123788	0.3939202

The additional computational effort required for determining the value of the test statistic  $T^2$  with the entries in this table is almost negligible for a user working in a matrix-oriented programming environment. Both of the equations (8.34) and (8.35) can be translated into a single assignment statement then. Executing a brief script containing these statements yields the result  $T^2 = 1.404995$ . Setting the equivalence margin to the theoretical squared Mahalanobis distance between both distributions equal to  $\varepsilon^2 = .25$ , the critical upper bound to which this value has to be compared, is read from Table 8.12 to be  $T_{.05; n,n}^{2;k}(\varepsilon) = 1.21902$ . Since the calculated value of  $T^2$  obviously fails to remain below this bound, the null hypothesis of relevant differences in terms of Mahalanobis distance cannot be rejected in the present case. According to Table 8.13, even if the population mean vectors were exactly the same for both groups, the chance of coming out with a positive decision was only a bit

larger than 10% given the fairly small size of both samples and the choice of the equivalence margin. Thus, with a view to power, the result is far from being surprising.

### 8.2.2 Behavior of the two-sample $T^2$ -test for equivalence under heteroskedasticity

The easiest way of handling settings where both population covariance matrices differ is to rely on the assumption that the equivalence version of Hotelling's two-sample test is sufficiently robust against such deviations from the standard model under which the test is exact. The first step which has to be taken when one wants to investigate to what extent this assumption is warranted, consists of generalizing the hypotheses formulation (8.33) by replacing ordinary Mahalanobis distance with the quadratic form in the average covariance matrix  $\bar{\Sigma} \equiv (1/2)(\Sigma_1 + \Sigma_2)$  evaluated, as before, at  $(\mu - \nu)$ . Of course, this requires that  $\bar{\Sigma}$  is of the same rank as  $\Sigma_1$  and  $\Sigma_2$ , namely  $k$ . Under that additional restriction, generalizing (8.33) in a natural way leads to considering

$$\begin{aligned} H : & (\mu - \nu)(\Sigma_1 + \Sigma_2)^{-1}(\mu - \nu)' \geq \varepsilon^2/2 \\ \text{versus } K : & (\mu - \nu)(\Sigma_1 + \Sigma_2)^{-1}(\mu - \nu)' < \varepsilon^2/2. \end{aligned} \quad (8.33^*)$$

The question for what deviations between  $\Sigma_1$  and  $\Sigma_2$  the test with rejection region (8.37) keeps being valid in the sense of maintaining the significance level at least approximately under the null hypothesis of (8.33 $^*$ ) has been addressed in another series of simulation experiments. All individual experiments performed for that purpose related to the same number of measurements taken per individual as in Example 8.2, namely  $k = 4$ . They differed with regard to which of the correlation structures, vectors of marginal variances and location-shift patterns specified in Table 8.15 were assigned to the populations in generating the data. The shift patterns were implemented in the following way: The distribution underlying the first sample was assumed to be centered about  $\mathbf{0}$ ; the vector  $\nu$  of expected values of the  $Y_{iv}$  was set equal to  $c_\varepsilon (\delta_1^{(s)}, \dots, \delta_k^{(s)})$  with  $c_\varepsilon$  being determined by solving the equation

$$(\delta_1^{(s)}, \dots, \delta_k^{(s)})(\Sigma_1 + \Sigma_2)^{-1}(\delta_1^{(s)}, \dots, \delta_k^{(s)})' = \varepsilon^2/(2c^2), \quad c > 0.$$

Table 8.16 shows the simulated rejection probabilities obtained for various combinations of sample sizes and parameter configurations generated by assigning one of the correlation structures, sets of marginal variances and location-shift patterns appearing in Table 8.15 to each of the multivariate normal distributions to be tested for equivalence with respect to the generalized Mahalanobis distance. The major conclusions to be drawn from these results are as follows:

- (i) As long as the design is balanced with respect to sample size, heteroskedasticity does not lead to serious exceedances of the rejection

Table 8.15 *Correlation matrices, marginal variances and location-shift patterns used for generating 4-dimensional data vectors in studying the size of the  $T^2$ -test for equivalence under heteroskedasticity.*

Correlation Matrix #		$\varrho_{12}^{(r)}$	$\varrho_{13}^{(r)}$	$\varrho_{14}^{(r)}$	$\varrho_{23}^{(r)}$	$\varrho_{24}^{(r)}$	$\varrho_{34}^{(r)}$
1		.25	.25	.25	.25	.25	.25
2		.90	.90	.90	.90	.90	.90
3		-.65	.35	.45	.35	.35	.95
Vector of Marginal Standard Deviations #		$\sigma_1^{(q)}$	$\sigma_2^{(q)}$	$\sigma_3^{(q)}$	$\sigma_4^{(q)}$		
1		1.000	1.000	1.000	1.000		
2		0.100	0.100	0.100	0.100		
3		100.0	100.0	100.0	100.0		
Location-Shift Pattern #		$\delta_1^{(l)}$	$\delta_2^{(l)}$	$\delta_3^{(l)}$	$\delta_4^{(l)}$		
1		1	1	1	1		
2		1	0	0	0		
3		1	2	3	4		

probability under the null hypothesis over the threshold specified as the nominal significance level. More marked is a tendency towards anticonservatism.

- (ii) Under gross imbalance regarding the sample sizes, huge discrepancies between actual size of the test and the nominal significance level can occur both in the conservative and the liberal direction. A grossly anticonservative behavior of the test has to be expected when the larger of both samples is taken from the population with large entries in the covariance matrix, in particular large marginal variances. Switching this assignment leads to gross overconservatism.

- (iii) Given the sample sizes and both covariance matrices, moving along the common boundary of the hypotheses (8.33\*) through changing the

Table 8.16 *Simulated rejection probabilities of the two-sample  $T^2$ -test for equivalence at the boundary of the hypotheses (8.33\*) with  $\varepsilon = .5$  for various specific parameter combinations of the correlation structures, marginal variances and location-shift patterns specified in Table 8.15. [The meaning of the symbols  $r$ ,  $q$  and  $l$  is the same as in the previous table; 100,000 replications per Monte Carlo experiment were carried out.]*

Sample 1			Sample 2			(l)	Rejection probability
(r)	(q)	m	(r)	(q)	n		
1	2	10	2	3	10	1	.03782
"	"	50	"	"	50	"	.04779
"	"	10	"	"	50	"	.23852
"	"	50	"	"	10	"	.00209
"	"	10	"	"	50	2	.23960
"	"	50	"	"	10	2	.00198
"	"	10	"	"	50	3	.23811
"	"	50	"	"	10	3	.00199
1	2	10	3	1	10	1	.04505
"	"	50	"	"	50	"	.04963
"	"	10	"	"	50	"	.17244
"	"	50	"	"	10	"	.00592
"	"	10	"	"	50	2	.17548
"	"	50	"	"	10	2	.00574
"	"	10	"	"	50	3	.13933
"	"	50	"	"	10	3	.00703
2	3	10	3	1	10	1	.03698
"	"	50	"	"	50	"	.04816
"	"	10	"	"	50	"	.00193
"	"	50	"	"	10	"	.23763
"	"	10	"	"	50	2	.00193
"	"	50	"	"	10	2	.23823
"	"	10	"	"	50	3	.00189
"	"	50	"	"	10	3	.23777

location-shift pattern only slightly affects the rejection probability of the test.

In summary, we may conclude that applying the two-sample  $T^2$ -test also under heteroskedasticity is an option as long as the study under consideration follows an approximately balanced design and Mahalanobis distance is generalized in the way made precise in (8.33\*). For considerable differences between both samples with regard to the number of data vectors contained, the homoskedasticity condition is crucial, and applying the test in presence of evidence against this assumption must be discouraged.

### 8.2.3 Multivariate two-sample tests for equivalence regions of rectangular shape

Even when all model assumptions are satisfied so that the  $T^2$ -test for equivalence is exactly valid and UMPI, the shape of the underlying theoretical equivalence region is basically the same as in the one-sample case. Accordingly, there seems to be again considerable need for tests for inclusion of the true parameter point in an equivalence region of more intuitive form being independent of the covariance structure. In principle, both approaches followed in the one-sample case for modifying the shape of the theoretical equivalence region in a way making it invariant against changes of the common covariance matrix, can easily be adopted for the two-sample setting as well. The two-sample analogue of the asymptotic test for equivalence with respect to squared Euclidean rather than Mahalanobis distance derived in § 8.1.2 entails essentially the same problems with the low speed of convergence of the distribution of the suitably standardized maximum-likelihood test statistic to  $\mathcal{N}(0, 1)$ . Hence, we do not pursue this approach here any further (for a detailed description of the procedure see Hoffelder, 2006) and confine ourselves to a brief consideration of the testing problem

$$H : |\mu_j - \nu_j|/\sigma_j \geq \varepsilon \text{ for at least one } j \in \{1, \dots, k\}$$

versus  $K : |\mu_j - \nu_j|/\sigma_j < \varepsilon \forall 1 \leq j \leq k,$  (8.39)

with arbitrarily fixed  $\varepsilon > 0$ .

Obviously, this is another example of a testing problem which can easily be solved by means of the intersection-union principle. This time, the univariate two-sample  $t$ -test for equivalence of § 6.1 is the natural choice for testing each of the elementary hypotheses making up (8.39), and the corresponding rejection rule reads:

Reject  $H$  in favor of  $K$  if and only if

$$\sqrt{mn/N} |\bar{X}_j - \bar{Y}_j|/\tilde{S}_j < [F_{1, N-2; \alpha}(mne^2/N)]^{1/2} \quad \forall 1 \leq j \leq k, \quad (8.40)$$

with

$$\tilde{S}_j^2 = \left( \frac{N}{mn(N-2)} \right) \left[ \sum_{u=1}^m (X_{ju} - \bar{X}_j)^2 + \sum_{v=1}^n (Y_{jv} - \bar{Y}_j)^2 \right], \quad N = m+n. \quad (8.41)$$

A prerequisite for applying the procedure in practice is some knowledge of its power against alternatives under which both multivariate Gaussian distributions from which the samples are taken, coincide. Except for the case of bivariate observations or observed vectors of any dimension with mutually independent components, exact power calculation is beyond the limits of computational feasibility. Hence, the entries in Table 8.17 which relates to multivariate normal observations of dimension  $k = 5$  and four different correlation structures, were generated by means of simulation.

Table 8.17 *Simulated<sup>†</sup> power of the intersection-union test (8.40) for equivalence of two  $k$ -variate Gaussian normal distributions against alternatives with  $\mu_j - \nu_j = 0 \forall j = 1, \dots, k$ , for  $k = 5$ ,  $\alpha = .05$ ,  $\varepsilon = .50$ , the correlation structures of Table 8.9, and common sample size  $m = n = 10(5)50(10)100$ . [100,000 replications per Monte Carlo experiment.]*

$n$	Correlation Matrix # (r)				
	0	1	2	3	4
10	.00001	.00001	.00042	.00009	.00026
15	.00003	.00005	.00167	.00024	.00053
20	.00015	.00029	.00598	.00107	.00184
25	.00062	.00083	.01907	.00385	.00596
30	.00234	.00274	.05261	.01060	.01575
35	.00757	.00965	.11506	.02734	.03439
40	.02015	.02438	.20062	.05784	.07055
45	.04430	.05340	.29159	.10509	.11824
50	.08247	.09320	.38326	.16536	.18056
60	.19782	.22196	.54069	.31782	.32350
70	.34400	.37019	.66487	.46842	.46632
80	.49256	.51864	.75902	.60409	.59982
90	.62420	.64728	.82955	.71136	.70476
100	.73078	.74530	.87757	.79342	.78905

<sup>†</sup>) For  $r = 0$  [ $\leftrightarrow$  independence], the tabulated power values were computed exactly applying formula (6.7) and exponentiating the result.

Not surprisingly, it becomes obvious from the above results that the power of the intersection-union test (8.40) highly depends on the correlation structure. Least favorable among the five correlation patterns investigated is pairwise independence of all components of the observed random vectors, and the largest power values are attained under the constellation indexed by  $r = 2$  [ $\leftrightarrow$  equal pairwise correlations of .90]. Under independence, 50 observations per group are required for excluding that the power of the test falls below the significance level. In order to guarantee a power of 80% under all correlation structures, both samples would have to consist of more than 100 data points.

---

## *Tests for establishing goodness of fit*

---

### 9.1 Testing for equivalence of a single multinomial distribution with a fully specified reference distribution

In their simplest form, problems of testing for goodness rather than lack of fit involve a fully specified multinomial distribution, together with some sufficiently small “neighborhood” of distributional models of the same kind which according to the alternative hypothesis to be established by means of the data, contains the true distribution from which the sample has been taken. In such a setting the primary data set consists of  $n$  mutually independent random vectors  $(Y_{11}, \dots, Y_{1k}), \dots, (Y_{n1}, \dots, Y_{nk})$  of dimension  $k \geq 2$  where  $(Y_{i1}, \dots, Y_{ik})$  indicates in what out of  $k$  mutually exclusive categories the  $i$ th sampling unit is observed to be placed. Thus,  $(Y_{i1}, \dots, Y_{ik})$  consists of exactly  $k - 1$  zeros and a single one, and the probability that the nonzero component appears at the  $j$ th position is assumed to be the same number  $\pi_j \in [0, 1]$  for each element of the sample. As usual in the analysis of categorical data, these vectors are aggregated to the corresponding cell counts  $(X_1, \dots, X_k)$  form the beginning defining  $X_j \equiv \#\{i \mid i \in \{1, \dots, n\}, Y_{ij} = 1\}$  for  $j = 1, \dots, k$ . Then, we have  $\sum_{j=1}^k X_j = n$ , and the distribution of  $(X_1, \dots, X_k)$  is multinomial with parameters  $n$  and  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_k)$  given by the probability mass function

$$P[X_1 = x_1, \dots, X_k = x_k] = n! \prod_{j=1}^k \pi_j^{x_j} / x_j! ,$$

$$(x_1, \dots, x_k) \in \left\{ (\tilde{x}_1, \dots, \tilde{x}_k) \mid \tilde{x}_j \in \mathbb{N}_0 \forall j = 1, \dots, k, \sum_{j=1}^k \tilde{x}_j = n \right\}. \quad (9.1)$$

The usual shorthand notation for this distribution is  $\mathcal{M}(n; \boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_k)$ .

The reference distribution of  $(X_1, \dots, X_k)$  to which the true one is asserted to fit sufficiently well under the alternative hypothesis to be established by assessing the observed cell counts, is simply given by some specific choice of the parameter (row-) vector  $\boldsymbol{\pi}$ , say  $\boldsymbol{\pi}^\circ = (\pi_1^\circ, \dots, \pi_k^\circ) \in \{(p_1, \dots, p_k) \in [0, 1]^k \mid \sum_{j=1}^k p_j = 1\}$ . In order to quantify the degree of dissimilarity between the true and the target distribution of  $(X_1, \dots, X_k)$ , it seems reasonable to

use the ordinary Euclidean distance between the associated parameter vectors which leads to the following formulation of the problem of testing for goodness of fit of  $\mathcal{M}(n; \pi_1, \dots, \pi_k)$  to  $\mathcal{M}(n; \pi_1^\circ, \dots, \pi_k^\circ)$ :

$$H : d^2(\boldsymbol{\pi}, \boldsymbol{\pi}^\circ) \geq \varepsilon^2 \quad \text{versus} \quad K : d^2(\boldsymbol{\pi}, \boldsymbol{\pi}^\circ) < \varepsilon^2. \quad (9.2)$$

As before [recall (7.24)],  $d^2(\cdot, \cdot)$  denotes the square of the metric in the Euclidean space of prespecified dimension. In other words, (9.2) has to be read with

$$d^2(\boldsymbol{\pi}, \boldsymbol{\pi}^\circ) = \sum_{j=1}^k (\pi_j - \pi_j^\circ)^2 \quad \forall \boldsymbol{\pi} \in \boldsymbol{\Pi} \quad (9.3)$$

where  $\boldsymbol{\Pi}$  denotes the parameter space for the underlying family of distributions, i.e., the hyperplane  $\{(p_1, \dots, p_k) \in [0, 1]^k \mid \sum_{j=1}^k p_j = 1\}$  in the  $k$ -dimensional unit-cube.

In the special case that there are just two different categories, (9.2) reduces to the problem considered in § 4.3, with  $(\pi_1^\circ - \varepsilon/\sqrt{2}, \pi_1^\circ + \varepsilon/\sqrt{2})$  as the equivalence range to the binomial parameter  $p = \pi_1$ . For larger values of  $k$ , there is no hope of being able to solve (9.2) in a way yielding an exact optimal testing procedure. Instead, we will rely again on asymptotic methods starting from the well-known fact (see, e.g., Bishop et al., 1975, Theorem 14.3–4) that the normalized vector

$$\sqrt{n}(\hat{\pi}_1 - \pi_1, \dots, \hat{\pi}_k - \pi_k) \equiv \sqrt{n}(X_1/n - \pi_1, \dots, X_k/n - \pi_k) \quad (9.4)$$

of relative frequencies converges in law (as  $n \rightarrow \infty$ ) to a random variable which follows a  $k$ -dimensional Gaussian distribution with mean  $\mathbf{0}$  and covariance matrix

$$\boldsymbol{\Sigma}_{\boldsymbol{\pi}} = \mathbf{D}_{\boldsymbol{\pi}} - \boldsymbol{\pi}' \boldsymbol{\pi} \equiv \begin{pmatrix} \pi_1 & 0 & \dots & 0 \\ 0 & \pi_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \pi_k \end{pmatrix} - \begin{pmatrix} \pi_1^2 & \dots & \pi_1 \pi_k \\ \pi_2 \pi_1 & \dots & \pi_2 \pi_k \\ \vdots & & \vdots \\ \pi_k \pi_1 & \dots & \pi_k^2 \end{pmatrix}. \quad (9.5)$$

Furthermore, it is obvious that  $\boldsymbol{\pi} \mapsto d^2(\boldsymbol{\pi}, \boldsymbol{\pi}^\circ)$  is totally differentiable on  $\boldsymbol{\Pi}$  with gradient vector  $\nabla d^2(\boldsymbol{\pi}, \boldsymbol{\pi}^\circ) = 2(\boldsymbol{\pi} - \boldsymbol{\pi}^\circ)$ . These facts imply that we can exploit once more the so-called  $\delta$ -method to derive the asymptotic distribution of the estimated squared distance  $d^2(\hat{\boldsymbol{\pi}}, \boldsymbol{\pi}^\circ)$  between true and target parameter vector. Writing

$$\sigma_a^2 [\sqrt{n} d^2(\hat{\boldsymbol{\pi}}, \boldsymbol{\pi}^\circ)] = 4(\boldsymbol{\pi} - \boldsymbol{\pi}^\circ)(\mathbf{D}_{\boldsymbol{\pi}} - \boldsymbol{\pi}' \boldsymbol{\pi})(\boldsymbol{\pi} - \boldsymbol{\pi}^\circ)', \quad (9.6)$$

we may thus conclude that

$$\sqrt{n}(d^2(\hat{\boldsymbol{\pi}}, \boldsymbol{\pi}^\circ) - d^2(\boldsymbol{\pi}, \boldsymbol{\pi}^\circ)) \xrightarrow{\mathcal{L}} Z \sim \mathcal{N}\left(0, \sigma_a^2 [\sqrt{n} d^2(\hat{\boldsymbol{\pi}}, \boldsymbol{\pi}^\circ)]\right) \quad \text{as } n \rightarrow \infty. \quad (9.7)$$

Next, the expression on the right-hand side of (9.6) is readily expanded into  $4\left[\sum_{j=1}^k(\pi_j - \pi_j^\circ)^2\pi_j - \sum_{j_1=1}^k\sum_{j_2=1}^k(\pi_{j_1} - \pi_{j_1}^\circ)(\pi_{j_2} - \pi_{j_2}^\circ)\pi_{j_1}\pi_{j_2}\right]$  which is obviously a polynomial in  $(\pi_1, \dots, \pi_k)$ . Since  $\hat{\pi}_j$  is a consistent estimator of  $\pi_j$  for each  $j = 1, \dots, k$ , the asymptotic variance (9.6) can be consistently estimated by replacing in this latter expression each cell probability with the homologous relative frequency. Denoting the resulting variance estimator by  $v_n^2(\hat{\pi}, \pi^\circ)$ , we have

$$\begin{aligned} v_n^2(\hat{\pi}, \pi^\circ) = 4 & \left[ \sum_{j=1}^k (\hat{\pi}_j - \pi_j^\circ)^2 \hat{\pi}_j - \right. \\ & \left. \sum_{j_1=1}^k \sum_{j_2=1}^k (\hat{\pi}_{j_1} - \pi_{j_1}^\circ)(\hat{\pi}_{j_2} - \pi_{j_2}^\circ) \hat{\pi}_{j_1} \hat{\pi}_{j_2} \right]. \end{aligned} \quad (9.8)$$

In view of (9.7) and the consistency of  $v_n^2(\hat{\pi}, \pi^\circ)$  for  $\sigma_a^2[\sqrt{n}d^2(\hat{\pi}, \pi^\circ)]$ , an asymptotically valid test for the goodness-of-fit problem (9.2) is obtained by treating  $d^2(\hat{\pi}, \pi^\circ)$  as a normally distributed statistic with known variance  $n^{-1}v_n^2(\hat{\pi}, \pi^\circ)$  and unknown expected value  $\theta$ , say, about which the one-sided null hypothesis  $\theta \geq \varepsilon^2$  has been formulated. Accordingly, an asymptotic solution to (9.2) can be based on the following decision procedure:

$$\begin{aligned} \text{Reject lack of fit of } & \mathcal{M}(n; \pi_1, \dots, \pi_k) \text{ to } \mathcal{M}(n; \pi_1^\circ, \dots, \pi_k^\circ) \\ \text{iff it turns out that } & d^2(\hat{\pi}, \pi^\circ) < \varepsilon^2 - u_{1-\alpha} v_n(\hat{\pi}, \pi^\circ)/\sqrt{n}, \end{aligned} \quad (9.9)$$

with  $u_{1-\alpha} = \Phi^{-1}(1 - \alpha)$ .

### Example 9.1

For the purpose of illustrating the use of the testing procedure derived in this section, it is tempting to revisit the elementary classroom example of assessing the compatibility of the results of casting the same play dice independently a given number  $n$  of times, with the Laplacean “equal likelihood” definition of probability. Suppose the number of casts performed in such an experiment was  $n = 100$ , and the following frequencies of the possible elementary outcomes were recorded:

Table 9.1 Absolute frequencies observed in a sequence of  $n = 100$  casts of a given play dice.

$j$	1	2	3	4	5	6	$\sum$
$x_j$	17	16	25	9	16	17	100

With these data, the traditional  $\chi^2$ -test for lack of fit yields 7.7600 as the observed value of the test statistic. The associated p-value, i.e., the upper

tail-probability of 7.7600 under a central  $\chi^2$ -distribution with 5 degrees of freedom, is computed to be  $p = .16997$ , indicating that the observed frequencies do not deviate significantly from those expected for an ideal dice. But is this really sufficient for asserting positively that the dice under assessment is approximately fair?

In order to obtain a conclusive answer to this question, we perform the testing procedure defined by (9.9), at the conventional significance level  $\alpha = .05$  and with tolerance  $\varepsilon = .15$  which corresponds in the binomial case to a theoretical equivalence interval of length  $\approx 2 \cdot .10$ . The necessary computational steps are readily carried out running the program `gofsimp`, another SAS macro provided in the **WKTSEQ2 Source Code Package**. (An R script serving the same purpose can be found there as well.) The squared distance between the empirical distribution shown in Table 9.1 and the theoretical distribution given by  $\pi_j^\circ = 1/6 \forall j = 1, \dots, 6$ , turns out to be  $d^2(\hat{\boldsymbol{\pi}}, \boldsymbol{\pi}^\circ) = .012933$  with an estimated standard error of  $n^{-1/2} v_n(\hat{\boldsymbol{\pi}}, \boldsymbol{\pi}^\circ) = .009200$ . Hence, with the suggested specifications of  $\alpha$  and  $\varepsilon$ , the critical upper bound which the observed value of  $d^2(\hat{\boldsymbol{\pi}}, \boldsymbol{\pi}^\circ)$  has to be compared to according to (9.9), is  $.15^2 - 1.644854 \cdot .009200 = .007367$ . Since the observed value of the test statistic clearly exceeds this bound, at the 5% level the data of Table 9.1 do not allow us to decide in favor of the (alternative) hypothesis that the dice under assessment is approximately fair. Thus, the example gives a concrete illustration of the basic general fact (obvious enough from a theoretical viewpoint) that the traditional  $\chi^2$ -test of goodness of fit to a fully specified multinomial distribution is inappropriate for *establishing* the hypothesis of (approximate) fit of the true to the prespecified distribution.

#### *Some simulation results on the finite-sample behavior of the test*

As is the case for every multiparameter problem with a composite null hypothesis specifying an uncountable subspace of parameter values, performing a high-resolution search for the maximum exact rejection probability on the common boundary of the hypotheses of (9.2) or, still more, on  $H$  as a whole in a simulation study would entail a tremendous computational effort. Therefore, for the purpose of throwing some light on the effective size of the critical region of the test (9.9) when applied to finite samples, we confine ourselves to present simulated rejection probabilities only under a small yet sufficiently diversified selection of parameter configurations  $\boldsymbol{\pi} \in \Pi$  with  $d(\boldsymbol{\pi}, \boldsymbol{\pi}^\circ) = \varepsilon$ . Tables 9.2a and b show such a representative set of rejection probabilities simulated under the null hypothesis for the “equal likelihood” problem with  $k = 4$  and  $k = 6$  categories, respectively, for the same specification of  $\alpha$  and  $\varepsilon$  as in the above dice-casting example.

Table 9.2a *Simulated exact rejection probability of the test (9.9) at nominal level  $\alpha = .05$  with  $n = 100$  observations, under selected null configurations for  $k = 4$ ,  $\pi_j^\circ = 1/4 \forall j = 1, \dots, 4$ , and  $\varepsilon = .15$  [40,000 replications of each Monte Carlo experiment].*

$\pi_1$	$\pi_2$	$\pi_3$	$\pi_4$	Rejection Probability
.28879	.13000	.25000	.33121	.03025
.29655	.15000	.21000	.34345	.04250
.21029	.20000	.21000	.37971	.06513
.25388	.22000	.16000	.36612	.05098
.14393	.25000	.25000	.35607	.04143
.17500	.32500	.17500	.32500	.04540

Table 9.2b *Analogue of Table 9.2a for  $k = 6$ .*

$\pi_1$	$\pi_2$	$\pi_3$	$\pi_4$	$\pi_5$	$\pi_6$	Rejection Probability
.21057	.05000	.15000	.15000	.20000	.23943	.00835
.12296	.05000	.20000	.20000	.20000	.22704	.00833
.17296	.10000	.10000	.15000	.20000	.27704	.02685
.10283	.15000	.10000	.15000	.25000	.24717	.02575
.10211	.15000	.15000	.15000	.15000	.29789	.03943
.10543	.22790	.10543	.22790	.10543	.22790	.02113

The results displayed in the above pair of tables suggest that, given the sample size, the signed difference between nominal significance level and exact size of the critical region increases quite rapidly with the number of cells of the underlying contingency table. In fact, whereas for  $k = 4$  the tabulated rejection probabilities indicate some tendency towards anticonservatism (whose extent seems to cause little harm for most practical purposes, to be sure), under the conditions of Example 9.1 the test fails to exhaust the target significance level by about 1%. Clearly, both of these undesirable effects can be removed at least in part by adjusting the nominal level to the maximum admissible degree, but working out the details is beyond our scope here.

Tables 9.3 a and b give results from simulation experiments performed in order to study the second basic property of the test (9.9), namely, the relationship between its power against the specific alternative of perfect coincidence between true and model-derived distribution of  $(X_1, \dots, X_k)$ , and size  $n$  of the available sample. These latter data point to an interesting relationship corroborated by the results of additional simulation experiments with

Table 9.3a *Simulated exact power  $POW_\circ$  of the test (9.9) at nominal level  $\alpha = .05$  against the alternative that  $\pi_j = \pi_j^\circ = 1/k \ \forall j = 1, \dots, k$ , for  $k = 4$ ,  $\varepsilon = .15$  and  $n = 50(25)200$  [40,000 replications of each experiment].*

$n$	50	75	100	125	150	175	200
$POW_\circ$	.23798	.44120	.63673	.77330	.87213	.92868	.96388

Table 9.3b *Analogue of Table 9.3a for  $k = 6$ .*

$n$	50	75	100	125	150	175	200
$POW_\circ$	.16030	.42943	.66795	.83120	.92620	.96913	.98710

other values of  $k$ : The larger the number of categories, the steeper the increase of the power against  $\boldsymbol{\pi} = \boldsymbol{\pi}^\circ$  as a function of  $n$ . Both in the tetra- and the hexanomial case,  $n = 150$  observations turn out sufficient to guarantee satisfactory power of the test (9.9) against the specific alternative of primary interest.

## 9.2 Testing for approximate collapsibility of multiway contingency tables

The concept of collapsibility plays an important role within the literature on log-linear models for multiway contingency tables (cf. Bishop et al., 1975, § 2.5). The precise meaning of the term in its strict sense is as follows: Let  $(X_{1\dots 1}, \dots, X_{k_1\dots k_q})$  be the vector of frequencies counted in the  $\prod_{\nu=1}^q k_\nu$  cells of a higher-order contingency table formed by cross-classifying each of  $n$  sampling units with respect to  $q$  categorical variables  $C_1, \dots, C_q$  with possible values  $\{1, \dots, k_1\}, \dots, \{1, \dots, k_q\}$ , respectively. Then, the set  $\{C_1, \dots, C_q\}$  is called strictly collapsible across  $C_\nu$  (for prespecified  $\nu \in \{1, \dots, q\}$ ) if the log-linear model for the reduced table  $(X_{1\dots +\dots 1}, \dots, X_{k_1\dots k_{\nu-1}+k_{\nu+1}\dots k_q})$  obtained by taking for each  $(i_1, \dots, i_{\nu-1}, i_{\nu+1}, \dots, i_q)$  the sum  $\sum_{i_\nu=1}^{k_\nu} X_{i_1\dots i_{\nu-1}i_\nu i_{\nu+1}\dots i_q}$  contains exactly the same parameters as that for the original table, except for those related to  $C_\nu$ . It can be shown (see Bishop et al., 1975, p. 47) that this holds true if and only if  $C_\nu$  is independent of  $\{C_1, \dots, C_{\nu-1}, C_{\nu+1}, \dots, C_q\}$ . Accordingly, strict collapsibility of a set of categorical variables defining a given multiway contingency table is a special case of independence of a couple

of such variables, so that the general problem raised in the title of this section can be solved by constructing a test for approximate independence of two categorical variables A and B say, with possible values  $1, \dots, r$  and  $1, \dots, s$ , respectively.

As usual, we denote by  $\pi_{ij}$  the probability that a specific sampling unit is observed to belong to cell  $(i, j)$  of the corresponding two-way table, and  $X_{ij}$  stands for the number of sampling units counted in that cell. Of course, independence of the two characteristics A and B holds if and only if we may write  $\pi_{ij} = \pi_{i+} \pi_{+j} \equiv (\sum_{\nu=1}^s \pi_{i\nu}) (\sum_{\mu=1}^r \pi_{\mu j}) \forall (i, j) \in \{1, \dots, r\} \times \{1, \dots, s\}$ . Thus, the target distribution to which the true distribution of the primary observations is asserted to be sufficiently similar under any hypothesis formalizing the notion of approximate independence, depends on the unknown parameters  $\pi_{ij}$  rather than being completely specified. Nevertheless, it will turn out to be comparatively easy to accommodate the approach of the preceding section to this more complicated situation.

In order to give the details of the construction, we have to adapt the notation for the primary parameter vector by defining  $\boldsymbol{\pi}$  as the vector of dimension  $r \times s$  formed by putting the rows of the matrix  $(\pi_{ij})_{(r,s)}$  of cell probabilities one after the other. Furthermore, it will be convenient to use the following extra symbols for additional parametric functions to be referred to repeatedly in defining the hypotheses, as well as in deriving a test statistic upon which an asymptotically valid test for approximate independence of A and B can be based:

$$\varrho_i(\boldsymbol{\pi}) \equiv \sum_{\nu=1}^s \pi_{i\nu}, \quad i = 1, \dots, r; \quad (9.10a)$$

$$\zeta_j(\boldsymbol{\pi}) \equiv \sum_{\mu=1}^r \pi_{\mu j}, \quad j = 1, \dots, s; \quad (9.10b)$$

$$\begin{aligned} \mathbf{g}(\boldsymbol{\pi}) \equiv & (\varrho_1(\boldsymbol{\pi}) \zeta_1(\boldsymbol{\pi}), \dots, \varrho_1(\boldsymbol{\pi}) \zeta_s(\boldsymbol{\pi}), \\ & \dots, \varrho_r(\boldsymbol{\pi}) \zeta_1(\boldsymbol{\pi}), \dots, \varrho_r(\boldsymbol{\pi}) \zeta_s(\boldsymbol{\pi})) . \end{aligned} \quad (9.10c)$$

By analogy with the way of approaching the problem of testing for consistency with a fully specified distribution considered in § 9.1, it seems sensible to base the criterion for approximate independence of the classifications under consideration on the squared Euclidean distance

$$d^2(\boldsymbol{\pi}, \mathbf{g}(\boldsymbol{\pi})) = \sum_{i=1}^r \sum_{j=1}^s (\pi_{ij} - \varrho_i(\boldsymbol{\pi}) \zeta_j(\boldsymbol{\pi}))^2 \quad (9.11)$$

between the vector  $\boldsymbol{\pi}$  of true cell probabilities and the associated vector  $\mathbf{g}(\boldsymbol{\pi})$  of products of marginal totals. Again, a natural estimator of that distance is obtained by substituting the vector  $\hat{\boldsymbol{\pi}}$  of observed relative frequencies for  $\boldsymbol{\pi}$  throughout, where the components of  $\hat{\boldsymbol{\pi}}$  have this time to be written  $\hat{\pi}_{ij} =$

$X_{ij}/n$ . Except for obvious changes in notation, the asymptotic distribution of  $\sqrt{n}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi})$  is the same as in the one-dimensional case [recall (9.5)]. This implies that  $\sqrt{n}(d^2(\hat{\boldsymbol{\pi}}, \mathbf{g}(\hat{\boldsymbol{\pi}})) - d^2(\boldsymbol{\pi}, \mathbf{g}(\boldsymbol{\pi}))$  is asymptotically normal with expectation 0 and variance  $\sigma_a^2[\sqrt{n}d^2(\hat{\boldsymbol{\pi}}, \mathbf{g}(\hat{\boldsymbol{\pi}}))]$ , say, which can be computed by means of the  $\delta$ -method as

$$\begin{aligned} \sigma_a^2[\sqrt{n}d^2(\hat{\boldsymbol{\pi}}, \mathbf{g}(\hat{\boldsymbol{\pi}}))] &= \sum_{i=1}^r \sum_{j=1}^s \left[ \frac{\partial}{\partial \pi_{ij}} d^2(\boldsymbol{\pi}, \mathbf{g}(\boldsymbol{\pi})) \right]^2 \pi_{ij} \\ &- \sum_{i_1=1}^r \sum_{j_1=1}^s \sum_{i_2=1}^r \sum_{j_2=1}^s \frac{\partial}{\partial \pi_{i_1 j_1}} d^2(\boldsymbol{\pi}, \mathbf{g}(\boldsymbol{\pi})) \frac{\partial}{\partial \pi_{i_2 j_2}} d^2(\boldsymbol{\pi}, \mathbf{g}(\boldsymbol{\pi})) \pi_{i_1 j_1} \pi_{i_2 j_2}. \end{aligned} \quad (9.12)$$

For the purpose of deriving explicit expressions for the individual components of the gradient of  $\boldsymbol{\pi} \mapsto d^2(\boldsymbol{\pi}, \mathbf{g}(\boldsymbol{\pi}))$ , we start with determining the partial derivatives of the  $(i, j)$ th term of the double sum appearing on the right-hand side of (9.11) [for arbitrarily fixed  $(i, j) \in \{1, \dots, r\} \times \{1, \dots, s\}$ ] with respect to  $\pi_{\mu\nu}$  for  $\mu = 1, \dots, r$ ,  $\nu = 1, \dots, s$ . From (9.10a), (9.10b) it is obvious that we have

$$\frac{\partial(\varrho_i(\boldsymbol{\pi})\zeta_j(\boldsymbol{\pi}))}{\partial \pi_{\mu\nu}} = \begin{cases} 0 & \mu \neq i, \nu \neq j \\ \zeta_j(\boldsymbol{\pi}) & \text{for } \mu = i, \nu \neq j \\ \varrho_i(\boldsymbol{\pi}) & \mu \neq i, \nu = j \\ \varrho_i(\boldsymbol{\pi}) + \zeta_j(\boldsymbol{\pi}) & \mu = i, \nu = j \end{cases}. \quad (9.12a)$$

Using (9.12a), it is readily verified that

$$\begin{aligned} \frac{1}{2} \frac{\partial(\pi_{ij} - \varrho_i(\boldsymbol{\pi})\zeta_j(\boldsymbol{\pi}))^2}{\partial \pi_{\mu\nu}} &= \\ \begin{cases} 0 & \mu \neq i, \nu \neq j \\ -(\pi_{ij} - \varrho_i(\boldsymbol{\pi})\zeta_j(\boldsymbol{\pi}))\zeta_j(\boldsymbol{\pi}) & \text{for } \mu = i, \nu \neq j \\ -(\pi_{ij} - \varrho_i(\boldsymbol{\pi})\zeta_j(\boldsymbol{\pi}))\varrho_i(\boldsymbol{\pi}) & \mu \neq i, \nu = j \\ (\pi_{ij} - \varrho_i(\boldsymbol{\pi})\zeta_j(\boldsymbol{\pi}))(1 - \varrho_i(\boldsymbol{\pi}) - \zeta_j(\boldsymbol{\pi})) & \mu = i, \nu = j \end{cases} \end{aligned} \quad (9.12b)$$

Summing (9.12b) over  $(i, j) \in \{1, \dots, r\} \times \{1, \dots, s\}$  gives:

$$\begin{aligned} \frac{\partial}{\partial \pi_{\mu\nu}} d^2(\boldsymbol{\pi}, \mathbf{g}(\boldsymbol{\pi})) &= 2 \left\{ (\pi_{\mu\nu} - \varrho_\mu(\boldsymbol{\pi})\zeta_\nu(\boldsymbol{\pi})) - \sum_{i=1}^r \left[ (\pi_{i\nu} - \varrho_i(\boldsymbol{\pi})\zeta_\nu(\boldsymbol{\pi}))\varrho_i(\boldsymbol{\pi}) \right] - \sum_{j=1}^s \left[ (\pi_{\mu j} - \varrho_\mu(\boldsymbol{\pi})\zeta_j(\boldsymbol{\pi}))\zeta_j(\boldsymbol{\pi}) \right] \right\}. \end{aligned} \quad (9.13)$$

Of course, what we eventually need is a consistent estimator, say  $v_n^2(\hat{\boldsymbol{\pi}}, \mathbf{g}(\hat{\boldsymbol{\pi}}))$ , of  $\sigma_a^2[\sqrt{n}d^2(\hat{\boldsymbol{\pi}}, \mathbf{g}(\hat{\boldsymbol{\pi}}))]$ . In view of (9.12) and (9.13), for achieving this goal, it is sufficient to plug in the homologous relative frequency at every

place on the right-hand side of both equations where a population parameter, i.e., a component of the vector  $\boldsymbol{\pi}$  appears. More precisely speaking, for any  $(\mu, \nu) \in \{1, \dots, r\} \times \{1, \dots, s\}$ , the  $(\mu, \nu)$ -component of the gradient vector of  $\boldsymbol{\pi} \mapsto d^2(\boldsymbol{\pi}, \mathbf{g}(\boldsymbol{\pi}))$  can be consistently estimated by

$$\hat{d}_{\mu\nu} \equiv 2 \left\{ (\hat{\pi}_{\mu\nu} - \varrho_\mu(\hat{\boldsymbol{\pi}})\zeta_\nu(\hat{\boldsymbol{\pi}})) - \sum_{i=1}^r [(\hat{\pi}_{i\nu} - \varrho_i(\hat{\boldsymbol{\pi}})\zeta_\nu(\hat{\boldsymbol{\pi}})) \cdot \right. \\ \left. \varrho_i(\hat{\boldsymbol{\pi}})] - \sum_{j=1}^s [(\hat{\pi}_{\mu j} - \varrho_\mu(\hat{\boldsymbol{\pi}})\zeta_j(\hat{\boldsymbol{\pi}}))\zeta_j(\hat{\boldsymbol{\pi}})] \right\}. \quad (9.14)$$

Finally, the desired estimator of the asymptotic variance of  $\sqrt{n} d^2(\hat{\boldsymbol{\pi}}, \mathbf{g}(\hat{\boldsymbol{\pi}}))$  is obtained by substituting  $\hat{d}_{11}, \dots, \hat{d}_{rs}$  and  $\hat{\pi}_{11}, \dots, \hat{\pi}_{rs}$  for  $\frac{\partial}{\partial \pi_{11}} d^2(\boldsymbol{\pi}, \mathbf{g}(\boldsymbol{\pi})), \dots, \frac{\partial}{\partial \pi_{rs}} d^2(\boldsymbol{\pi}, \mathbf{g}(\boldsymbol{\pi}))$  and  $\pi_{11}, \dots, \pi_{rs}$ , respectively, on the right-hand side of (9.12), yielding

$$v_n^2(\hat{\boldsymbol{\pi}}, \mathbf{g}(\hat{\boldsymbol{\pi}})) = \sum_{i=1}^r \sum_{j=1}^s \hat{d}_{ij}^2 \hat{\pi}_{ij} - \\ \sum_{i_1=1}^r \sum_{j_1=1}^s \sum_{i_2=1}^r \sum_{j_2=1}^s \hat{d}_{i_1 j_1} \hat{d}_{i_2 j_2} \hat{\pi}_{i_1 j_1} \hat{\pi}_{i_2 j_2}. \quad (9.15)$$

Now we are ready to formulate the decision rule of an asymptotically valid test of

$$H : d^2(\boldsymbol{\pi}, \mathbf{g}(\boldsymbol{\pi})) \geq \varepsilon^2 \text{ versus } K : d^2(\boldsymbol{\pi}, \mathbf{g}(\boldsymbol{\pi})) < \varepsilon^2 \quad (9.16)$$

for fixed  $\varepsilon > 0$  to be chosen *a priori*. The alternative hypothesis we want to establish according to (9.16) specifies that the true distribution is in acceptably close agreement with that hypothesized by the model of independence given both vectors of marginal probabilities, which seems to be an adequate translation of the notion of approximate validity of that model into a statement about the parameter characterizing the distribution from which the data are taken. After the technical preparations made in the preceding paragraphs, we know that asymptotically, the estimated squared distance of  $\boldsymbol{\pi}$  from  $\mathbf{g}(\boldsymbol{\pi})$  can be viewed as a normally distributed variable with  $d^2(\boldsymbol{\pi}, \mathbf{g}(\boldsymbol{\pi}))$  as unknown expected value and fixed known variance  $n^{-1} v_n^2(\hat{\boldsymbol{\pi}}, \mathbf{g}(\hat{\boldsymbol{\pi}}))$ . Hence, we get a test of asymptotic level  $\alpha$  for the problem (9.16) by applying the decision rule:

Reject existence of relevant discrepancies between true distribution and independence model iff

$$d^2(\hat{\boldsymbol{\pi}}, \mathbf{g}(\hat{\boldsymbol{\pi}})) < \varepsilon^2 - u_{1-\alpha} v_n(\hat{\boldsymbol{\pi}}, \mathbf{g}(\hat{\boldsymbol{\pi}})) / \sqrt{n}. \quad (9.17)$$

*Example 9.2*

In one out of four different arms of a large multicenter trial of drugs used as first-choice medication for mild arterial hypertension, classifying the patients by gender and outcome of short-term treatment gave the  $2 \times 4$  contingency table shown below as Table 9.4. It covers all patients who were randomized for administration of the calcium-entry blocking agent nitrendipine and treated in compliance with the study protocol. At a certain stage of their work, the statisticians responsible for the analysis of the results of the trial were confronted also with the problem of identifying those treatment arms for which a relevant gender-by-outcome interaction could be excluded. For the nitrendipine group, this question can be suitably addressed by applying decision rule (9.17) to the data of Table 9.4. Once more, we set the significance level to  $\alpha = .05$ , and specify  $\varepsilon = .15$  as the largest tolerable distance between the vector of true cell probabilities and the associated vector of products of marginal probabilities.

Table 9.4  $2 \times 4$  contingency table relating gender and treatment outcome on nitrendipine monotherapy in patients suffering from mild arterial hypertension. [Source: Philipp et al. (1997); Numbers in ( ): relative frequencies with respect to the overall total  $n$ .]

		Outcome Category				$\sum$
		1*)	2**)	3***)	4†)	
Gender	Female	9 (.042)	13 (.060)	13 (.060)	48 (.221)	83 (.382)
	Male	24 (.111)	18 (.083)	20 (.092)	72 (.332)	134 (.618)
	$\sum$	33 (.152)	31 (.143)	33 (.152)	120 (.553)	217 (1.000)

\*)...\*\*\*) Response at lowest, middle and highest dose step, respectively

†) Failure to reach target level of diastolic blood pressure at all three dose steps

With  $X_{11} = 9, \dots, X_{24} = 72$ , the estimated distance between  $\boldsymbol{\pi} = (\pi_{11}, \dots, \pi_{24})$  and  $\mathbf{g}(\boldsymbol{\pi}) = (\pi_{1+} + \pi_{+1}, \dots, \pi_{2+} + \pi_{+4})$  is readily computed to be  $d(\hat{\boldsymbol{\pi}}, \mathbf{g}(\hat{\boldsymbol{\pi}})) = .028396$ . Evaluating formula (9.15) for the squared standard error of the statistic  $\sqrt{n} d^2(\hat{\boldsymbol{\pi}}, \mathbf{g}(\hat{\boldsymbol{\pi}}))$  by hand or by means of a pocket calculator is a rather tedious process, even if the total number of cells of the contingency table under analysis is comparatively small. So we leave this job to another computer program ready for being downloaded from **WKSHEQ2 Source Code Package**. Once more, it exists both in a SAS and a R version, with `gofind_t`

as program name. Running this macro, we read from the output  $v_n = .017016$ . Thus, at the 5%-level and with  $\varepsilon = .15$ , the critical upper bound which the observed value of  $d^2(\hat{\pi}, g(\hat{\pi}))$  has to be compared to, is computed to be  $.15^2 - 1.644854 \cdot .017016 / \sqrt{217} = .020600$ . Since  $.028396^2 = .000806$  is clearly smaller than this critical bound, we have to decide in favor of approximate independence between gender and treatment outcome in hypertensive patients taking nitrendipine according to the prescriptions of the protocol of the trial from which the data of Table 9.4 are taken (for details see Philipp et al., 1997).

### *Example 9.3*

In the light of the introductory remarks at the beginning of this section, we now turn to illustrating the use of decision rule (9.17) for testing for approximate collapsibility in the strict sense of a higher-order contingency table across a prespecified element of the set  $\{C_1, \dots, C_q\}$  of categorical variables involved. For this purpose, we confine ourselves to a brief discussion of a synthetic data set. Suppose the format of the table under analysis is  $2 \times 2 \times 2$  with observed cell frequencies as shown below, and the variable with respect to which we aim to establish collapsibility, is the first one. Suppose further that  $\alpha$  and  $\varepsilon$  have been specified as in the previous example.

Table 9.5a *Observed  $2 \times 2 \times 2$  table to be tested for approximate collapsibility in the strict sense across the first binary classification.*

		$C_2$	$C_3$	
		+	-	
		+	8	13
$C_1$		-	15	6
		+	19	21
		-	31	7

In a preliminary step we have to rearrange the original array in a  $2 \times 4$  table with the combination of  $C_2$  and  $C_3$  defining the columns. Using the entries in the rearranged Table 9.5b as input data to the program `gofind_t` we obtain  $d(\hat{\pi}, g(\hat{\pi})) = .030332$  and  $v_n = .019190$ . Hence, the condition for rejecting the null hypothesis of relevant deviations from collapsibility is in the present case that  $d^2(\hat{\pi}, g(\hat{\pi})) < .15^2 - 1.644854 \times .019190 / \sqrt{120} = .019619$  which in view of  $.030332^2 < .001$  is clearly satisfied.

Table 9.5b *Rearrangement of Table 9.5a in a  $2 \times 4$  array.*

		B ( $\leftrightarrow C_2 \times C_3$ )				$\sum$	
		1	2	3	4		
A ( $\leftrightarrow C_1$ )	1	8	13	15	6	42	
	2	19	21	31	7	78	
		$\sum$	27	34	46	13	120

*Simulation results on size and power*

Table 9.6 is an analogue of Tables 9.2 a, b for the testing procedure (9.17) as applied to a  $2 \times 4$  contingency table containing roughly the same total number of observations as had been available in the real application described in Example 9.2. Of course, for the problem of testing for approximate independence of two categorical variables, a numerically accurate search for the maximum rejection probability on the common boundary of the hypotheses is still farther beyond the scope of computational feasibility than for the problem of establishing goodness of fit to a fully specified multinomial distribution. In fact, in the independence case, the alternative hypothesis is an uncountable union of hyperspheres of radius  $\varepsilon$  rather than a single set of that form. Accordingly, the simulation results displayed in the rightmost column of the next table allow only a rough qualitative assessment of the finite-sample behavior of the testing procedure (9.17) under the null hypothesis  $H$  of (9.16).

Table 9.6 *Simulated exact rejection probability of the test (9.17) at nominal level  $\alpha = .05$ , with  $r = 2$ ,  $s = 4$ ,  $\varepsilon = .15$  and  $n = 200$ , under various parameter configurations on the common boundary of the hypotheses (9.16) [40,000 replications per Monte Carlo experiment].*

$\pi_{11}$	$\pi_{12}$	$\pi_{13}$	$\pi_{14}$	$\pi_{21}$	$\pi_{22}$	$\pi_{23}$	$\pi_{24}$	P [Rejection]
.05	.05	.05	.05	.65923	.05	.05	.04077	.07153
.05	.05	.05	.10	.48723	.05	.15	.06277	.06718
.05	.15	.10	.35	.07878	.05	.15	.07122	.06218
.10	.05	.25	.15	.18082	.10	.05	.11918	.05405
.10	.10	.10	.15	.24850	.10	.20	.00150	.05680
.05	.10	.15	.20	.23229	.05	.10	.11771	.05803
.25	.25	.05	.05	.04030	.20	.15	.00970	.05615
.15	.15	.15	.15	.25308	.05	.05	.04692	.06573
.20	.05	.05	.05	.16601	.30	.15	.03399	.06993
.20	.05	.05	.20	.27247	.05	.15	.02753	.05895

Again we find some anticonservative tendency whose extent seems small enough for being ignored in the majority of practical applications, but is clearly outside the range of deviations which can be accounted for by the simulation error. Since the results of additional simulation experiments not reproduced in detail here suggest that the convergence of the exact size of the asymptotic test to the target significance level is quite slow, it is recommended to rely on an adjusted nominal level  $\alpha^* < \alpha$  whenever strict maintenance of the prespecified level is felt to be an indispensable requirement. For instance, replacing  $\alpha = .05$  with  $\alpha^* = .025$  suffices for ensuring that even for the least favorable of the parameter constellations covered by Table 9.6 [ $\rightarrow$  row 1], the (simulated) exact rejection probability with a sample of size  $n = 200$  no longer exceeds the 5%-bound.

For completeness, we finish our discussion of (asymptotic) testing procedures tailored for establishing the approximate validity of customary models for contingency tables, with presenting some simulation results on the power of the procedure (9.17) to detect that the true parameter configuration  $\pi$  fits the model of independence exactly. Obviously, given the format of the two-way tables to be analyzed, there is an uncountable manifold of specific alternatives of that type. Singling out once more the case  $(r, s) = (2, 4)$ , we studied the power under three configurations of marginal probabilities  $\pi_{1+}, \pi_{2+}; \pi_{+1}, \dots, \pi_{+4}$  determining extremely different joint distributions of the  $X_{ij}$ . Table 9.7 shows the rejection probabilities simulated under these parameter configurations for  $n \in \{50, 100, 150\}$ ,  $\varepsilon = .15$  and nominal levels  $\alpha^*$  which, according to our results on the size of the test (9.17), are small enough to ensure that the target level of 5% is maintained at least at all points on the common boundary of the hypotheses covered by Table 9.6.

Table 9.7 *Simulated exact power of the test (9.17) against three alternatives satisfying  $\pi_{ij} = \pi_{i+}\pi_{+j} \forall (i, j)$  for  $2 \times 4$  arrays, with  $\varepsilon = .15$ ,  $n = 50, 100, 150$  and corrected nominal significance level  $\alpha^*$  [40,000 replications per Monte Carlo experiment].*

$n$	$\alpha^*$	$\pi_{1+}$	$\pi_{2+}$	$\pi_{+1}$	$\pi_{+2}$	$\pi_{+3}$	$\pi_{+4}$	P [Rejection]
50	.01	.25	.75	.10	.40	.30	.20	.39788
"	"	.50	.50	.25	.25	.25	.25	.25395
"	"	.33	.67	.15	.15	.15	.55	.43115
100	.02	.25	.75	.10	.40	.30	.20	.87578
"	"	.50	.50	.25	.25	.25	.25	.84405
"	"	.33	.67	.15	.15	.15	.55	.90630
150	.025	.25	.75	.10	.40	.30	.20	.97925
"	"	.50	.50	.25	.25	.25	.25	.98458
"	"	.33	.67	.15	.15	.15	.55	.98758

Not surprisingly, the power against alternatives  $\pi$  such that  $d(\pi, g(\pi))$  vanishes turns out to be very sensitive against gross changes to the shape of the marginal distributions. Furthermore, even if the nominal significance level  $\alpha^*$  is downgraded as far as necessary for ensuring that the rejection probability keeps below the target level of 5% under all null constellations studied for assessing the level properties of the test, a sample size of 100 seems sufficient for achieving reasonable power against alternatives satisfying the model of independence exactly.

---

### 9.3 Establishing goodness of fit of linear models for normally distributed data

#### 9.3.1 An exact optimal test for negligibility of interactions in a two-way ANOVA layout

Presumably, more often than not, existence of interaction effects will be considered as an undesirable lack of fit to the model which an investigator applying standard ANOVA techniques would like to rely upon. The obvious reason is that in presence of interactions, the main effects do not admit a direct interpretation. In fact, only in absence of interactions, a positive main effect associated with some level of a given factor justifies the conclusion that on an average, all experimental units assigned to that level do better than those assigned to another level with a negative effect of the same factor. Adopting this perspective, it is natural to consider that part of the analysis of an ANOVA two-way layout which deals with the interaction effects, as a preliminary check carried out with the intention to establish that the “ideal” additive model fits the data sufficiently well.

For the moment, we restrict the description of a testing procedure, which allows us to exclude relevant deviations from additivity of the main effects, to the case of a strictly balanced design. In other words, we assume that the data set under analysis consists of  $r \cdot s$  independent samples of common size  $n$  from normal distributions with the same unknown variance  $\sigma^2 > 0$ . Denoting the  $k$ th observation in group  $(i, j)$  by  $X_{ijk}$ , we suppose that for each  $(i, j) \in \{1, \dots, r\} \times \{1, \dots, s\}$ , the cell mean  $\mu_{ij} = E(X_{ij1}) = \dots = E(X_{ijn})$  can be additively decomposed in the usual way leading to the representation

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} . \quad (9.18)$$

Of course, the parameters coming into play according to this basic model are assumed to satisfy the usual side conditions  $\sum_{i=1}^r \alpha_i = 0 = \sum_{j=1}^s \beta_j$ ,  $\sum_{j=1}^s \gamma_{ij} = 0 \forall i = 1, \dots, r$ ,  $\sum_{i=1}^r \gamma_{ij} = 0 \forall j = 1, \dots, s$ .

Now, it seems reasonable to consider the interaction effects negligible as soon as it can be taken for granted that the standardized  $\gamma_{ij}$  are the coordinates of a point in  $rs$ -dimensional Euclidean space which lies in a spherical neighborhood of  $\mathbf{0}$  of sufficiently small radius  $\varepsilon > 0$ . Accordingly, our alternative hypothesis of absence of relevant interactions specifies that  $\sum_{i=1}^r \sum_{j=1}^s (\gamma_{ij}/\sigma)^2 < \varepsilon^2$ , so that the testing problem as a whole reads

$$H : d^2(\boldsymbol{\gamma}/\sigma, \mathbf{0}) \geq \varepsilon^2 \quad \text{versus} \quad K : d^2(\boldsymbol{\gamma}/\sigma, \mathbf{0}) < \varepsilon^2, \quad (9.19)$$

provided  $\boldsymbol{\gamma}$  stands for the vector consisting of the rows of the matrix  $(\gamma_{ij})_{(r,s)}$ ,  $\mathbf{0}$  for the null vector in  $rs$ -dimensional space, and  $d(\mathbf{u}, \mathbf{v})$  denotes the Euclidean distance between arbitrary points  $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{rs}$ . It is easily verified that this pair of hypotheses remains invariant under the same group of 1 : 1 transformations of the sample space  $\mathbb{R}^{nrs}$  onto itself as the problem of testing the conventional null hypothesis that all interaction parameters vanish. (As is generally the case with problems exhibiting the same invariance structure as a pair of linear hypotheses about the mean of an independent vector of homoskedastic Gaussian variables, a precise characterization of this transformation group requires the introduction of a canonical coordinate system for the observed vector, which would lead the exposition away from what our real issue is here.) Hence, any invariant test of (9.19) is necessarily a function of the classical  $F$ -ratio for testing  $\gamma_{ij} = 0 \forall (i,j)$  vs.  $\gamma_{ij} \neq 0$  for some  $(i,j) \in \{1, \dots, r\} \times \{1, \dots, s\}$ . Discarding the constant factor  $n/(r-1)(s-1)$ , this statistic can be viewed as the squared Euclidean distance from the origin of the least squares estimator of  $\boldsymbol{\gamma}$  standardized with respect to the length of the residual vector for the nonrestricted model as the standard estimator of  $\sigma$ . Using a notation which is to be suggestive of this interpretation of the rescaled  $F$ -ratio, we write

$$d^2(\hat{\boldsymbol{\gamma}}/S, \mathbf{0}) = \frac{\sum_{i=1}^r \sum_{j=1}^s (\bar{X}_{ij} - \bar{X}_{i..} - \bar{X}_{.j} + \bar{X}_{...})^2}{\sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^n (X_{ijk} - \bar{X}_{ij})^2 / (n-1)rs} , \quad (9.20)$$

with  $\bar{X}_{ij} \equiv \sum_{k=1}^n X_{ijk}/n$ ,  $\bar{X}_{i..} \equiv \sum_{j=1}^s \bar{X}_{ij}/s$ ,  $\bar{X}_{.j} \equiv \sum_{i=1}^r \bar{X}_{ij}/r$ , and  $\bar{X}_{...} \equiv \sum_{i=1}^r \sum_{j=1}^s \bar{X}_{ij}/rs$ .

The distribution of  $nd^2(\hat{\boldsymbol{\gamma}}/S, \mathbf{0})/((r-1)(s-1))$  is well known (see, e.g., Lehmann and Romano, 2005, p. 292) to be noncentral  $F$  with  $(r-1)(s-1)$ ,  $(n-1)rs$  degrees of freedom and noncentrality parameter  $\lambda_{nc}^2 = nd^2(\boldsymbol{\gamma}/\sigma, \mathbf{0})$ , under any parameter constellation  $(\mu, \alpha_1, \dots, \alpha_r, \beta_1, \dots, \beta_s, \gamma_{11}, \dots, \gamma_{rs}) \in \mathbb{R}^{(r+1)(s+1)}$ . Since the class of all distributions of that type constitutes a family with strictly monotone likelihood ratios, it follows that the class of all level- $\alpha$  tests of (9.19) based upon  $d^2(\hat{\boldsymbol{\gamma}}/S, \mathbf{0})$  contains a uniformly most powerful element which rejects  $H$  if and only if  $nd^2(\hat{\boldsymbol{\gamma}}/S, \mathbf{0})/((r-1)(s-1))$  turns

out to be less than or equal to the  $\alpha$ -quantile of a  $F$ -distribution with the same numbers of degrees of freedom and noncentrality parameter  $\lambda_{nc}^2 = n\varepsilon^2$ . In view of the maximal invariance of the statistic  $d^2(\hat{\gamma}/S, \mathbf{0})$  for the hypotheses (9.19) we can thus conclude that

$$\left\{ d^2(\hat{\gamma}/S, \mathbf{0}) < ((r-1)(s-1)/n) F_{(r-1)(s-1), (n-1)rs; \alpha}(n\varepsilon^2) \right\} \quad (9.21)$$

is the critical region of an UMPI level- $\alpha$  test for negligibility of interactions in a two-way ANOVA layout with  $rs$  cells. Of course, in order to render this statement true, the threefold subscribed symbol  $F$  must be defined in the same way as in the preceding chapters so that we have

$$\begin{aligned} F_{(r-1)(s-1), (n-1)rs; \alpha}(n\varepsilon^2) = & \text{lower } 100\alpha\text{-percentage point of} \\ & \text{an } F\text{-distribution with } (r-1)(s-1), (n-1)rs \text{ degrees} \\ & \text{of freedom and noncentrality parameter } n\varepsilon^2. \end{aligned} \quad (9.22)$$

For the power of the test with rejection region (9.21) against any specific alternative with some given value, say  $\psi^2$ , of  $d^2(\boldsymbol{\gamma}/\sigma, \mathbf{0})$ , we get a direct analogue of formula (7.12) referring to the noncentral  $F$ -test for equivalence of  $k$  homoskedastic Gaussian distributions as discussed in § 7.2. More precisely speaking, if we denote the rejection probability of the test under any parameter constellation  $(\mu, \alpha_1, \dots, \alpha_r, \beta_1, \dots, \beta_s, \gamma_{11}, \dots, \gamma_{rs})$  such that  $d^2(\boldsymbol{\gamma}/\sigma, \mathbf{0}) = \psi^2$  by  $\beta(\psi^2)$ , then we can write

$$\beta(\psi^2) = P[\mathcal{F}_{(r-1)(s-1), (n-1)rs}(n\psi^2) < F_{(r-1)(s-1), (n-1)rs; l, \alpha}(n\varepsilon^2)] \quad (9.23)$$

with  $\mathcal{F}_{(r-1)(s-1), (n-1)rs}(n\psi^2)$  representing a random variable whose distribution differs from that to be used for determining the critical constant only by changing the value of the noncentrality parameter from  $n\varepsilon^2$  to  $n\psi^2$ .

### *Example 9.4*

In a randomized clinical trial of the antimuscarinic effects (to be brought about for the purpose of reversing scopolamine-induced pupil dilation at the end of cataract surgery) of acetylcholine and thymoxamine, a total of 228 eyes was subdivided into the following four treatment arms:

- (1) placebo (pure solvent)
- (2) acetylcholine
- (3) thymoxamine
- (4) acetylcholine + thymoxamine.

Randomization was strictly uniform so that  $n = 57$  eyes were assigned to each of these treatment groups. The primary endpoint for assessing the efficacy of

all treatments was pupillary constriction [mm] as measured 5 minutes after application. The following table summarizes the data obtained from this trial (as cellwise arithmetic means and standard deviations, respectively):

Table 9.8 *Arithmetic means and standard deviations (in parentheses) of pupillary restriction [mm] measured in the four arms of a clinical trial of the antimuscarinic effects of acetylcholine and thymoxamine following a classical  $2 \times 2$ -design [ $n = 57$  eyes treated in each group]. (Data from Pfeiffer et al., 1994.)*

		Acetylcholine	
		—	+
Thymox-	—	−0.0386 (.2433)	1.8561 (.6690)
	+	1.5281 (.7975)	3.4193 (.9801)

With these values, we calculate  $\bar{X}_{1.} = 0.9088$ ,  $\bar{X}_{.1} = 0.7448$ ,  $\bar{X}_{2.} = 2.4737$ ,  $\bar{X}_{.2} = 2.6377$ ,  $\bar{X}_{..} = 1.6912$  yielding for the interaction parameters the estimates

$$\hat{\gamma}_{11} = -0.0386 - (0.9088 + 0.7448 - 1.6912) = -0.0009,$$

$$\hat{\gamma}_{12} = -\hat{\gamma}_{11} = 0.0009, \hat{\gamma}_{21} = -\hat{\gamma}_{11} = 0.0009 \text{ and } \hat{\gamma}_{22} = \hat{\gamma}_{11} = -0.0009.$$

Furthermore, using the parenthesized entries in Table 9.8, we find that the residual variance of the data set under analysis with respect to the unrestricted model (9.18) is  $S^2 = (.2433^2 + .6690^2 + .7975^2 + .9801^2)/4 = .5258$ . Thus, as estimated squared distance of  $\boldsymbol{\gamma}/\sigma$  from the origin, we find here

$$d^2(\hat{\boldsymbol{\gamma}}/S, \mathbf{0}) = 4 \cdot .0009^2 / .5258 = 6.16 \cdot 10^{-6}.$$

Let us now suppose that the true interaction effects are to be considered negligible if  $\boldsymbol{\gamma}/\sigma$  lies in a sphere of radius  $\varepsilon = .25$  around  $\mathbf{0}$  and that we want to perform the test at significance level  $\alpha = .05$ . Then, using the appropriate intrinsic SAS function, the critical upper bound which the estimated standardized distance of  $\boldsymbol{\gamma}/\sigma$  from  $\mathbf{0}$  has to be compared to according to (9.21), is computed to be

$$\begin{aligned} & \frac{(r-1)(s-1)}{n} F_{(r-1)(s-1), (n-1)rs; \alpha}(n\varepsilon^2) = \frac{1}{57} \text{ finv(.05, 1, 224, 3.5625)} \\ & = \frac{1}{57} \cdot .125142 = .002195. \end{aligned}$$

Since this is clearly larger than the observed value of  $d^2(\hat{\boldsymbol{\gamma}}/S, \mathbf{0})$ , the data of Table 9.8 allow us to reject the null hypothesis that there is a nonnegligible

interaction between the antimuscarinic effects of both substances. Although a common sample size of almost 60 for each treatment combination is unusually large for a multiarm trial performed in a rather specialized field, it is still much too small for ensuring that the power of the UMPI test exceeds 80% against the alternative that all interaction effects vanish. In fact, for  $\psi^2 = 0$ , evaluating the right-hand side of (9.23) by means of the SAS function `probft` yields under our present specifications

$$\beta(0) = P[\mathcal{F}_{1,224}(0) \leq 0.125142] = \text{probft}(.125142, 1, 224) = .27614.$$

Keeping  $\alpha$  and  $\varepsilon$  fixed at  $\alpha = .05$  and  $\varepsilon = .25$ , respectively, at least  $n = 138$  observations per cell would be necessary for raising the power against the same type of alternative to 80% or more. With  $\varepsilon = .50$ ,  $n = 35$  would suffice for meeting the requirement  $\beta(0) \geq .80$ .

#### *Generalization of the UMPI test for negligibility of interactions to nonbalanced designs*

In an unbalanced two-way layout, the numbers  $n_{ij}$  of observations taken under the various treatment combinations are arbitrary natural numbers (discarding situations where some cells contain no data at all) allowed to be even pairwise different from each other. As holds generally true for analyses of classical linear models for multiway layouts, nonbalancedness induces complications both with regard to the computational techniques required for determining the test statistics and, still more important, interpretation of hypotheses.

From a computational perspective, the crucial change refers to the numerator of the (rescaled)  $F$ -statistic (9.20) which in the unbalanced case can no longer be written simply as the sum of squares of the best linear estimates of the  $\gamma_{ij}$  in the full model (9.18). In fact,  $\bar{X}_{ij} - (\bar{X}_{i\cdot} + \bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot})$  has to be replaced with  $\bar{X}_{ij} - (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j)$  where  $(\hat{\mu}, \hat{\alpha}_i, \hat{\beta}_j)$  denotes the least squares estimate of  $(\mu, \alpha_i, \beta_j)$  in the interaction-constrained model  $\mu_{ij} = \mu + \alpha_i + \beta_j$ . With unbalanced data, the  $\hat{\alpha}_i$ ,  $\hat{\beta}_j$  and  $\hat{\mu}$  admit no explicit representation as linear functions of the cell means  $\bar{X}_{ij}$  but must be determined by solving a system of linear equations whose coefficient matrix exhibits no closed-form inverse (for details see, e.g., Searle, 1987, § 4.9). Of course, instead of the ordinary we have to form the weighted sum of squares of the  $\hat{\gamma}_{ij} \equiv \bar{X}_{ij} - (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j)$  using weight-factors  $n_{ij}/\bar{n}$  where

$$\bar{n} \equiv N/rs \equiv \frac{1}{rs} \sum_{i=1}^r \sum_{j=1}^s n_{ij} . \quad (9.24)$$

Making the dependence of the numerator of the test statistic on these weights explicit by adding the subscript **n** [symbolizing the so-called incidence matrix

$(n_{ij})_{(r,s)}$ ] to  $d^2(\cdot, \cdot)$ , we have to replace (9.20) in the unbalanced case with

$$d_{\mathbf{n}}^2(\hat{\gamma}/S, \mathbf{0}) = \frac{\sum_{i=1}^r \sum_{j=1}^s (n_{ij}/\bar{n})(\bar{X}_{ij} - (\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j))^2}{\sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^{n_{ij}} (X_{ijk} - \bar{X}_{ij})^2/(N - rs)} . \quad (9.25)$$

From a practical point of view, it makes almost no difference whether computation of the test statistic has to be based on formula (9.25) or can be simplified to (9.20). In fact, any well-designed program package for use in statistical applications contains a procedure for computing  $F$ -statistics for all hypotheses of interest in analyzing ANOVA higher-way layouts with arbitrary incidence matrices. Denoting the  $F$ -ratio for testing the null hypothesis of no interactions by  $\mathcal{F}_{(r,s)}^{(\mathbf{n})}(\gamma|\mu, \boldsymbol{\alpha}, \boldsymbol{\beta})$ , we have just to read the value of the latter from the output list generated by running a suitable program (like `proc glm` in SAS) and multiply it by  $(r-1)(s-1)/\bar{n}$  to obtain the observed value of  $d_{\mathbf{n}}^2(\hat{\gamma}/S, \mathbf{0})$  as defined in (9.25). Clearly, this suggests using

$$\left\{ d_{\mathbf{n}}^2(\hat{\gamma}/S, \mathbf{0}) < ((r-1)(s-1)/\bar{n}) F_{(r-1)(s-1), N-rs; \alpha}(\bar{n}\varepsilon^2) \right\} . \quad (9.26)$$

as the *rejection region* of the desired test for negligibility of interactions in the case of an arbitrary incidence matrix  $\mathbf{n} \neq n\mathbf{1}_{(r,s)}$ .

However, the hypotheses for which (9.26) gives a UMPI critical region can no longer be expressed in terms of a spherical neighborhood of the origin in the space of the standardized true  $\gamma_{ij}$ . To see this, one has to examine the noncentrality parameter  $\lambda_{nc}^2$  of the distribution of  $\mathcal{F}_{(r,s)}^{(\mathbf{n})}(\gamma|\mu, \boldsymbol{\alpha}, \boldsymbol{\beta}) = (\bar{n}/(r-1)(s-1))d_{\mathbf{n}}^2(\hat{\gamma}/S, \mathbf{0})$  under an arbitrary parameter configuration with  $\gamma \neq \mathbf{0}$ . Discarding the factor  $\bar{n}$ ,  $\lambda_{nc}^2$  is obtained by substituting in the numerator of (9.25) every single observation  $X_{ijk}$  and hence every cell mean  $\bar{X}_{ij}$  by its expectation  $\mu_{ij}$ , and dividing the resulting expression by  $\sigma^2$ . Since  $\hat{\gamma}$  is a vector of linear functions, say  $\hat{\mathbf{g}}_{\mathbf{n}}(\bar{X}_{11}, \dots, \bar{X}_{rs})$ , of the cell means with coefficients determined by the  $n_{ij}$ , it follows that  $\mathcal{F}_{(r,s)}^{(\mathbf{n})}(\gamma|\mu, \boldsymbol{\alpha}, \boldsymbol{\beta})$  is noncentral  $F$  with  $(r-1)(s-1), N - rs$  degrees of freedom and  $\lambda_{nc}^2 = \bar{n}d_{\mathbf{n}}^2(\hat{\mathbf{g}}_{\mathbf{n}}(\mu_{11}, \dots, \mu_{rs})/\sigma, \mathbf{0})$  so that the hypotheses actually tested when using (9.26) as the critical region, are

$$H : d_{\mathbf{n}}^2(\hat{\mathbf{g}}_{\mathbf{n}}(\mu_{11}, \dots, \mu_{rs})/\sigma, \mathbf{0}) \geq \varepsilon^2 \quad \text{versus} \quad K : d_{\mathbf{n}}^2(\hat{\mathbf{g}}_{\mathbf{n}}(\mu_{11}, \dots, \mu_{rs})/\sigma, \mathbf{0}) < \varepsilon^2 \quad (9.27)$$

On the one hand, for any configuration  $(\mu_{11}, \dots, \mu_{rs}) \in \mathbb{R}^{rs}$  of true cell means satisfying the no-interaction model exactly,  $d_{\mathbf{n}}^2(\hat{\mathbf{g}}_{\mathbf{n}}(\mu_{11}, \dots, \mu_{rs})/\sigma, \mathbf{0})$  vanishes and can thus be viewed as a generalized measure of distance between true and constraint model. On the other, the parameter sets corresponding to the hypotheses of (9.27) depend in a complicated way on the

entries in the incidence matrix  $\mathbf{n}$  and will be comparable in form to those considered in the balanced-data case only for small to moderate values of  $(\max_{i,j} n_{ij} / \min_{i,j} n_{ij}) - 1$ . In presence of gross differences between the cell frequencies, a more satisfactory solution to the problem of testing for negligibility of interactions might be obtained by keeping the hypotheses unchanged and constructing a large-sample test for (9.19) based on the asymptotic distribution of  $\sqrt{N}(d^2(\hat{\gamma}/S, \mathbf{0}) - d^2(\gamma/\sigma, \mathbf{0}))$ . Up to now, this approach has not been pursued in the literature so that nothing can be said about how the resulting procedure compares with that based on (9.26).

### 9.3.2 Establishing negligibility of carryover effects in the analysis of two-period crossover trials

In this subsection, we suppose that the data set under analysis stems from a standard two-period crossover trial involving two different treatments, say  $A$  and  $B$ . Trials of this type are very popular in various branches of clinical research and give in particular the standard case of a comparative bioavailability study to which the final chapter of this book is devoted in full length. The decisive feature of the two-period crossover design is that each experimental unit receives both treatments under comparison in temporal succession. Since there are of course two possible sequences in which the treatments can be administered the total sample of  $N = m + n$  subjects is randomly split into sequence groups  $A/B$  and  $B/A$ , say. In the first of these groups, each subject receives treatment  $A$  at the beginning of the first period of the trial, and is switched to treatment  $B$  in a second period. Usually, between the end of Period 1 and the beginning of Period 2, a sufficiently long washing-out period is scheduled. The only difference in the second sequence group  $B/A$  as compared to the first one is that the order of application of the treatments is reversed. It is furthermore assumed that the same univariate quantitative outcome measure is used in both periods to describe each subject's biological condition after administration of the respective treatment. A convenient notation is obtained by defining  $X_{ki}$  and  $Y_{kj}$  to be the outcome observed in period  $k = 1, 2$  in the  $i$ th and the  $j$ th subject of sequence group  $A/B$  and  $B/A$ , respectively. Thus, a complete description of the layout is as follows:

Sequence Group	Period 1	Period 2
$A/B$	$X_{11}, \dots, X_{1m}$	$X_{21}, \dots, X_{2m}$
$B/A$	$Y_{11}, \dots, Y_{1n}$	$Y_{21}, \dots, Y_{2n}$

At first sight, such a design looks simply like a special case of a two-way ANOVA layout with both factors varying over just two levels. However, there is a marked difference deserving keen attention: In the crossover

case, the observations from different cells are not mutually independent but form pairs within each row. In other words, the data set obtained from a two-period crossover trial consists of two unrelated samples of bivariate observations rather than four samples made up of one-dimensional variables. By far the most popular approach to analyzing the data of a standard crossover trial is based on a simple parametric model originally introduced by Grizzle (1965). Of course, it allows in particular for correlations within all pairs  $(X_{1i}, X_{2i}), (Y_{1j}, Y_{2j})$  and can be written

$$X_{ki} = \mu_k + S_i^{(1)} + \varepsilon_{ki}^{(1)}, \quad i = 1, \dots, m, \quad k = 1, 2, \quad (9.28a)$$

$$Y_{kj} = \nu_k + S_j^{(2)} + \varepsilon_{kj}^{(2)}, \quad j = 1, \dots, n, \quad k = 1, 2, \quad (9.28b)$$

where all  $S_i^{(1)}, S_j^{(2)}, \varepsilon_{ki}^{(1)}, \varepsilon_{kj}^{(2)}$  are mutually independent with,

$$S_i^{(1)}, S_j^{(2)} \sim \mathcal{N}(0, \sigma_S^2) \quad (9.29)$$

and

$$\varepsilon_{1i}^{(1)}, \varepsilon_{1j}^{(2)} \sim \mathcal{N}(0, \tau_1^2), \quad \varepsilon_{2i}^{(1)}, \varepsilon_{2j}^{(2)} \sim \mathcal{N}(0, \tau_2^2). \quad (9.30)$$

Usually,  $S_i^{(1)}$  and  $S_j^{(2)}$  is called the (random) effect of the  $i$ th and  $j$ th subject in sequence group  $A/B$  and  $B/A$ , respectively,  $\sigma_S^2$  the between-subject variance, and  $\tau_k^2$  the within-subject variance for the  $k$ th period.

For the expected values  $\mu_k, \nu_k$  of the  $X_{ki}, Y_{kj}$ , Grizzle's model assumes that these allow to be additively decomposed into a overall mean  $\omega$ , a "direct treatment effect"  $\phi_A$  or  $\phi_B$ , a period effect  $\pi_k$ , and for  $k = 2$ , a carryover effect denoted here  $\lambda^{(1)} [\leftrightarrow A/B]$  and  $\lambda^{(2)} [\leftrightarrow B/A]$ , respectively. The corresponding model equations are

$$\mu_1 = \omega + \phi_A + \pi_1, \quad \mu_2 = \omega + \phi_B + \pi_2 + \lambda^{(1)}, \quad (9.31a)$$

$$\nu_1 = \omega + \phi_B + \pi_1, \quad \nu_2 = \omega + \phi_A + \pi_2 + \lambda^{(2)}. \quad (9.31b)$$

Clearly, the carryover effects reflect deviations from strict additivity of treatment and period effects and are thus another instance of interaction parameters. If they have some common value, say  $\lambda^{(0)}$ , they can be totally dropped from the above model equations since then  $\omega + \lambda^{(1)}$  and  $\omega + \lambda^{(2)}$  are simply relabeled versions of the same overall mean. Thus, with regard to the carryover effects in a two-period crossover setting, negligibility means equality except for irrelevant differences.

In authoritative expositions of inferential methods for the analysis of crossover trials (see, e.g., Jones and Kenward, 2003) in addition to Grizzle's pioneering paper of 1965), it is much more clearly understood than in the literature on classical ANOVA methods, that the primary aim of inference on interactions consists in ruling out the possibility that such effects must be taken into account. The standard argument which is usually presented in favor of this view is that otherwise none of the differential main effects  $\phi_A - \phi_B$

and  $\pi_1 - \pi_2$  admits unbiased estimation from the full data set. Nevertheless, also in this field, very little has changed over the years with regard to the usual practice of assessing the difference between both carryover effects: A preliminary test of the null hypothesis  $\lambda^{(1)} = \lambda^{(2)}$  is carried out and a non-significant result considered sufficient to warrant the conclusion that the  $\lambda^{(k)}$  can be discarded when assessing the main effects. As a tried way around the flaw that this test is obviously directed to the wrong side, Grizzle suggested to raise the nominal significance level to  $\alpha = .10$ , and this advice was adopted by Jones and Kenward in the first edition of their book.

Of course, it is our aim here to show how this makeshift solution of the problem of establishing approximate additivity of treatment and period effects in a crossover layout can be replaced with a rigorous inferential procedure. It will turn out that the problem can readily be reduced in such a way that its solution requires no new ideas but simply an application of some of the methods developed in Chapter 6. In order to see this, let us recall the rationale behind the standard parametric test for the traditional two-sided problem  $\lambda^{(1)} = \lambda^{(2)}$  vs.  $\lambda^{(1)} \neq \lambda^{(2)}$ . The initial step, from which the rest follows almost automatically, consists of forming for each subject the total response by summing up his responses to both treatments. Defining

$$X_i^+ = X_{1i} + X_{2i} \quad \forall i = 1, \dots, m, \quad Y_j^+ = Y_{1j} + Y_{2j} \quad \forall j = 1, \dots, n, \quad (9.32)$$

it is a straightforward exercise to verify by means of (9.28)–(9.31) that the  $X_i^+$  and  $Y_j^+$  make up two independent samples from Gaussian distributions with common variance

$$\sigma_+^2 = 4\sigma_S^2 + \tau_1^2 + \tau_2^2 \quad (9.33)$$

and expected values  $\mu_+ = E(X_i^+)$ ,  $\nu_+ = E(Y_j^+)$  such that

$$\mu_+ - \nu_+ = \lambda^{(1)} - \lambda^{(2)}. \quad (9.34)$$

In view of (9.33) and (9.34), it is clearly appropriate to formulate the problem of establishing negligibility of carryover effects in the model due to Grizzle (1965) as the problem of testing

$$H_+ : |\mu_+ - \nu_+|/\sigma_+ \geq \varepsilon \quad \text{versus} \quad K_+ : |\mu_+ - \nu_+|/\sigma_+ < \varepsilon \quad (9.35)$$

by means of the within-subject totals (9.32). Since the latter satisfy

$$X_i^+ \sim \mathcal{N}(\mu_+, \sigma_+^2) \quad \forall i, \quad Y_j^+ \sim \mathcal{N}(\nu_+, \sigma_+^2) \quad \forall j, \quad (9.36)$$

(9.35) is simply a special case of (6.2) [→ p. 119]. Thus, there is a UMPI level- $\alpha$  test which rejects  $H_+$  if and only if the absolute value of the ordinary two-sample statistic  $T_+$ , say, calculated with the  $X_i^+$  and  $Y_j^+$  as the raw data, turns out smaller than  $\sqrt{F_{1,N-2;\alpha}(mn\varepsilon^2/N)}$  [for the definition of  $F_{1,N-2;\alpha}$ , recall (9.22)].

*Example 9.5*

It is instructive to compare the exact optimal test for (9.35) with the “inverted” two-sided  $t$ -test through applying both procedures to the same data set. Table 9.9 shows the results of a classical crossover trial of the efficacy of a certain oxygen gel [ $\leftrightarrow (B)$ ] in the improvement of dental plaque and gingival-inflammation, as compared with placebo and used at several places in the biostatistical literature for illustrating various techniques useful for the analysis of crossover trials (Brown, 1980; Jones and Kenward, 1989). The values displayed in the table represent improvements in a summary score of oral hygiene during each trial period.

Table 9.9 *Improvements in a summary score of oral hygiene observed in a crossover trial comparing a new oxygen gel (B) with placebo (A). (Data originally reported by Zinner et al. (1970); reproduced from Brown (1980), with kind permission by Wiley-Blackwell Publishing, Oxford.)*

*Group A/B*

$i$	1	2	3	4	5	6	7	8	9	10	11	12
$X_{1i}$	0.83	1.00	0.67	0.5	0.50	0.83	1.0	0.67	0.67	0.33	0.00	1.17
$X_{2i}$	1.83	2.17	1.67	1.5	2.33	1.83	0.5	0.33	0.50	0.67	0.83	1.33
$X_i^+$	2.66	3.17	2.34	2.0	2.83	2.66	1.5	1.00	1.17	1.00	0.83	2.50
	13	14	15	16	17	18	19	20	21	22	23	24
0.00	0.50	0.33	0.33	0.50	1.00	0.00	0.5	-0.50	0.17	1.00	1.00	
0.67	1.83	1.50	1.50	1.17	1.67	1.33	1.5	2.83	2.33	1.33	1.67	
0.67	2.33	1.83	1.83	1.67	2.67	1.33	2.0	2.33	2.50	2.33	2.67	
	25	26	27	28	29	30	31	32	33	34		
1.33	0.33	2	4.00	0.83	0.50	0.5	0.50	2.17	0.67			
0.67	0.83	1	0.17	1.67	1.33	1.5	1.67	1.33	1.17			
2.00	1.16	3	4.17	2.50	1.83	2.0	2.17	3.50	1.84			

*Group B/A*

$j$	1	2	3	4	5	6	7	8	9	10	11	12
$Y_{1j}$	1.67	2.5	1.00	1.67	1.83	0.50	1.33	1.33	0.50	2.17	1.67	1.5
$Y_{2j}$	0.33	0.5	-0.17	0.50	0.50	0.33	0.67	0.00	0.17	0.83	0.33	0.0
$Y_j^+$	2.00	3.0	0.83	2.17	2.33	0.83	2.00	1.33	0.67	3.00	2.00	1.5

Table 9.9 (*continued*)

13	14	15	16	17	18	19	20	21	22	23	24
1.33	1.5	1.33	0.67	1.67	2.50	1.83	0.83	2.33	1.17	1.33	1.33
0.50	0.5	0.00	-0.17	0.50	0.67	0.00	0.67	0.17	0.50	0.00	0.83
1.83	2.0	1.33	0.50	2.17	3.17	1.83	1.50	2.50	1.67	1.33	2.16
<hr/>											
25	26	27	28	29	30						
0.33	2.17	1.00	0.33	1.17	0.5						
1.33	1.17	0.33	1.00	0.17	0.5						
1.66	3.34	1.33	1.33	1.34	1.0						

As argued by Jones and Kenward (1989, p. 39), the 28th subject of Group  $A/B$  should be removed as an outlier. With the remaining  $N = 33 + 30$  within-subject totals, the  $t$ -statistic is computed to be

$$\begin{aligned} T_+ &= \sqrt{\frac{33 \cdot 30}{63}} (2.0552 - 1.7883) / [(32 \cdot .699331^2 + 29 \cdot .730654^2) / 61]^{1/2} \\ &= 1.4810. \end{aligned}$$

In the conventional two-sided test, the associated p-value is .1438 so that the null hypothesis  $\lambda^{(1)} = \lambda^{(2)}$  has to be accepted also at the 10%-level. According to Grizzle's rule there is thus no reason to suspect that the standard test for assessing the direct treatment and period effects might lead to invalid conclusions. Performing the  $t$ -test for equivalence of both distributions of within-subject totals instead, it seems appropriate to rely on a rather strict specification of the tolerance  $\varepsilon$  determining the hypotheses of (9.35). Following the recommendations put forward in § 1.5 this amounts to setting  $\varepsilon = .36$ , and with that value of  $\varepsilon$  the critical constant of the UMPI test for (9.35) at level  $\alpha = .05$  based on samples of size  $m = 33$  and  $n = 30$  is found to be (by means of the SAS intrinsic function for quantiles of arbitrary  $F$ -distributions)

$$\begin{aligned} [F_{1,61; .05}(33 \cdot 30 \cdot .36^2 / 63)]^{1/2} &= \text{finv}(.05, 1, 61, 2.0366)**.05 \\ &= 0.173292. \end{aligned}$$

Since the observed value of  $|T_+|$  is far beyond this bound, our formal test for absence of relevant carryover effects does not lead to a positive decision in the case of the dental hygiene study of Zinner et al. Having a look at the power of the test, this result is far from being surprising: Even under the null alternative  $\lambda^{(1)} = \lambda^{(2)}$ , the rejection probability is as low as 13.7%, and in a balanced trial, 133 subjects in each sequence group would be required for

raising the chance of detecting perfect additivity to 80%. From a different perspective, this result corroborates the conclusion drawn by Brown (1980) from his analysis of the efficiency of the carryover relative to a simple parallel-group design: The frequently invoked superiority of the former holds only in those cases where additivity of the main effects can be taken for granted *a priori* and need not be established by means of a valid preliminary test providing reasonable power to detect that the restriction  $\lambda^{(1)} = \lambda^{(2)}$  is actually satisfied. Moreover, even if the preliminary test for carryover effects is performed in the logically adequate version treating the assumption of nonexistence of relevant carryover effects as the alternative hypothesis and the sample sizes are sufficiently large for ensuring reasonable power against the specific alternative  $\lambda^{(1)} = \lambda^{(2)}$ , serious doubts have to be raised regarding its use as part of a two-stage procedure. In fact, the results obtained by Freeman (1989) on the traditional two-stage procedure using the intraindividual differences between both periods for assessing the treatment effects only conditionally on a sufficiently large  $p$ -value of the conventional two-sided test for differences between the carryover effects, are likely to apply also to the equivalence version of the latter. In other words, a two-stage procedure starting with carrying out the UMPI test for (9.35) as a pre-test for making sufficiently sure that the assumption of negligibility of carryover effects holds at least approximately, tends to be anticonservative. Consequently, preliminary testing for negligibility of the carryover effects can only be recommended when a separate pilot sample has been made available not to be used in the main testing step, whether or not the test for differences between treatment effects will eventually be chosen to involve the observations taken in both periods.

### *Alternative approaches*

- (i) *Assessing raw rather than standardized effects.* Sometimes, it might be preferable to measure the extent of the deviation from strict additivity of the main effects through the raw difference  $\lambda^{(1)} - \lambda^{(2)}$  between the carryover effects. In such cases, it is reasonable to replace the testing problem (9.35) with

$$H'_+ : |\mu_+ - \nu_+| \geq \varepsilon' \quad \text{versus} \quad K'_+ : |\mu_+ - \nu_+| < \varepsilon' \quad (9.37)$$

and to apply the interval inclusion approach of § 3.1 to the latter. A guidance for specifying the equivalence range  $(-\varepsilon', \varepsilon')$  in this alternative formulation of the problem of testing for negligibility of the carryover effects can be found in Brown (1980): Considering the power of the traditional two-sided test, this author recommends choosing  $\varepsilon' = \delta'/2$  with  $\delta' > 0$  denoting the true difference one tries to detect in the standard test of equality of direct treatment effects.

Suppose the authors of the study considered in the above Example 9.5 eventually aimed at detecting an effect of  $\delta' = 1.00$  of the oxygen gel

as compared with placebo. Then, in the interval inclusion test at level  $\alpha = .05$  for (9.37), rejection of  $H'_+$  would require that  $(-.50, .50) \supseteq (\bar{X}^+ - \bar{Y}^+ - t_{61;.95}\tilde{S}^+, \bar{X}^+ - \bar{Y}^+ + t_{61;.95}\tilde{S}^+)$ , with  $\tilde{S}^+$  denoting the standard error of  $\bar{X}^+ - \bar{Y}^+$  as estimated from the pooled sample. Since  $\tilde{S}^+$  can obviously be written  $\tilde{S}^+ = |\bar{X}^+ - \bar{Y}^+|/|T^+|$  and we had  $\bar{X}^+ = 2.0552$ ,  $\bar{Y}^+ = 1.7883$ ,  $T^+ = 1.4810$ , the 95%-confidence bounds to  $\mu_+ - \nu_+$  are computed to be  $0.2668 \pm 1.67022 \cdot .2668/1.4810 = 0.2668 \pm .3009$ . Thus, with  $\varepsilon' = .50$  and at level  $\alpha = .05$ , the null hypothesis of a nonnegligible absolute difference between the carryover effects must likewise be accepted.

- (ii) *Testing for negligibility of the carryover effects without relying on parametric distributional assumptions.* In cases of doubt about the adequacy of the distributional assumptions (9.29), (9.30), it is generally advisable (cf. Jones and Kenward, 1989, § 2.8) to replace  $t$ -tests with their nonparametric counterparts. Of course, there is no reason for exempting the test for negligibility of the carryover effects from this recommendation. As we know from Chapter 6, there is a distribution-free analogue of the  $t$ -test for equivalence of two homoskedastic Gaussian distributions with respect to the standardized shift in location. With the computational tools presented in § 6.2, the Mann-Whitney test for equivalence of both distributions of within-subject totals is almost as easily carried out as the noncentral  $t$ -test. Specifying the equivalence range  $(1/2 - \varepsilon'_1, 1/2 + \varepsilon''_2) = (.40, .60)$  for the target functional  $P[X_i^+ > Y_j^+]$  (which, according to the results of § 1.7, is approximately the same as requiring  $|\mu_+ - \nu_+|/\sigma_+ < .36$  in the Gaussian case), we obtain by means of the SAS macro `mawi` (or the R version of this program) with the data of Example 9.5:

$$W_+ = .59495, \hat{\sigma}[W_+] = .070076, C_{MW}(.05; .10, .10) = .17259.$$

In view of  $|W_+ - 1/2|/\hat{\sigma}[W_+] = 1.3550 > C_{MW}(.05; .10, .10)$ , the decision is identical to that we were lead to take using the UMPI test at the same level for the parametric version of the problem.

## 9.4 Testing for approximate compatibility of a genotype distribution with the Hardy-Weinberg condition

### 9.4.1 Genetical background, measures of HW disequilibrium

The so-called Hardy-Weinberg law plays a central role both in classical population genetics and modern genetic epidemiology of complex diseases. Since

genotyping is nowadays almost exclusively done at the highest level of resolution yielding data on single DNA bases, it suffices to restrict discussion to the Hardy-Weinberg model for biallelic markers called single nucleotide polymorphisms (SNPs) in contemporary genetic terminology.

Except for loci on the sex chromosomes, genotyping in terms of SNPs generates trinomial frequency distributions as basic data. More precisely speaking, whenever a total of  $n$  individuals have been genotyped at some single nucleotide SNP involving two alleles denoted by A and B, the empirical genotype distribution found in the sample under consideration is given by the entries in the upper row of the following one-way contingency table:

Genotype	AA	AB	BB	$\sum$
Absolute Count	$X_1$	$X_2$	$X_3$	$n$
Population frequency	$\pi_1$	$\pi_2$	$\pi_3$	1.000

In traditional genetic notation, the corresponding allele frequencies are denoted by  $p$  [ $\leftrightarrow$  A] and  $q$  [ $\leftrightarrow$  B] where

$$p \equiv \pi_1 + \pi_2/2, \quad q \equiv \pi_2/2 + \pi_3. \quad (9.38)$$

The law discovered independently in 1908 by the British mathematician G.H. Hardy and the German physician W. Weinberg states that the distribution of a genotype involving two different alleles on each chromosome is reproduced in a 1:1 manner within the next generation if and only if the genotype frequencies  $(\pi_1, \pi_2, \pi_3)$  can perfectly be reconstructed from the allele frequencies according to the following equations:

$$\pi_1 = p^2 = (\pi_1 + \pi_2/2)^2; \quad (9.39a)$$

$$\pi_2 = 2pq = 2(\pi_1 + \pi_2/2)(\pi_2/2 + \pi_3); \quad (9.39b)$$

$$\pi_3 = q^2 = (\pi_2/2 + \pi_3)^2. \quad (9.39c)$$

It is important to note that, if the genotype frequencies for the next generation are denoted by  $(\pi'_1, \pi'_2, \pi'_3)$ , then according to the basic rule of Mendelian population genetics we *always* have  $\pi'_1 = (\pi_1 + \pi_2/2)^2$ ,  $\pi'_2 = 2(\pi_1 + \pi_2/2)(\pi_2/2 + \pi_3)$ , and  $\pi'_3 = (\pi_2/2 + \pi_3)^2$ . Thus, Hardy-Weinberg's law implies that, without "disturbing" influences from "outside," a genotype distribution reaches equilibrium already in the next to current generation.

Keeping in mind the boundary condition  $\pi_1 + \pi_2 + \pi_3 = 1$ , it is easily verified that each of the above equations (9.39a–c) implies the other two. Thus, the genotype distribution under assessment is in strict Hardy-Weinberg equilibrium (HWE) if and only if there holds

$$\pi_2 = 2\left(\sqrt{\pi_1} - \pi_1\right) \quad \forall \pi_1 \in (0, 1). \quad (9.40)$$

The curve in the parameter space of  $(\pi_1, \pi_2)$  given by this equation is usually graphically represented in a so-called DeFinetti diagram constructed through applying the following simple transformation of coordinates:

$$(\pi_1, \pi_2) \mapsto (\pi_1 + \pi_2/2, \pi_2). \quad (9.41)$$

One of the reasons why this transformation is very popular among statistical geneticists and genetic epidemiologists (see, e.g., Ziegler and König, 2006, § 2.4) is that it maps that curve into the symmetric parabola with equation  $y = 2x(1 - x)$ .

The assessment of departure from HWE is still the most important and frequently used tool of quality control in association studies on genetic risk factors for complex diseases involving unrelated individuals. A standard strategy is to filter markers that do not conform with HWE prior to the conduct of genetic association tests (for details see Ziegler and König, 2006, Ch. 4). Unfortunately, the traditional statistical procedures to assess HWE are not tailored for establishing compatibility with the model which is the actual aim in the majority of applications of that type. Specifically, this statement holds true for the standard Pearson  $\chi^2$ -test which to call a goodness-of-fit test is as misleading in the present context as in the settings considered in the first two sections of this chapter. Actually, the corresponding null hypothesis is that the distribution underlying the data is in agreement with HWE, and a significant test result indicates incompatibility of the observed data with the model. Accordingly, this approach allows only the identification of markers with the most severe deviations from HWE so that two basic issues remain. First, one cannot conclude that the remaining markers, for which the test does not reject its null hypothesis, do conform to the HW proportions of (9.39), since it may simply lack the power to detect even a gross violation of the equilibrium condition. The second and perhaps more troubling problem is that, given the large sample sizes that are nowadays typically used for studies of complex diseases, the null hypothesis of perfect HWE in the underlying population may be rejected because of high power for detecting a small, irrelevant deviation from HWE.

A promising way around these difficulties consists of constructing a test of the null hypothesis that the true population frequencies  $(\pi_1, \pi_2)$  correspond to a point in the parameter space which lies outside a suitable indifference zone around the curve determined by the family of all genotype distributions being in exact HWE, versus the alternative of equivalence with HWE. The first step to be taken on the way towards accomplishing this goal is to identify a function of the true genotype frequencies  $(\pi_1, \pi_2, \pi_3)$  which gives a reasonable measure of HW *disequilibrium* (HWD). The specific measure of HWD we will use in the sequel combines mathematical convenience with ease of genetic interpretability. We start with elaborating briefly on the latter aspect.

Clearly, there are many equivalent ways of re-expressing the complete set (9.39a–c) of conditions being jointly necessary and sufficient for HWE, through

a single restriction on the components of the parameter of the underlying trinomial distribution. For the present purpose, the most interesting of them is given by the statement

$$\pi_2 = 2\sqrt{\pi_1 \pi_3} \quad \forall (\pi_1, \pi_2, \pi_3) \in \left\{ (u_1, u_2, u_3) \in (0, 1)^3 \mid \sum_{j=1}^3 u_j = 1 \right\}. \quad (9.42)$$

Equivalence between the statements (9.39a–c) and (9.42) (for a short formal proof see Wellek et al., 2009, Section A.1) implies that the degree of Hardy-Weinberg disequilibrium (HWD) is reflected by the extent to which the actual proportion  $\pi_2$  of heterozygotes differs from the proportion  $2\sqrt{\pi_1 \pi_3}$  of heterozygotes expected in a population which conforms exactly to HWE. This justifies considering the ratio

$$\omega \equiv (\pi_2/2)/\sqrt{\pi_1 \pi_3} \quad (9.43)$$

as a measure of the relative excess heterozygosity (REH) and declaring the actual genotype distribution under analysis for being sufficiently close to HWE if and only if it can be taken for granted that there holds  $1 - \varepsilon_1 < \omega < 1 + \varepsilon_2$  for suitably chosen equivalence margins  $0 < \varepsilon_1 < 1$ ,  $\varepsilon_2 > 0$ .

In § 9.4.3, we will argue that a choice which can well be recommended for routine applications in genetic association studies is to set the equivalence limits to  $\omega$  equal to  $1 - \varepsilon_1 = 1/(1 + \varepsilon_o)$ ,  $1 + \varepsilon_2 = 1 + \varepsilon_o$ , with  $\varepsilon_o = .4$ . Figure 9.1 visualizes the corresponding equivalence region in the parameter space of  $(\pi_1, \pi_2)$  in a DeFinetti diagram. The basic features of this band are

- symmetry of both boundary curves about the line  $\{(1/2, \pi_2) \mid 0 \leq \pi_2 \leq 1\}$
- increasingness of the vertical width in the allele frequency  $p$  on the left and in  $(1 - p)$  on the right of the axis of symmetry.

The subsequent subsection is devoted to the construction of an exact optimal test for establishing the (alternative) hypothesis that the genotype distribution generated through determining a given SNP in the population under study corresponds to a point in the  $(p, \pi_2)$ -plane lying inside the equivalence band shown in Figure 9.1. In § 9.4.3, the same testing problem will be addressed by means of the principle of confidence interval inclusion. In its asymptotic version, the interval inclusion test will turn out to be computationally very simple, making it better suited for applications in genomewide association studies involving many hundred thousands of SNPs as compared with the exact UMPU test.

#### 9.4.2 Exact conditional tests for absence of substantial disequilibrium

First of all, we introduce a bit of additional notation setting

$$\theta \equiv \frac{\pi_2^2}{\pi_1(1 - \pi_1 - \pi_2)}, \quad (9.44)$$

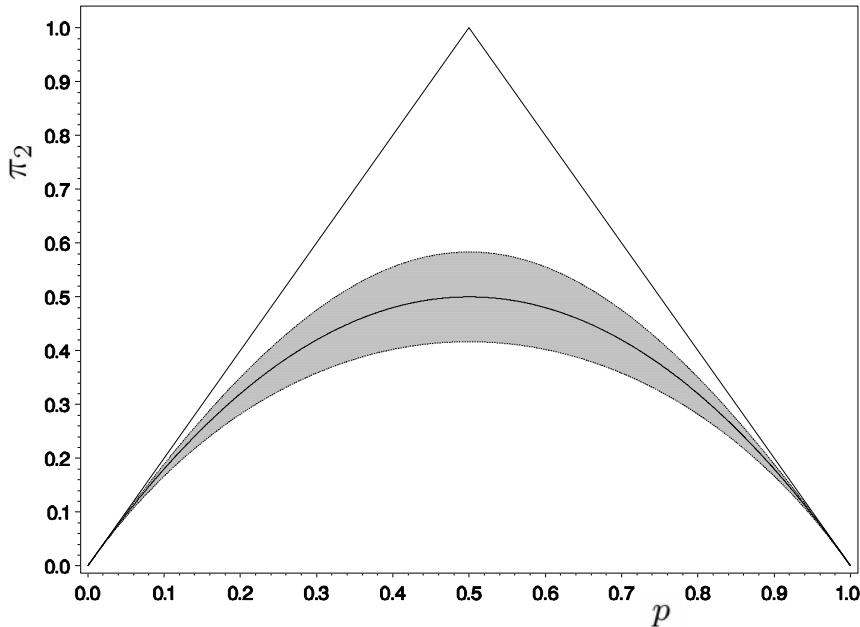


Figure 9.1 Equivalence band obtained by applying the condition  $1 - \varepsilon_1 < \omega < 1 + \varepsilon_2$  with  $\varepsilon_1 = 2/7$ ,  $\varepsilon_2 = 2/5$ . [Middle solid line: HWE-curve (9.40), transformed by means of (9.41).]

$$\vartheta \equiv \pi_1 / (1 - \pi_1 - \pi_2), \quad (9.45)$$

and

$$c_n(\theta, \vartheta) \equiv (1 + \sqrt{\theta\vartheta} + \vartheta)^{-n}. \quad (9.46)$$

The starting-point for the construction of an exact UMPU test for approximate compatibility of a SNP-based genotype distribution with HWE is given by the fact that the joint probability mass function of the absolute genotype counts  $X_1$  [ $\leftrightarrow$  AA] and  $X_2$  [ $\leftrightarrow$  AB] in a random sample of size  $n$  admits the representation

$$p_{\theta, \vartheta}(x_1, x_2) = \frac{n!}{x_1! x_2! (n - x_1 - x_2)!} \theta^{x_2/2} \vartheta^{(2x_1+x_2)/2} c_n(\theta, \vartheta). \quad (9.47)$$

Obviously, (9.47) implies that the possible distributions of  $(X_1, X_2)$  form a two-parameter exponential family in  $(\log \sqrt{\theta}, \log \sqrt{\vartheta})$  and  $(T, S)$ , with the sufficient statistics being given by  $T = X_2$ ,  $S = 2X_1 + X_2$ . Comparing (9.44)

with (9.43), it is likewise obvious that the two parametric functions  $\omega$  and  $\theta$  are related through the simple equation

$$\theta = 4\omega^2, \quad (9.48)$$

implying that each of them is a continuous and strictly increasing function of the other. Accordingly, the testing problem

$$H : \omega \leq 1 - \varepsilon_1 \text{ or } \omega \geq 1 + \varepsilon_2 \quad \text{versus} \quad K : 1 - \varepsilon_1 < \omega < 1 + \varepsilon_2 \quad (9.49)$$

is the same as

$$H : \theta \leq 4(1 - \delta_1) \text{ or } \theta \geq 4(1 + \delta_2) \quad \text{versus} \quad K : 4(1 - \delta_1) < \theta < 4(1 + \delta_2), \quad (9.50)$$

provided, the  $\delta$ 's are specified to be the following transforms of the  $\varepsilon$ 's:

$$\delta_1 = 1 - (1 - \varepsilon_1)^2, \quad \delta_2 = (1 + \varepsilon_2)^2 - 1. \quad (9.51)$$

In view of (9.47) and the continuity and strict increasingness of the transformation  $\theta \mapsto \log \sqrt{\theta}$ , it is clear that the testing problem (9.50) is another special case to which the general result stated in the Appendix as Theorem A.2.2 applies. Hence, we can conclude that a UMPU level- $\alpha$  test exists and is defined by the following decision rule to be followed for any fixed value  $s$  of the total number  $S$  of A alleles:

$$\begin{cases} \text{reject } H \text{ if } C_{\alpha;s}^1 + 2 \leq X_2 \leq C_{\alpha;s}^2 - 2; \\ \text{reject } H \text{ with probability } \gamma_{\alpha;s}^\nu \text{ if } X_2 = C_{\alpha;s}^\nu, \quad \nu = 1, 2; \\ \text{accept } H \text{ if } X_2 \leq C_{\alpha;s}^1 - 2 \text{ or } X_2 \geq C_{\alpha;s}^2 + 2. \end{cases} \quad (9.52)$$

The critical constants  $C_{\alpha;s}^\nu$ ,  $\gamma_{\alpha;s}^\nu$  ( $\nu = 1, 2$ ) have to be determined through solving the equations

$$\sum_{x_2=C_{\alpha;s}^1+2}^{C_{\alpha;s}^2-2} p_{\theta_1;s}(x_2) + \sum_{\nu=1}^2 \gamma_{\alpha;s}^\nu p_{\theta_1;s}(C_{\alpha;s}^\nu) = \alpha = \sum_{x_2=C_{\alpha;s}^1+2}^{C_{\alpha;s}^2-2} p_{\theta_2;s}(x_2) + \sum_{\nu=1}^2 \gamma_{\alpha;s}^\nu p_{\theta_2;s}(C_{\alpha;s}^\nu), \quad (9.53)$$

where  $p_{\theta_1;s}(\cdot)$  and  $p_{\theta_2;s}(\cdot)$  stands for the probability mass function of the conditional distribution of  $X_2$  given  $S = s$  under  $\theta = \theta_1 \equiv 4(1 - \delta_1)$  and  $\theta = \theta_2 \equiv 4(1 + \delta_2)$ , respectively.

The conditional distribution of  $X_2$  which has been investigated by Stevens already in the 1930s, has a comparatively simple mathematical form. In order to provide an explicit formula for its mass function under any value of  $\theta$ , we have first to note that, given any fixed value  $s \in \{0, 1, \dots, 2n\}$  of  $S$ , the set

of possible values of  $X_2$ , say  $\mathcal{W}_s^n(X_2)$ , depends both on whether  $s$  is odd or even and smaller or larger than  $n$ . Actually, there holds:

$$\mathcal{W}_s^n(X_2) = \begin{cases} 0, 2, 4, \dots, s & \text{for even } s \leq n \\ 1, 3, 5, \dots, s & " \text{ odd } s \leq n \\ 0, 2, 4, \dots, 2n-s & " \text{ even } s \geq n \\ 1, 3, 5, \dots, 2n-s & " \text{ odd } s \geq n \end{cases}. \quad (9.54)$$

Furthermore, let us define:

$$C_s^n(j) = \frac{n!}{((s-j)/2)!j!(n-j/2-s/2)!} \quad \text{for all } j \in \mathcal{W}_s^n(X_2). \quad (9.55)$$

Then, the conditional probability of observing  $x_2$  heterozygotes in a sample where exactly  $s$  A-alleles are counted, is given by

$$P_\theta[X_2 = x_2 | S = s] = C_s^n(x_2) \theta^{x_2/2} / \sum_{x'_2 \in \mathcal{W}_s^n(X_2)} C_s^n(x'_2) \theta^{x'_2/2}. \quad (9.56)$$

An algorithm for computing the critical constants of the UMPU level- $\alpha$  test (9.52) for the equivalence problem (9.50) from Equations (9.53) is obtained through adapting the computational scheme presented in §3.3 to the special case given by (9.56). An implementation allowing for sample sizes not exceeding 600 has been made available as a SAS macro in a Web appendix to Wellek (2004). An improved version with no limitation concerning the sample size is provided as accompanying material to Goddard et al. (2009) and can be found in the **WKTSEQ2 Source Code Package** under the program name **gofhwex** both as a SAS macro and a R function.

It is of some importance to note that the test given by the decision rule (9.52), is not only UMPU but also remains invariant under exchanging the counts  $X_1$  and  $X_3$  of both homozygous genotypes. The reason why this is a desirable property of any test for assessing compatibility of some biallelic genotype distribution for compatibility with HWE comes from the fact that labeling of the two alleles discriminated in the course of genotyping is completely arbitrary so that the same is true for assigning the two possible homozygous variants to the outer cells of the underlying contingency table. Accordingly, in a reasonable formulation of the testing problem, both hypotheses should be invariant against exchanging  $\pi_1$  and  $\pi_3$ , and it follows immediately from (9.43) that (9.49) satisfies this condition for any choice of the equivalence margins  $\varepsilon_\nu$ . It is not difficult to see that the conditional test obtained above has the analogous property in terms of  $X_1$  and  $X_3$ : Conditioning on  $\{2X_1 + X_2 = s\}$  leads to the same distribution of  $X_2$  as fixing the value of the sufficient statistic  $S = 2X_1 + X_2$  at  $2n - s$ , and it is obvious that  $2X_1 + X_2 = 2n - s$  occurs if and only if  $s$  is the value taken on by  $2X_1 + X_2$ .

Another nice property of the UMPU test (9.52) for approximate compatibility of a genotype distribution with the Hardy-Weinberg law is the possibility of carrying out exact power calculations. For the purpose of computing

the rejection probability  $\beta_{\text{HWE}}(\boldsymbol{\pi})$ , say, under an arbitrary configuration  $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)$  of genotype frequencies in the population, it is particularly convenient to use the representation

$$\beta_{\text{HWE}}(\boldsymbol{\pi}) = \sum_{(x_1, x_2, x_3) \in \mathcal{X}_n} \phi(x_1, x_2, x_3) \frac{n!}{x_1! x_2! x_3!} \pi_1^{x_1} \pi_2^{x_2} \pi_3^{x_3}, \quad (9.57a)$$

with  $\mathcal{X}_n = \{(x_1, x_2, x_3) \mid x_j \in \mathbb{N}_0, \forall j = 1, 2, 3, x_1 + x_2 + x_3 = n\}$  and

$$\phi(x_1, x_2, x_3) = \begin{cases} 1 & x_2 \in (C_{\alpha; 2x_1+x_2}^1, C_{\alpha; 2x_1+x_2}^2) \\ \gamma_{\alpha; 2x_1+x_2}^\nu & \text{for } x_2 = C_{\alpha; 2x_1+x_2}^\nu, \nu = 1, 2 \\ 0 & x_2 < C_{\alpha; 2x_1+x_2}^1 \text{ or } x_2 > C_{\alpha; 2x_1+x_2}^2 \end{cases}. \quad (9.57b)$$

[For the definition of the critical constants  $C_{\alpha; 2x_1+x_2}^\nu, \gamma_{\alpha; 2x_1+x_2}^\nu$  recall (9.53).]

Of course, the alternatives of major interest are those which specify that the population sampled is in perfect HWE. The power against any such alternative is obtained by evaluating (9.57) at  $\boldsymbol{\pi} = (p^2, 2p(1-p), (1-p)^2)$  for some  $p \in [0, 1]$ . Given the significance level and the equivalence margins, the form of the corresponding function of  $p$  (called the power function of the test in the sequel, despite the fact that it is actually only a section through the power function as the whole being defined on the two-dimensional domain  $\{(\pi_1, \pi_2) \mid 0 \leq \pi_1 + \pi_2 \leq 1\}$ ) highly depends on the sample size. Figures 9.2a-c show the power curves both of the exact UMPU test and its nonrandomized version [obtained through setting  $\gamma_{\alpha; 2x_1+x_2}^\nu = 0$  for  $\nu = 1, 2$  and all  $(x_1, x_2)$  such that  $(x_1, x_2, n - x_1 - x_2) \in \mathcal{X}_n$  in (9.57a)] for  $\alpha = .05$ , the same specification of the equivalence band as in Figure 9.1, and  $n \in \{200, 400, 800\}$ . For each choice of the sample size, the power curves for both versions of the test are strictly unimodal and symmetric about  $p = 1/2$ . Furthermore, it becomes obvious from the graphs that the power of both the exact UMPU test and its nonrandomized counterpart approaches unity everywhere as  $n \rightarrow \infty$ , with a speed of convergence which highly depends on the population allele frequency  $p$ . For large  $n$ , the differences between the nonrandomized and the exact version mainly reduces to the fact that the power converges to zero rather than  $\alpha$  when  $p$  approaches one of the boundary points of the unit interval.

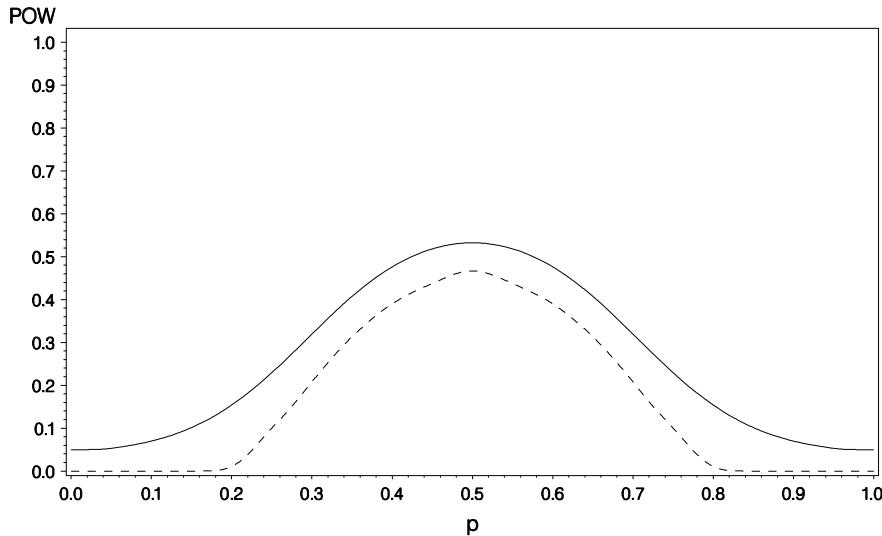


Figure 9.2a *Power of the conditional test for approximate compatibility with HWE against specific alternatives under which HWE is satisfied exactly, with  $\varepsilon_1 = 2/7$ ,  $\varepsilon_2 = 2/5$ ,  $\alpha = .05$  and  $n = 200$ . [Solid line: exact test; dashed line: nonrandomized version of the procedure.]*

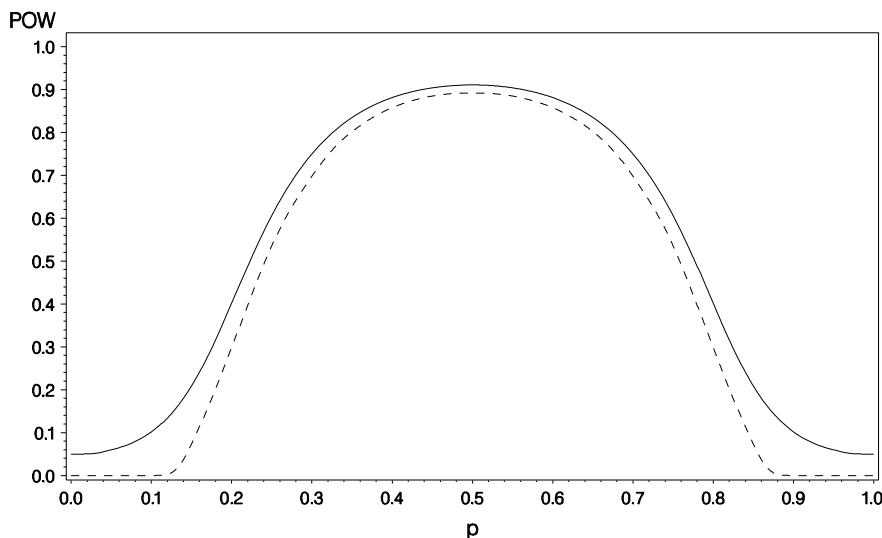
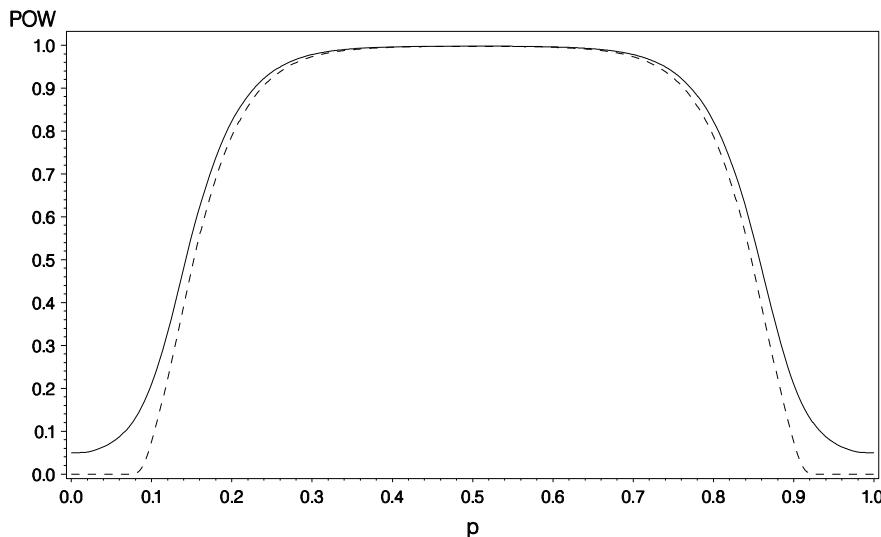


Figure 9.2b *Analogue of Figure 2b for sample size  $n = 400$ .*

Figure 9.2c Analogue of Figure 2a for  $n = 800$ .*Example 9.6*

- a) As regards the biological mechanism behind the genetic variant under study, the data set shown in Table 9.10 refers to a nonstandard case since the mutation consists in a loss of a single nucleotide rather than an exchange of some DNA base against another one. The total number of nondeleted alleles is 429, and for  $\alpha = .05$ ,  $\varepsilon = .40$ ,  $n = 475$ ,  $S = 429$ , the critical constants to be used in the UMPU test for approximate compatibility of the underlying population with HWE, are computed to be (by means of the program *gofhwex*)  $C_{\alpha;s}^1 = 215$ ,  $\gamma_{\alpha;s}^1 = .9353$ ,  $C_{\alpha;s}^2 = 257$ ,  $\gamma_{\alpha;s}^2 = .3552$ . Since the observed number of heterozygotes is  $X_2 = 233$ , the decision is in favor of equivalence of  $\omega$  to unity so that the SNP

Table 9.10 *Genotype distribution observed through determining the deletion (D) allele of the angiotensin I converting enzyme gene (ACE) in the sample of controls in a study of possible associations between ACE and dementia. (Data from Richard et al., 2001.)*

Genotype	II	ID	DD	$\sum$
Absolute Count	98	233	144	475
Population frequency	.2063	.4905	.3032	1.000

passes the check for approximate compatibility with HWE (regarding the associated genotype distribution in the population of nondemented controls).

- b) The data set shown in Table 9.11 refers to an ordinary SNP involving exchanges rather than deletions of nucleotides. Setting the significance level and the equivalence margins to the same values as before, the critical interval for  $X_2$  now to be used in the UMPU test (9.52) is found to be  $(C_{.05;511}^1, C_{.05;511}^2) = (167, 191)$ . Since the observed value of the number of patients genotyped GT falls neither in the interior nor to the boundary of this interval, we have to accept the null hypothesis of relevant deviations from HWE in the underlying distribution of genotypes.

Table 9.11 *Genotype distribution with respect to a SNP within the endothelial nitric oxide synthase (NOS3) gene obtained in  $n = 393$  individuals recruited as controls for a study on genetic risk factors for ischemic stroke. (Data from MacLeod et al., 1999.)*

Genotype	GG	GT	TT	$\Sigma$
Absolute Count	154	203	36	393
Population frequency	.3919	.5165	.0916	1.000

*Noninferiority version of the test: Excluding existence of a substantial deficit of heterozygotes*

In genetic association studies, an important issue is the risk of inducing spurious associations between some genotype and the disease status under study through recruiting the subjects from mixture populations exhibiting marked genetic heterogeneity. With a view to this problem (usually called the problem of population stratification in genetic epidemiology — see, e.g., Ziegler and König, 2006, § 10.4), a particularly relevant concern is to determine if there is excess homozygosity in the sample compared to the expected homozygosity under HWE. It is an immediate consequence of the concavity of the HWE curve [see Figure 9.1] that mixing different populations conforming with HWE always results in an excess of homozygotes, or in terms of the parametric function  $\omega$  selected here, to a deficit of heterozygotes. Thus, for purposes of ruling out that population stratification has to be taken into account, it suffices to carry out a test tailored to establish the (alternative) hypothesis that the underlying genotype distribution corresponds to a point in the  $(\pi_1, \pi_2)$ -plane which does not fall below the lower boundary of the equivalence band shown in Figure 9.1.

Wherever such considerations apply, the testing problem (9.49) should be replaced by what is called in clinical applications of equivalence assessment

procedures the noninferiority version of that testing problem. Specializing the general considerations of Chapter 2 to the present setting, this leads to rewriting the hypotheses of interest as

$$H_1 : \omega \leq 1 - \varepsilon_0 \text{ versus } K_1 : \omega > 1 - \varepsilon_0. \quad (9.58)$$

The modifications to the decision rule (9.52) and the equations (9.53) determining the optimal critical constants which are required for obtaining a UMPU solution of (9.58) are largely analogous to those which were required in the binomial two-sample setting with the odds ratio as the distributional parameter of interest [recall § 6.6]. More precisely speaking, the decision rule we have to apply in carrying out a UMPU level- $\alpha$  test for (9.58) reads:

$$\begin{cases} \text{reject } H_1 \text{ if } X_2 \geq C_{\alpha;s} + 2; \\ \text{reject } H_1 \text{ with probability } \gamma_{\alpha;s} \text{ if } X_2 = C_{\alpha;s}; \\ \text{accept } H_1 \text{ if } X_2 \leq C_{\alpha;s} - 2. \end{cases} \quad (9.59)$$

For determining the critical constants, there is only a single equation left to solve, namely

$$\sum_{x_2 \geq C_{\alpha;s} + 2} p_{\theta_0;s}(x_2) + \gamma_{\alpha;s} p_{\theta_0;s}(C_{\alpha;s}), \quad (9.60)$$

where  $p_{\theta_0;s}(x_2)$  has to be defined as in (9.53) with substitution of  $\theta_0 = 4(1 - \varepsilon_0)^2$  for  $\theta$ . For solving (9.60), another SAS macro is available in the **WKTSHEQ2 Source Code Package** under the filename **gofhwex\_1s.sas**. The R-code of the same computational procedure can be found at the same place opening the file **gofhwex\_1s.R**.

### *Example 9.7*

We illustrate the use of the noninferiority version (9.59) of the exact conditional test for absence of relevant deviations from HWE with an example taken from an association study of the impact of two APOE (Apolipoprotein E) promoter polymorphisms on the risk of Alzheimer's disease. Table 9.12 shows the results of genotyping one of the control groups involved in this study for 491 A→T, the fist of the SNPs which were selected for that purpose.

Table 9.12 *Results of genotyping a sample of  $n = 133$  controls in a study of the contribution of APOE promoter polymorphisms to the risk of developing Alzheimer's disease. (Data from Lambert et al., 2002, Table 3.)*

Genotype	TT	AT	AA	$\Sigma$
Absolute Count	6	53	74	133
Population frequency	.0451	.3985	.5564	1.000

For  $n = 133$  and  $s = 2 \cdot 6 + 53 = 65$ , the critical constants to be used in (9.59) at significance level  $\alpha = .05$  and for the same (left-hand) equivalence margin as before, i.e.,  $\varepsilon_0 = 2/7$  are obtained by means of the program `gofhwex_1s` to be  $C_{\alpha;s} = 51$ ,  $\gamma_{\alpha;s} = .9376$ . Thus, the observed value of  $X_2$  exceeds its critical lower bound implying that the null hypothesis of a relevant deficit of heterozygotes can be rejected. In contrast, under the same specifications, in the test (9.52) for approximate compatibility with HWE in the two-sided sense, the null hypothesis of relevant deviations from the model could not be rejected with the data of Table 9.12. Actually, running the program `gofhwex` yields  $(C_{.05;65}^1, C_{.05;65}^2) = (49, 51)$  for  $n = 133$  and thus a critical interval whose intersection with the support  $\mathcal{W}_s^n(X_2)$  of the conditional distribution of  $X_2$  [recall (9.54)] is empty.

The exact power function of the one-sided test can be computed by essentially the same rationale being behind the graphs displayed in Figures 9.2a–c. The only change is that the critical function  $\phi(\cdot)$  has to be specified on the basis of (9.60) rather than (9.52). Figure 9.3 shows the resulting power curve for a sample of size  $n = 400$ , together with that of the two-sided equivalence test. Except for the most extreme values of the allele frequency  $p$  in the population, the gain in power from discarding the upper boundary of the equivalence region is considerable.

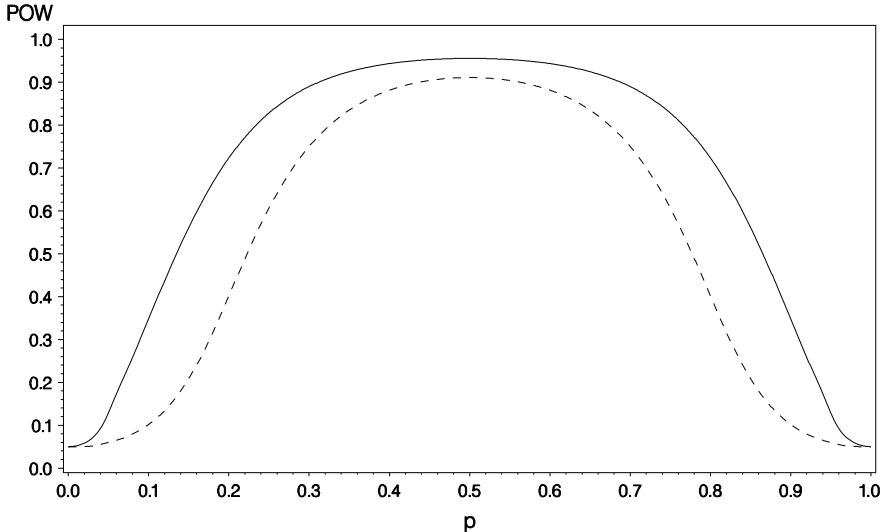


Figure 9.3 Comparison of the power curve for the one-sided version of the test [solid line] with that of the test for approximate compatibility with HWE in the two-sided sense, for  $\varepsilon_0 = \varepsilon_1 = 2/7$ ,  $\varepsilon_2 = 2/5$ ,  $\alpha = .05$  and  $n = 400$ .

### 9.4.3 Confidence-interval-based procedures for assessing HWE

Although the exact conditional testing procedure made available in § 9.4.2 is fully satisfactory from the statistical point of view, and its use for the analysis of association studies has been demonstrated in the genetic epidemiology literature (Goddard et al., 2009), the rationale behind it might not be easy to understand for applied researchers, and its practical implementation requires access to advanced statistical software. By such reasons the acceptance which the idea of replacing the classical lack-of-fit with goodness-of-fit tests for routine checks for HWE in large-scale genetic association studies will find, can be expected to improve substantially by providing procedures which are based on the principle of confidence interval inclusion as stated and proven as a general result in Chapter 3.1 of this book. An additional advantage of the availability of confidence procedures is that the resulting bounds and intervals can be used for decision making concerning large classes of testing problems involving hypotheses of very different forms. In particular, instead of assessing approximate compatibility of the underlying genotype distribution with HWE, each pair of confidence bounds can also be used for testing the traditional null hypothesis specifying that HWE holds exactly.

*Procedure based on the asymptotic distribution of the plug-in estimator of  $\omega$*

One natural approach to deriving an interval estimation procedure for the parameter  $\omega$  consists in studying the asymptotic distribution of its natural estimator. This is obtained by replacing each population genotype frequency with its empirical counterpart in the expression on the right-hand side of Equation (9.43). Denoting this plug-in estimator by  $\hat{\omega}_n$ , we have

$$\hat{\omega}_n = \frac{\hat{\pi}_2}{2\sqrt{\hat{\pi}_1\hat{\pi}_3}} = \frac{X_2/n}{2\sqrt{(X_1/n)(X_3/n)}}, \quad (9.61)$$

with  $\hat{\pi}_j = X_j/n$ ,  $j = 1, 2, 3$ . Since the parameter space of  $\omega$  is bounded on the left by zero, it is more promising to use the logarithm of  $\hat{\omega}_n$  as a pivot for determining the desired confidence bounds.

Based on the asymptotic distribution of  $\log \hat{\omega}_n$  which is again easily derived by means of the  $\delta$ -method (full details of this derivation are given in the Appendix to Wellek et al., 2009), one obtains the following pair of confidence bounds for  $\log \omega$  at one-sided asymptotic confidence level  $1 - \alpha$  each:

$$\begin{aligned} \underline{C}_{\alpha}^l(\hat{\pi}_n) &= \log \hat{\omega}_n - u_{1-\alpha} \sqrt{\frac{1}{n} \left( \frac{1 - \hat{\pi}_2}{4\hat{\pi}_1\hat{\pi}_3} + \frac{1}{\hat{\pi}_2} \right)}, \\ \bar{C}_{\alpha}^l(\hat{\pi}_n) &= \log \hat{\omega}_n + u_{1-\alpha} \sqrt{\frac{1}{n} \left( \frac{1 - \hat{\pi}_2}{4\hat{\pi}_1\hat{\pi}_3} + \frac{1}{\hat{\pi}_2} \right)}, \end{aligned} \quad (9.62)$$

where  $u_{1-\alpha}$  denotes the  $(1 - \alpha)$ -quantile of  $\mathcal{N}(0, 1)$ . The corresponding confidence interval for the relative heterozygosity  $\omega$  on the original scale is given by

$$(\underline{C}_\alpha(\hat{\pi}_n), \bar{C}_\alpha(\hat{\pi}_n)) = \left( \hat{\omega}_n / \exp \left\{ u_{1-\alpha} \sqrt{\frac{1}{n} \left( \frac{1 - \hat{\pi}_2}{4\hat{\pi}_1\hat{\pi}_3} + \frac{1}{\hat{\pi}_2} \right)} \right\}, \right. \\ \left. \hat{\omega}_n \exp \left\{ u_{1-\alpha} \sqrt{\frac{1}{n} \left( \frac{1 - \hat{\pi}_2}{4\hat{\pi}_1\hat{\pi}_3} + \frac{1}{\hat{\pi}_2} \right)} \right\} \right). \quad (9.63)$$

For applications, it is desirable to represent the rejection region of the interval inclusion test for the problem (9.49) based on (9.62) or (9.63) as a subset of the sample space of  $(\hat{\pi}_1, \hat{\pi}_2)$  or, alternatively, by way of graphical visualization in a DeFinetti diagram. Again, following the better part of the genetic epidemiology literature, we adopt the latter frame of reference. The desired geometrical representation is obtained in the following way: For each possible value  $s/2n$  ( $s = 0, \dots, 2n$ ) of the allele frequency  $\hat{p}$  observed in a sample of size  $n$ , the confidence bounds of (9.62) are considered as functions of  $\hat{\pi}_2$ , say  $\underline{c}_{\hat{p}; \alpha}(\cdot)$  and  $\bar{c}_{\hat{p}; \alpha}(\cdot)$ , which in view of  $\hat{p} = \hat{\pi}_1 + \hat{\pi}_2/2$  are given by

$$\underline{c}_{\hat{p}; \alpha}(\hat{\pi}_2) = \log \hat{\pi}_2 - (1/2) [\log(\hat{p} - \hat{\pi}_2/2) + \log(1 - \hat{p} - \hat{\pi}_2/2)] \\ - u_{1-\alpha} \sqrt{\frac{1}{n} \left( \frac{1 - \hat{\pi}_2}{4(\hat{p} - \hat{\pi}_2/2)(1 - \hat{p} - \hat{\pi}_2/2)} + \frac{1}{\hat{\pi}_2} \right)}, \quad (9.64a)$$

$$\bar{c}_{\hat{p}; \alpha}(\hat{\pi}_2) = \log \hat{\pi}_2 - (1/2) [\log(\hat{p} - \hat{\pi}_2/2) + \log(1 - \hat{p} - \hat{\pi}_2/2)] \\ + u_{1-\alpha} \sqrt{\frac{1}{n} \left( \frac{1 - \hat{\pi}_2}{4(\hat{p} - \hat{\pi}_2/2)(1 - \hat{p} - \hat{\pi}_2/2)} + \frac{1}{\hat{\pi}_2} \right)}. \quad (9.64b)$$

In order to determine the set of all  $\hat{\pi}_2$  such that the asymptotic confidence interval assigned to the point  $(\hat{p}, \hat{\pi}_2)$  satisfies the condition for rejecting the null hypothesis  $H : \omega \leq 1 - \varepsilon_1 \vee \omega \geq 1 + \varepsilon_2$ , we have to solve the equations

$$\underline{c}_{\hat{p}; \alpha}(\hat{\pi}_2) = \log(1 - \varepsilon_1), \quad 0 \leq \hat{\pi}_2 \leq 2 \min\{\hat{p}, 1 - \hat{p}\} \quad (9.65a)$$

and

$$\bar{c}_{\hat{p}; \alpha}(\hat{\pi}_2) = \log(1 + \varepsilon_2), \quad 0 \leq \hat{\pi}_2 \leq 2 \min\{\hat{p}, 1 - \hat{p}\}. \quad (9.65b)$$

Neither one of the functions  $\underline{c}_{\hat{p}; \alpha}(\cdot)$  and  $\bar{c}_{\hat{p}; \alpha}(\cdot)$  is strictly monotonic so that the above equations typically admit several solutions. The natural way of determining the limits of the  $\hat{p}$ -section of the rejection region consists of selecting the solution that is nearest to the ordinate  $2\hat{p}(1 - \hat{p})$  of the corresponding point on the HWE curve on both sides. Denoting these values by  $\hat{\pi}_{2;l}(\hat{p}; \alpha)$  [lower critical bound] and  $\hat{\pi}_{2;u}(\hat{p}; \alpha)$  [upper critical bound], respectively, it is possible (in particular for small sample sizes) that the interval determined by them is empty. Furthermore, for values of  $\hat{p}$  near the endpoints of the unit interval, a solution to equation (9.65a) and/or (9.65b) need not exist. In all these degenerate cases, we use the convention of setting  $\hat{\pi}_{2;l}(\hat{p}; \alpha) =$

$\hat{\pi}_{2;u}(\hat{p}; \alpha) = 2\hat{p}(1 - \hat{p})$ . Figure 9.4 shows the result of this construction, again for the choice  $\varepsilon_1 = 5/7$ ,  $\varepsilon_2 = 7/5$  of the equivalence margins. The graph relates to the usual significance level  $\alpha = .05$  and a sample of size  $n = 200$ . In addition to the critical region of the asymptotic test based on (9.65a-b), it also shows the equivalence region specified under the alternative hypothesis.

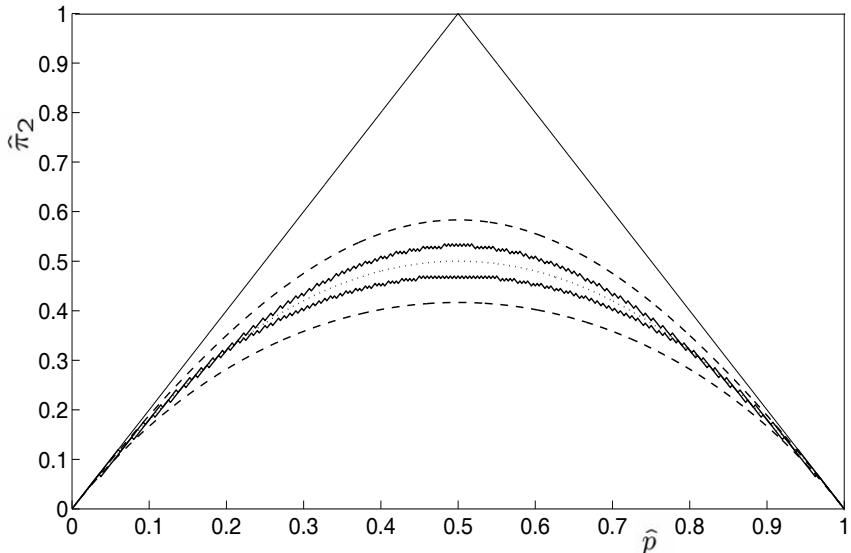


Figure 9.4 *Rejection region of the test for goodness of fit with HWE based on pairs of asymptotic 95%-confidence bounds for the relative excess heterozygosity parameter  $\omega$  computed from a sample of size  $n = 200$ . [Dotted central line characterizes distributions being in exact HWE; outer dashed lines demarcate the theoretical equivalence band for  $\omega$  obtained by specifying  $\varepsilon_1 = 2/7$ ,  $\varepsilon_2 = 2/5.$ ]*

#### *Exact conditional confidence limits*

The conditional distributions of the number  $X_2$  of heterozygotes given an arbitrarily fixed value  $s \in \{0, \dots, 2n\}$  for the total number  $S$  of A-alleles, constitute a one-parameter exponential family having in particular monotonically increasing likelihood ratios. Hence, exact confidence bounds of any prespecified level  $1 - \alpha$  can be computed for  $\theta$  (and thus also  $\omega$ ) applying the usual procedure for STP<sub>2</sub> families (cf. Lehmann and Romano, 2005, §3.5) to the distribution given by (9.56) for each realized value  $s$  of the statistic  $S$  being sufficient for the nuisance parameter  $\vartheta$  as defined in (9.45). Since each such distribution is of the discrete type and we do not want to rely on extraneous randomization as would be required for exhausting the non-coverage proba-

bility  $\alpha$ , we perform the construction in a conservative version fully analogous to that carried out in Bickel and Doksum (2001, pp. 244–6).

In order to give a precise description of the procedure, we start with choosing an arbitrarily fixed point  $\theta_0$  in the range space of the basic parameter  $\theta$ . Considering the  $p$ -values of the nonrandomized conditional tests of the one-sided null hypotheses  $H_l : \theta \geq \theta_0$  and  $H_r : \theta \leq \theta_0$  corresponding to this value as functions of  $\theta_0$  and denoting them by  $g_{x_2; s, n}^{(l)}(\theta_0)$  and  $g_{x_2; s, n}^{(r)}(\theta_0)$ , respectively, we have

$$g_{x_2; s, n}^{(l)}(\theta_0) = \sum_{x'_2 \in \mathcal{X}_{n, s}^{(2)} \cap [0, x_2]} K_{n, s}(x'_2) \theta_0^{x'_2/2} / \sum_{x'_2 \in \mathcal{X}_{n, s}^{(2)}} K_{n, s}(x'_2) \theta_0^{x'_2/2}, \quad (9.66a)$$

$$g_{x_2; s, n}^{(r)}(\theta_0) = \sum_{x'_2 \in \mathcal{X}_{n, s}^{(2)} \cap [x_2, \infty)} K_{n, s}(x'_2) \theta_0^{x'_2/2} / \sum_{x'_2 \in \mathcal{X}_{n, s}^{(2)}} K_{n, s}(x'_2) \theta_0^{x'_2/2}. \quad (9.66b)$$

The monotone likelihood property of the underlying family of (conditional) distributions ensures (by another classical result stated, e.g., in Lehmann and Romano, 2005, as Theorem 3.4.1(ii)) that the second function is strictly increasing, implying strict decreasingness of the first function. Furthermore, it is obvious from their definitions that they are both continuous. Thus, each of the equations

$$g_{x_2; s, n}^{(l)}(\theta_0) = \alpha, \quad 0 \leq \theta_0 < \infty; \quad g_{x_2; s, n}^{(r)}(\theta_0) = \alpha, \quad 0 \leq \theta_0 < \infty \quad (9.67)$$

has a unique solution. The first of them yields the desired upper confidence bound for  $\theta$  with exact conditional coverage probability  $\geq 1 - \alpha$ . The solution to  $g_{x_2; s, n}^{(r)}(\theta_0) = \alpha$  is a lower confidence limit with the same property. Of course, both values depend on  $x_2$ ,  $s$  and  $n$  in addition to  $\alpha$  which is made explicit by using the notation  $\bar{\theta}_{\alpha; n}(x_2, s)$  and  $\underline{\theta}_{\alpha; n}(x_2, s)$  for the exact conditional upper and lower confidence bound, respectively. The corresponding confidence bounds  $\underline{\omega}_{\alpha; n}(x_2, s)$ ,  $\bar{\omega}_{\alpha; n}(x_2, s)$  for the relative excess heterozygosity  $\omega$  are eventually obtained through rescaling, i.e., as  $\underline{\omega}_{\alpha; n}(x_2, s) = \sqrt{\underline{\theta}_{\alpha; n}(x_2, s)}/2$ ,  $\bar{\omega}_{\alpha; n}(x_2, s) = \sqrt{\bar{\theta}_{\alpha; n}(x_2, s)}/2$ . Details of the corresponding computational procedure are given in Wellek et al. (2009), and a set of SAS/IML modules for its implementation can be found in the **WKTSHEQ2** Source Code Package [filename: `cf_reh_exact.sas`].

#### *Confidence limits based on exact conditional mid-p-values*

The confidence bounds obtained by means of the procedure described above are exact but conservative, in the sense that the conditional coverage probabilities  $P_\theta[\underline{\theta}_{\alpha; n}(X_2, S) < \theta | S = s]$ ,  $P_\theta[\bar{\theta}_{\alpha; n}(X_2, S) > \theta | S = s]$  might be larger than required, i.e.,  $> 1 - \alpha$ . Such conservatism is clearly an intrinsic

property of confidence bounds based on families of nonrandomized tests for parameters of discrete distributions. A frequently recommended (see, e.g., Agresti, 2002, § 1.4.5) approach to reducing or even eliminating this conservatism consists of modifying the  $p$ -values involved by assigning a weight of  $1/2$  (instead of unity) to the probability that the test statistic takes on the observed value. In the present context, this so-called mid- $p$ -value approach leads to replacing  $g_{x_2; s, n}^{(l)}(\theta_0)$  and  $g_{x_2; s, n}^{(r)}(\theta_0)$  in (9.66a–b) with

$$\begin{aligned} g_{x_2; s, n}^{(l*)}(\theta_0) &\equiv \frac{1}{2} \left[ g_{x_2-2; s, n}^{(l)}(\theta_0) + g_{x_2; s, n}^{(l)}(\theta_0) \right] \quad \text{and} \\ g_{x_2; s, n}^{(r*)}(\theta_0) &\equiv \frac{1}{2} \left[ g_{x_2; s, n}^{(r)}(\theta_0) + g_{x_2+2; s, n}^{(r)}(\theta_0) \right], \end{aligned} \quad (9.68)$$

respectively.

The derivation of a confidence procedure from these modified  $p$ -values is fully analogous to the construction of exact confidence limits. In what follows, the resulting confidence bounds for  $\omega$  are denoted  $\underline{\omega}_{\alpha; n}^*(x_2, s)$  and  $\bar{\omega}_{\alpha; n}^*(x_2, s)$ . The source code of a SAS/IML program for computing these statistics is stored in the file named `cf_reh_midp.sas`. It should be noted that the mid- $p$ -value-based bounds are not guaranteed to maintain the target confidence level of  $1 - \alpha$ . However, numerical results from simulation studies of a variety of discrete families of distributions show that they are typically still conservative but to a markedly lesser extent than their exact counterparts.

### *Illustration*

For illustration, we reanalyze the controls samples of a selection of the genetic association studies that were extracted by Wittke-Thompson et al. (2005) for consideration of the appropriate use of the classical tests for lack of fit to HWE. The sample sizes range from  $n = 58$  to  $n = 801$ , and the observed relative excess heterozygosities from  $\hat{\omega}_n = 0.7667$  to  $\hat{\omega}_n = 1.2964$ . The results of estimating in these samples 95%-confidence bounds for the relative excess heterozygosity by means of the three methods described above are shown in Table 9.13 in ascending order with respect to the observed value of the point estimate  $\hat{\omega}_n$  of the relative excess heterozygosity. Specifying the hypothetical equivalence range for  $\omega$  as before, approximate compatibility with HWE can be concluded by means of the asymptotic version of the interval inclusion test at nominal significance level .05 for 10 out of 19 samples listed in the table. Remarkably, only two of these equivalence decisions (relating to studies #8 and 12) is affected by changing the method of constructing the confidence bounds to  $\omega$ . More extensive comparisons between the asymptotic, exact and mid- $p$ -based confidence limits revealed that the differences between the results of the mid- $p$  and the asymptotic approach are distinctly smaller than

those between either of them and the exact approach. This corroborates the rule stated by Agresti (2002, p. 21) that the mid-p method is markedly less conservative than the exact approach.

Table 9.13 95%-Confidence bounds for the relative excess heterozygosity in the control subjects of 19 association studies reported by Wittke-Thompson et al. (2005).

#	Study	$X_1/X_2/X_3$	$\hat{\omega}_n$	Lower Confidence Bound			Upper Confidence Bound		
				asympt.	exact	mid-p	asympt.	exact	mid-p
1	3/23/75	0.7667	0.4235	0.3851	0.4168	1.3878	1.6439	1.4570	
2	75/35/6	0.8250	0.5280	0.4946	0.5213	1.2888	1.3937	1.3057	
3	61/212/237	0.8816	0.7487	0.7403	0.7470	1.0381	1.0465	1.0370	
4	66/80/31	0.8843	0.6841	0.6647	0.6790	1.1431	1.1638	1.1390	
5	297/258/71	0.8883	0.7651	0.7581	0.7637	1.0314	1.0384	1.0304	
6	57/192/197	0.9059	0.7632	0.7538	0.7612	1.0754	1.0848	1.0739	
7	197/349/186	0.9116	0.8071	0.8016	0.8058	1.0296	1.0337	1.0285	
8	58/97/48	0.9192	0.7292	0.7119	0.7246	1.1588	1.1752	1.1547	
9	77/41/6	0.9537	0.6186	0.5835	0.6127	1.4706	1.5891	1.4917	
10	193/393/215	0.9646	0.8587	0.8534	0.8574	1.0837	1.0876	1.0825	
11	98/233/144	0.9807	0.8421	0.8334	0.8400	1.1421	1.1492	1.1402	
12	23/83/67	1.0572	0.8082	0.7844	0.8024	1.3828	1.4122	1.3791	
13	62/65/15	1.0657	0.7797	0.7504	0.7727	1.4566	1.5020	1.4550	
14	16/76/79	1.0688	0.7966	0.7707	0.7909	1.4341	1.4755	1.4336	
15	112/132/33	1.0856	0.8740	0.8573	0.8704	1.3485	1.3678	1.3462	
16	27/26/5	1.1189	0.6691	0.6095	0.6522	1.8710	2.0542	1.8902	
17	114/560/533	1.1359	1.0179	1.0130	1.0171	1.2676	1.2725	1.2671	
18	174/45/2	1.2061	0.6397	0.6109	0.6554	2.2742	3.0015	2.5415	
19	98/77/9	1.2964	0.9206	0.8886	0.9176	1.8256	1.9182	1.8401	

#### Power, sample size and equivalence margin specification

A major advantage of the interval inclusion test with asymptotic confidence bounds is that it allows for simple yet sufficiently accurate approximations to its power against arbitrary alternatives under which there holds  $\omega = 1$ . In fact, if we denote the probability of rejecting the null hypothesis  $H$  of (9.49) in the test based on an asymptotic confidence interval computed from (9.63) in a sample from a population satisfying  $\pi_2 = 2\sqrt{\pi_1(1 - \pi_1 - \pi_2)}$  by  $\beta_{\varepsilon_1, \varepsilon_2; \alpha}^*(\pi_1, n)$ , then it can be shown (for a rigorous proof see Wellek et al., 2009) that for sufficiently large  $n$ , we can write

$$\begin{aligned} \beta_{\varepsilon_1, \varepsilon_2; \alpha}^*(\pi_1, n) &\approx \Phi[\sqrt{n}\tilde{\varepsilon}_2 \cdot 2\sqrt{\pi_1}(1 - \sqrt{\pi_1}) - u_{1-\alpha}] + \\ &\quad \Phi[\sqrt{n}\tilde{\varepsilon}_1 \cdot 2\sqrt{\pi_1}(1 - \sqrt{\pi_1}) - u_{1-\alpha}] - 1, \end{aligned} \quad (9.69)$$

where

$$\tilde{\varepsilon}_1 = -\log(1 - \varepsilon_1), \quad \tilde{\varepsilon}_2 = \log(1 + \varepsilon_2). \quad (9.70)$$

Choosing the equivalence interval symmetric on the log-scale through specifying  $\varepsilon_2 = \varepsilon$ ,  $\varepsilon_1 = 1 - 1/(1 + \varepsilon)$  with  $\varepsilon > 0$  and treating (9.69) like an ordinary (exact) equation to be solved for  $n$ , the minimum sample size required for ensuring a power of  $1 - \beta$ , say, can be computed from

$$n = \frac{(u_{1-\beta/2} + u_{1-\alpha})^2}{(2\sqrt{\pi_1}(1 - \sqrt{\pi_1}))^2 \tilde{\varepsilon}^2}. \quad (9.71)$$

The sample size formula (9.71) is of considerable interest under a number of different aspects. In particular, it provides a basis for rationally justifying the recommendation of setting the equivalence margin  $\varepsilon$  at the value .40 which was used throughout this subsection in all examples etc. Elaborating this remark we start with making the largely plain statement that a positive result of any equivalence testing procedure is scientifically the more satisfactory the smaller the equivalence margins have been chosen. Notwithstanding this fact, it is clearly not reasonable to narrow the equivalence region to an extent that an unacceptably large proportion of SNPs would fail to be selected as HWE-compatible although they perfectly satisfy the model. These general considerations suggest to choose the equivalence margin as small as we can do under the following *restrictions concerning* (i) *power*, (ii) *allele frequency*, and (iii) *sample size*:

- (i) Over the whole range of values of frequency  $f_{AMI}$  of the allele of major interest to be taken into account, the probability of rejecting lack of fit under a genotype distribution which exactly conforms to the model, must not be smaller than 90%.
- (ii) The power to detect an association in a case-control or cohort study is low for low values of  $f_{AMI}$ . In current genomewide association studies (GWAs), the proportion of SNPs with  $f_{AMI} < 0.1$  associated with a disease is extremely low (see Johnson and O'Donnell, 2009). For SNPs defined eligible for inclusion in association studies, we therefore require  $0.1 \leq f_{AMI} \leq 0.5$ , corresponding to  $0.01 \leq \pi_1 \leq 0.25$  under HWE.
- (iii) Current GWAs have up to 3,000 control subjects. The only exception are studies from Iceland, where more than 10,000 subjects served as controls for a variety of investigated diseases. Therefore, the required sample size should not exceed 3,000.

Now, as a function of  $\sqrt{\pi_1}$  which under HWE coincides with  $f_{AMI}$ , the expression on the right-hand side of (9.71) is clearly decreasing on  $(0, 1/2]$ . Hence, the smallest equivalence margin leading to a sample size of  $n = 3000$  for SNPs satisfying  $0.1 \leq f_{AMI} \leq 0.5$ , is calculated to be  $\tilde{\varepsilon} = \frac{u_{1-\beta/2} + u_{1-\alpha}}{2 \cdot .1 \cdot .9 \cdot \sqrt{3000}}$

$= \frac{2 \cdot u_{.95}}{2 \cdot .09 \cdot \sqrt{3000}} = .3337$ . Retransforming the equivalence range  $-.3337 < \log \omega < .3337$  to the original scale yields an equivalence margin of  $e^{.3337} - 1 = .3961$  for  $\omega$  itself so that we indeed end up with the proposal to specify  $\varepsilon = .4$ .

### Discussion

As compared with the UMPU conditional test, the confidence-interval-based approach derived here provides the following major advantages:

- conceptual lucidity;
- ease of computational implementation;
- flexibility with regard to changes of the equivalence margins and the form of the hypothesis to be established;
- existence of simple closed formula for power and sample size.

The first of these points is not specific to the problem of establishing approximate compatibility of a theoretical genotype distribution with the HWE model, and in general, conceptual simplicity of the confidence interval inclusion approach to equivalence testing does not ensure simplicity of computational implementation, as the exact and mid-p-value versions of the procedure show. However, the asymptotic version involves only elementary algebraic operations which makes it almost ideally suited as a tool for the working genetic epidemiologist. Moreover, as additional numerical results presented in Wellek et al. (2009) suggest, the corresponding rejection region is surprisingly similar to that of the exact conditional test even for moderate sample sizes. Even the most conservative of the confidence-limit based testing procedures yields rejection regions which are only slightly narrower than the critical region of the nonrandomized version of the UMPU test. This implies that in the present case, the loss in power entailed as the price to pay for computational simplicity is largely negligible from a practical point of view.

One of the major advantages of the confidence-limit-based approach to HWE assessment over that via application of the exact conditional test is the ease of adaptation to different specifications of the margins under the hypothesis to be established. Actually, all that happens in applying one of the interval inclusion rules when the margins are changed, is that in the very final step a potentially different decision has to be taken. Computationally, interval inclusion rules remain completely invariant against numerical re-specification of the margins. In contrast, carrying out the exact conditional test with some other value of the equivalence margins requires re-execution of the full computational procedure for determining the bounds of the critical interval to  $X_2$ .

Last but not least, for the UMPU test, power calculation is computationally quite demanding even for small or moderate values of  $n$ , and no analogue of formula (9.69) is available.

# 10

---

## *The assessment of bioequivalence*

---

### 10.1 Introduction

Although the overwhelming majority of the existing literature on statistical tests for equivalence hypotheses refers to the analysis of comparative bioavailability trials (often also named bioequivalence trials, for the sake of brevity), the testing problems to be dealt with in the context of bioequivalence (BE) assessment are rather special in nature. This is because the regulatory authorities set up strict guidelines for the planning and analysis of bioequivalence studies quite soon after this type of trial had begun to play an important role in pharmaceuticals and clinical pharmacology. According to these rules, every bioequivalence study has to be conducted as a phase-I clinical trial, which means that the subjects are healthy volunteers rather than patients suffering from a disease considered to be an indication for administering the drug under study. Furthermore, each subject has to be given several formulations of that drug in an experiment following a crossing-over scheme allowing for two different trial periods at least. The outcome measure to be used for characterizing the biological activity of the drug administered at the beginning of the respective period, has to be chosen from a very short list of pharmacokinetic parameters. The distributions of these outcome measures are of the continuous type throughout, modeled in the majority of cases in a fully parametric way, usually after logarithmic transformation of the primary measurements obtained during the trial.

A complete exposition covering all statistical facets of the planning and analysis of BE trials gives sufficient material for a monograph of its own (see Chow and Liu, 2008; Hauschke et al., 2007; Patterson and Jones, 2005) so that we confine ourselves in this chapter to a discussion of the standard case of a bioequivalence trial involving two periods and as many different drug formulations. Actually, the restriction to this standard design is far from severe for practical purposes since real BE trials involving more than two periods and/or treatments are still a rare exception. Furthermore, even with regard to the  $2 \times 2$  case, the controversy about the most appropriate statistical criterion for BE assessment is still very far from being settled and seems to be a never ending story within (bio-)statistical science. Not surprisingly, the major issue of the ongoing discussion refers to the conceptual level. In fact, from

the mathematical and statistical perspective, it can be stated that there are very satisfactory solutions to an impressive manifold of problems put forward in this context. However, only a small number of them yield an intuitively plausible translation of the bio-medical question to be answered at the end of a BE trial into the formalism of decision making between statistical hypotheses. Thus, it can be argued that statistical BE assessment has developed into a field where more mathematically sophisticated solutions exist than problems worth the effort entailed in deriving solutions. In view of this situation, we prefer to concentrate on a careful discussion of the pros and cons of the most frequently applied approaches. Even within the scope of methods tailored for the analysis of BE trials following the standard design, we do not aim at a complete coverage of the existing literature. In return, the reader will find also some new ideas on the topic which lead beyond the solutions encouraged by the official guidelines or may even suggest to question the adequacy of the rationale behind the latter.

The general frame of reference for the whole chapter is given by a conventional nonreplicated crossover design as had been considered from a totally different perspective in § 9.3.2. In the present context, the treatments under comparison are two drug products denoted “Test” ( $T$ ) and “Reference” ( $R$ ), adopting the usage almost universally followed in the field of BE assessment. Typically and most commonly,  $T$  is a generic drug to be approved for the market, and  $R$  a reference listed drug containing chemically identical active ingredients. The total sample of subjects recruited to a standard BE trial consists of  $N$  healthy volunteers randomly assigned to one of the two possible sequence groups  $T/R$  and  $R/T$ . Although the randomization is usually performed under the restriction of generating a strictly balanced layout, by technical difficulties it might happen that a few subjects have been lost from one or the other group due to providing incomplete or otherwise insufficient pharmacokinetic data. Thus, we allow for arbitrary sample sizes  $m$  ( $\leftrightarrow T/R$ ) and  $n$  ( $\leftrightarrow R/T$ ) in both sequence groups.

The basic data set consists of two vectors per subject, whose components give the ordinates of the concentration-time profile recorded in both periods of the trial. (In the majority of BE studies, the concentration of the active ingredient of interest is measured in the blood plasma or serum.) Figure 10.1 shows a schematic example of such a pair of observed concentration-time profiles. Instead of analyzing the individual profiles as multivariate data, they are usually reduced in a preliminary step to a single real value by calculating one of the following pharmacokinetic characteristics:

- area under the profile as a whole ( $AUC$ )
- maximum concentration ( $C_{\max}$ )
- time at which the peak concentration  $C_{\max}$  was measured ( $t_{\max}$ ).

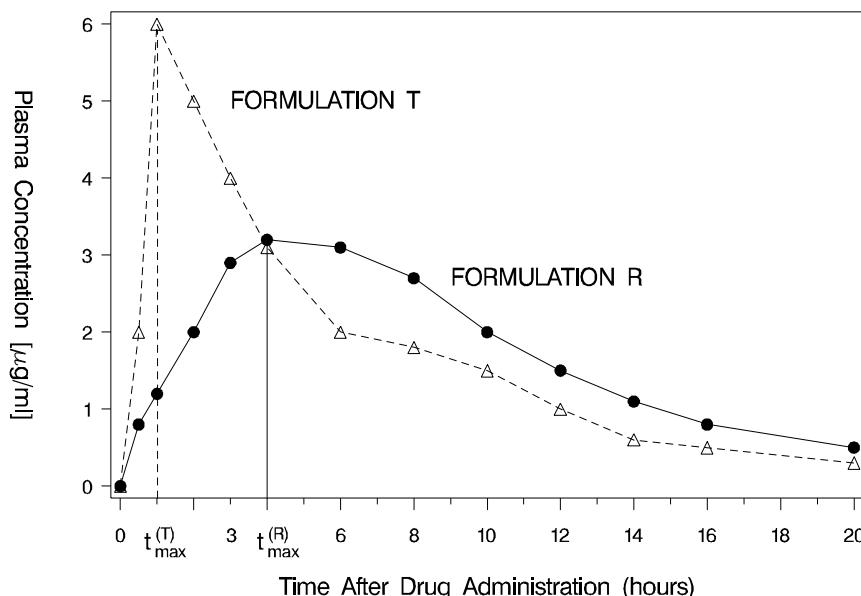


Figure 10.1 Example of a pair of concentration-time profiles observed in a given subject during both periods of a standard bioequivalence trial. (Redrawn from Rodda, 1990, with kind permission by Marcel Dekker, Inc.)

Commonly,  $AUC$  and  $C_{\max}$  are considered as alternative estimates of the extent of the absorption process induced in a trial subject, whereas  $t_{\max}$  is interpreted as measuring its rate. All three pharmacokinetic parameters are accepted as a reasonable choice for selecting a reference measure of bioavailability to be used as the endpoint variable in both periods and all subjects enrolled in the trial.

The data set eventually to be analyzed for the purpose of BE assessment consists of two independent samples  $(X_{11}, X_{21}), \dots, (X_{1m}, X_{2m}), (Y_{11}, Y_{21}), \dots, (Y_{1n}, Y_{2n})$  of pairs of one-dimensional random variables such that

$X_{ki}$  = bioavailability measured in the  $i$ th subject ( $i = 1, \dots, m$ )  
of sequence group  $T/R$  during the  $k$ th period ( $k = 1, 2$ );

$Y_{kj}$  = bioavailability measured in the  $j$ th subject ( $j = 1, \dots, n$ )  
of sequence group  $R/T$  during the  $k$ th period ( $k = 1, 2$ ).

In order to ensure at least approximate normality of the distributions of these variables, it is generally recommended (see FDA, 2001, § VI.A) to determine their values by taking logarithms of the bioavailabilities computed from the individual concentration-time profiles. Thus, if the  $AUC$  has been selected as

the measure of bioavailability of interest, then we have  $X_{11} = \log(A_{11})$ ,  $Y_{11} = \log(B_{11})$  with  $A_{11}$  and  $B_{11}$  denoting the area under the profile obtained in Period 1 in the 1st subject of group  $T/R$  and  $R/T$ , respectively, and so on.

If not explicitly stated otherwise, we assume throughout (following once more the current official guidance for conducting *in vivo* BE studies) that the log-bioavailabilities  $X_{ki}$ ,  $Y_{kj}$  satisfy the basic parametric model for the analysis of  $2 \times 2$  crossover trials [see (9.28–31)] with the following slight modifications: The within-subject variability (usually interpreted as reflecting measurement error) is this time allowed to depend on treatment rather than period so that we let

$$\text{Var} \left[ \varepsilon_{1i}^{(1)} \right] = \text{Var} \left[ \varepsilon_{2j}^{(2)} \right] \equiv \sigma_{eT}^2 \quad (10.1a)$$

and

$$\text{Var} \left[ \varepsilon_{2i}^{(1)} \right] = \text{Var} \left[ \varepsilon_{1j}^{(2)} \right] \equiv \sigma_{eR}^2 \quad (10.1b)$$

for all  $i = 1, \dots, m$  and  $j = 1, \dots, n$ . Furthermore, the direct treatment effects are now called formulation effects and written  $\phi_T, \phi_R$  instead of  $\phi_A, \phi_B$ . Except for § 10.5, we adopt the conventional view (supported by empirical evidence — see D’Angelo et al., 2001) that in analyzing a properly conducted BE trial, nonexistence of carryover effects can be taken for granted *a priori*, due to a careful procedure of eliminating the residual activity of the drug administered in the initial period during the washing-out interval. Of course, having renamed the treatment effects, this leads to the simplified model equations

$$\begin{aligned} E(X_{1i}) &= \mu_1 = \omega + \phi_T + \pi_1, & E(X_{2i}) &= \mu_2 = \omega + \phi_R + \pi_2 \\ && (i = 1, \dots, m) & \end{aligned} \quad (10.2a)$$

$$\begin{aligned} E(Y_{1j}) &= \nu_1 = \omega + \phi_R + \pi_1, & E(Y_{2j}) &= \nu_2 = \omega + \phi_T + \pi_2 \\ && (j = 1, \dots, n). & \end{aligned} \quad (10.2b)$$

In a broad sense, it is clear what bioequivalence of the two drug formulations means: sufficient similarity of the distributions of bioavailabilities measured after administration of  $T$  and  $R$ , respectively. However, there are obviously many ways of making this idea precise in terms of a hypothesis to be established by means of some suitable test of significance. The still most frequently used criterion focuses on the difference  $\phi_T - \phi_R$  of formulation effects. In the guidelines and the biostatistical literature on BE assessment, testing procedures allowing to decide whether or not a criterion of this type is satisfied, are called tests for “average bioequivalence.” Due to their overwhelming importance for biostatistical practice, the best established approaches to the assessment of average BE are discussed in the first of the core sections of this chapter. A totally different philosophy underlies the methods presented in § 10.3: “Individual bioequivalence” in the “probability-based” sense requires that in the underlying population, there is a sufficiently high proportion of

subjects such that the bioavailabilities induced by  $T$  and  $R$  in the same individual are identical except for some pharmacologically irrelevant difference. Methods of testing for “population bioequivalence,” which will be discussed in § 10.4, aim at establishing that the distributions of bioavailabilities associated with  $T$  and  $R$ , are similar *both* with respect to location and variability. Finally, in § 10.5 we argue that any marked difference between the *joint* distribution of the  $(X_{1i}, X_{2i})$  and that of the  $(Y_{1j}, Y_{2j})$  is at variance with equivalence of both drug formulations with respect to the given measure of bioavailability. Accordingly, the testing procedure derived in this final section, are bivariate analogues of the univariate two-sample tests for continuous data discussed in Chapter 6.

---

## 10.2 Methods of testing for average bioequivalence

### 10.2.1 Equivalence with respect to nonstandardized mean bioavailabilities

The traditional approach to BE assessment recommended even in the most recent official guidelines on statistical methods for the analysis of comparative bioavailability trials (FDA, 2001) as a standard procedure, focuses on the raw formulation effects  $\phi_T, \phi_R$  as appearing in the model equations (10.2). In the simplest version of the concept, average BE is defined to be satisfied if and only if the true value of  $\phi_T - \phi_R$  lies in a sufficiently small neighborhood of zero. By the so-called 80–125% convention, the radius of this neighborhood is specified

$$\delta_0 = \log(5/4) = .2231, \quad (10.3)$$

which can be motivated as follows: Suppose the period effects can also be dropped from the model equations (which, according to experience, holds true for the majority of real BE studies), and the within-subject variance is the same under both treatments (which is a standard assumption in the classical model for the  $2 \times 2$  crossover). Then, the distribution of the non-logarithmic bioavailability induced by formulation  $T$  and  $R$  is lognormal with parameters  $\omega + \phi_T, \sigma_S^2 + \sigma_e^2$  and  $\omega + \phi_R, \sigma_S^2 + \sigma_e^2$ , respectively. Hence (cf. Johnson et al., 1994, p. 212) the expected values, say  $\mu_T^*$  and  $\mu_R^*$ , of these distributions are given by  $\mu_T^* = \exp\{\omega + \phi_T + (\sigma_S^2 + \sigma_e^2)/2\}$ ,  $\mu_R^* = \exp\{\omega + \phi_R + (\sigma_S^2 + \sigma_e^2)/2\}$ , so that their ratio is  $\mu_T^*/\mu_R^* = \exp\{\phi_T - \phi_R\}$ . This shows that under the additional restrictions made explicit above, the condition  $|\phi_T - \phi_R| < \log(5/4)$  is the same as  $4/5 < \mu_T^*/\mu_R^* < 5/4$ , and it has become part of good pharmaceutical practice to consider the latter sufficient for neglecting the differences in the response to both formulations of the drug.

Keeping the specification (10.3) in mind, assessment of average BE in the usual sense requires that we are able to perform a valid test of

$$H : |\phi_T - \phi_R| \geq \delta_o \quad \text{versus} \quad K : |\phi_T - \phi_R| < \delta_o. \quad (10.4)$$

In order to see how such a test can be constructed, it is helpful to recall the standard parametric procedure of testing the conventional null hypothesis  $\phi_T = \phi_R$  in a  $2 \times 2$  crossover design satisfying strict additivity of direct treatment and period effects. The basic step enabling an easy derivation of such a test consists in reducing the bioavailabilities  $(X_{1i}, X_{2i}), (Y_{1j}, Y_{2j})$  observed in the successive trial periods, to within-subject differences measuring the change in response level from Period 1 to Period 2 (to be computed in that order throughout, irrespective of the sequence group to which a given subject has been assigned). Denoting this difference by  $X_i^-$  ( $\leftrightarrow$  Group  $T/R$ ) and  $Y_j^-$  ( $\leftrightarrow$  Group  $R/T$ ), respectively, we have

$$X_i^- = X_{1i} - X_{2i} \quad \forall i = 1, \dots, m, \quad Y_j^- = Y_{1j} - Y_{2j} \quad \forall j = 1, \dots, n, \quad (10.5)$$

and it readily follows from (10.1) and (10.2) that

$$X_i^- \sim \mathcal{N}(\phi_T - \phi_R + \pi_1 - \pi_2, \sigma^2) \quad \forall i = 1, \dots, m, \quad (10.6a)$$

$$Y_j^- \sim \mathcal{N}(\phi_R - \phi_T + \pi_1 - \pi_2, \sigma^2) \quad \forall j = 1, \dots, n, \quad (10.6b)$$

with

$$\sigma^2 = \sigma_{eT}^2 + \sigma_{eR}^2. \quad (10.7)$$

Clearly, (10.6) implies that the within-subject differences  $X_i^-$  and  $Y_j^-$  satisfy the assumptions of the ordinary two-sample  $t$ -test where the shift in location of both Gaussian distributions admits the representation

$$\mu_- - \nu_- \equiv E(X_i^-) - E(Y_j^-) = 2(\phi_T - \phi_R). \quad (10.8)$$

Thus, treating the  $X_i^-$  and  $Y_j^-$  as the raw data, (10.4) is nothing but a special instance of the problem of testing for equivalence of two homoskedastic Gaussian distributions with respect to the absolute shift in location. As we know from § 3.1, the simplest solution of an equivalence testing problem of that kind is through an application of the principle of confidence interval inclusion.

For a concise description of the corresponding decision rule, it is convenient to introduce the following notation making explicit that all statistics involved have to be computed from the inter-period differences defined by (10.5):

$$\begin{aligned} \bar{X}^- &= (1/m) \sum_{i=1}^m X_i^-, \quad \bar{Y}^- = (1/n) \sum_{j=1}^n Y_j^-; \\ \tilde{s}^- &= \left[ \left( \left( \sum_{i=1}^m (X_i^- - \bar{X}^-)^2 + \sum_{j=1}^n (Y_j^- - \bar{Y}^-)^2 \right) / (m+n-2) \right) \left( \frac{1}{m} + \frac{1}{n} \right) \right]^{1/2}. \end{aligned}$$

In terms of these symbols, a pair of confidence bounds for  $\phi_T - \phi_R$  of the same one-sided level  $1 - \alpha$  reads

$$(1/2) \cdot (\bar{X}^- - \bar{Y}^- \mp \tilde{S}^- t_{N-2; 1-\alpha}) \quad (10.9)$$

with  $N = m + n$  and  $t_{N-2; 1-\alpha}$  denoting the  $(1 - \alpha)$ -quantile of a central  $t$ -distribution with  $N - 2$  degrees of freedom. Correspondingly, the level- $\alpha$  interval inclusion test of (10.4) is given by the decision rule:

Reject average bioequivalence if and only if it turns

$$\text{out that } |\bar{X}^- - \bar{Y}^-| < 2\delta_o - \tilde{S}^- t_{N-2; 1-\alpha}. \quad (10.10)$$

Except for relying on confidence bounds of one-sided level  $1 - \alpha$  rather than  $1 - \alpha/2$ , this solution to the problem of average BE assessment has been proposed as early as in the seminal paper of Westlake (1972). Although the underlying criterion for the equivalence of two normal distributions with the same variance can be criticized for missing a crucial point (as will be explained in § 10.2.2), the method is still strongly recommended in the guidelines as a standard procedure. One of the undisputed advantages of the decision rule (10.10) is computational simplicity. Unfortunately, this advantage is more or less canceled out by the fact that exact power computations are rather complicated requiring numerical integration even if the null alternative  $\phi_T = \phi_R$  is selected. The power of the interval inclusion test (10.10) for average BE against an arbitrary specific alternative  $|\phi_T - \phi_R| = \delta < \delta_o$  can easily be shown to admit the representation

$$POW_{\delta_o}(\delta, \sigma) = \int_0^{v(\delta_o/\sigma; m, n, \alpha)} \left[ \Phi\left(\sqrt{\frac{mn}{N}} 2(\delta_o - \delta)/\sigma - vt_\alpha^*\right) - \Phi\left(-\sqrt{\frac{mn}{N}} \cdot 2(\delta_o + \delta)/\sigma + vt_\alpha^*\right) \right] \sqrt{N-2} f_{N-2}^\chi(\sqrt{N-2} v) dv, \quad (10.11)$$

where

$$t_\alpha^* = t_{N-2; 1-\alpha}, \quad v(\delta_o/\sigma; m, n, \alpha) = \sqrt{mn/N} 2(\delta_o/\sigma)/t_\alpha^*, \quad (10.12)$$

and  $f_{N-2}^\chi(\cdot)$  stands for the density function of a so-called  $\chi$ -distribution with  $N - 2$  degrees of freedom. By the usual definition (cf. Johnson et al., 1994, p. 417), the latter is explicitly given as

$$f_{N-2}^\chi(u) = 2^{2-N/2} e^{-u^2/2} \cdot u^{N-3}/\Gamma(N/2 - 1), \quad u > 0. \quad (10.13)$$

Numerical evaluation of the integral (10.11) can be carried out by means of the same technique recommended earlier [see § 3.2] for exact computation

of marginal posterior probabilities for equivalence ranges. In the WKTSEQ2 Source Code Package a couple of files with common basename `pow_abe` can be found containing R and SAS scripts of a program which enables its user to determine the power of the interval inclusion test (10.10) with higher accuracy than required for practical applications. The program allows for arbitrary specifications of the nominal significance level  $\alpha$ , the sample sizes  $m, n$  in both sequence groups, and the parameters  $\delta \in \mathbb{R}_+ \cup \{0\}$ ,  $\sigma \in \mathbb{R}_+$ . Since  $\delta$  can in particular be chosen as any point in  $[\delta_0, \infty)$ , the procedure can also be used for studying the exact size of the test (10.10).

### *Example 10.1*

We illustrate the standard procedure of testing for average BE by applying it to the areas under the serum-concentration curves from time 0 to the last time with measurable concentration obtained in a bioequivalence study whose results have been available at the FDA web-site since December 1997, on a file named `gen10.txt`. Although the original trial aimed primarily at detecting a possible gender-by-treatment interaction, the data are analyzed here as in a conventional comparative bioavailability study. Table 10.1 displays the log-AUC's observed in each subject during both periods, together with the inter-period differences.

Computing the empirical means and variances with the entries in the bottom lines of the two parts of the above table gives:

$$\begin{aligned}\bar{X}^- &= -0.06058; & S_{X^-}^2 &= .0277288; \\ \bar{Y}^- &= -0.04515; & S_{Y^-}^2 &= .0336990.\end{aligned}$$

Using these values of the sample variances, the pooled estimate of the standard error of  $\bar{X}^- - \bar{Y}^-$  is obtained as

$$\begin{aligned}\tilde{S}^- &= [(11 \cdot .0277288 + 12 \cdot .0336990)/23]^{1/2} \cdot [1/12 + 1/13]^{1/2} \\ &= .070306.\end{aligned}$$

Since the 95th percentile of the central  $t$ -distribution with  $N - 2 = 23$  degrees of freedom is  $t_{23,.95} = 1.713872$ , the critical upper bound to  $|\bar{X}^- - \bar{Y}^-|$  to be used in the interval inclusion test for average BE at level  $\alpha = 5\%$  is computed to be:

$$2 \log(5/4) - .070306 \cdot 1.713872 = .102648.$$

With the observed values of the sample means, we have  $|\bar{X}^- - \bar{Y}^-| = .015430$ , so that applying rejection rule (10.10) leads to a decision in favor of average BE.

Taking the observed value of  $\tilde{S}^- / \sqrt{1/m + 1/n}$  for the true value of the population standard deviation  $\sigma$  of a single within-subject difference gives  $\sigma = .175624$ , and under this assumption, the exact power of the test against  $\delta = \delta_0/2$  and  $\delta = 0$  turns out (running program `pow_abe`) to be .92415 and

.99993, respectively. If the true value of  $\sigma$  is assumed to be twice as large, i.e., .351249, these power values drop to .45679 and .84831, respectively.

Table 10.1 *Logarithmically transformed AUC-values and inter-period differences computed from the FDA (1997) sample bioequivalence data set gen10.*

*Sequence Group T/R*

<i>i</i>	1	2	3	4	5	6	7
$X_{1i}$	4.639	4.093	4.222	4.549	4.241	4.279	4.309
$X_{2i}$	4.501	4.353	4.353	4.580	4.064	4.618	4.380
$X_i^-$	0.138	-0.260	-0.131	-0.031	0.177	-0.339	-0.071
	8	9	10	11	12		
	4.436	4.572	4.546	3.882	4.561		
	4.565	4.492	4.577	4.121	4.452		
	-0.129	0.080	-0.031	-0.239	0.109		

*Sequence Group R/T*

<i>j</i>	1	2	3	4	5	6	7
$Y_{1j}$	4.746	4.521	4.009	4.818	4.040	4.099	4.140
$Y_{2j}$	4.560	4.486	3.953	5.001	4.114	4.511	3.816
$Y_j^-$	0.186	0.035	0.056	-0.183	-0.074	-0.412	0.324
	8	9	10	11	12	13	
	4.369	4.445	4.714	4.018	4.145	4.449	
	4.363	4.598	4.831	4.199	4.129	4.539	
	0.006	-0.153	-0.117	-0.181	0.016	-0.090	

### Modifications and extensions

The interval inclusion test (10.10) has been considered by many authors from very different perspectives. Mandallaz and Mau (1981) showed that it admits a representation as a Bayesian testing procedure with respect to an noninformative reference prior for  $(\mu^-, \nu^-, \sigma)$ , and Schuirmann (1987) pointed out its equivalence with a “combination” of two one-sided *t*-tests with shifted common boundary of the hypotheses [see also § 3.1]. From a mathematical point of view, it is perhaps its most striking property (becoming immediately ob-

vious from (10.11), (10.12)) that, given any value of the (absolute) difference  $\delta = \phi_T - \phi_R$  between the true formulation effects, its rejection probability converges to zero as  $\sigma \rightarrow \infty$ , implying that there are specific alternatives against which the power is much smaller than the significance level  $\alpha$ .

Various attempts have been made to reduce or even totally eliminate this lack of unbiasedness of the standard procedure of testing for average BE. Anderson and Hauck (1983) proposed to use the rejection region

$$\left\{ F_{N-2}^T((|\bar{D}^-| - 2\delta_0)/\tilde{S}^-) - F_{N-2}^T(-|\bar{D}^-| - 2\delta_0)/\tilde{S}^- < \alpha \right\}, \quad (10.14)$$

where  $F_{N-2}^T(\cdot)$  denotes the cdf of a central  $t$ -distribution with  $N-2$  degrees of freedom and  $\bar{D}^- \equiv \bar{X}^- - \bar{Y}^-$ . The level and power properties of (10.14) were extensively studied by Frick (1987) and Müller-Cohrs (1988) using numerical methods. First of all, the results obtained from these investigations show that the modification suggested by Anderson and Hauck leads to an anticonservative solution of the problem of testing for average BE which is not surprising. In fact, the function  $F_{N-2}^T(\cdot)$  is positive everywhere on  $\mathbb{R}$  so that (10.14) is a proper superset of  $\{F_{N-2}^T((|\bar{D}^-| - 2\delta_0)/\tilde{S}^-) < \alpha\} = \{(|\bar{D}^-| - 2\delta_0)/\tilde{S}^- < (F_{N-2}^T)^{-1}(\alpha)\} = \{|\bar{D}^-| < 2\delta_0 + \tilde{S}^- t_{N-2; \alpha}\} = \{|\bar{D}^-| < 2\delta_0 - \tilde{S}^- t_{N-2; 1-\alpha}\}$  and thus of the critical region of the interval inclusion test which is well known to have exact size  $\alpha$ . Aligning both tests with respect to size, the Anderson-Hauck procedure has better power against alternatives it detects with probability 60% at most. However, it still fails to be unbiased, but the bias is not as extreme as that of the interval inclusion test: For fixed  $\delta \in [0, \delta_0]$ , the power of the Anderson-Hauck test against the alternative  $(\delta, \sigma)$  converges to the nominal level rather than zero, as  $\sigma \rightarrow \infty$ . Munk (1993) proposed a “mixture” of both tests defined by the rule that (10.10) has to be used for values of  $\tilde{S}^-$  falling short of some cut-off value  $k^*$  (depending on the target significance level  $\alpha$  and the number  $N-2$  of degrees of freedom), and (10.14) with corrected nominal level  $\alpha^* < \alpha$  for  $\tilde{S}^- > k^*$ .

Finally, the testing problem (10.4) admits also of constructing a critical region which is strictly unbiased. This has been shown by Brown et al. (1997) adapting Hodges and Lehmann’s (1954) approach to deriving an unbiased test for the dual problem to be written  $H' : |\phi_T - \phi_R| \leq \delta_0$  vs.  $K' : |\phi_T - \phi_R| > \delta_0$  in the present context. Unfortunately, the form of the unbiased critical region exhibits still more counterintuitive features than that of Anderson and Hauck’s test. In fact, as a function of the realized value  $\tilde{s}^-$  of  $\tilde{S}^-$ , the width of its horizontal sections is not only nonmonotonic, but even jagged. Like (10.14), the unbiased critical region contains a large subregion of points in the  $(\bar{d}^-, \tilde{s}^-)$ -plane whose distance from the vertical axis is larger than the theoretical equivalence limit to the target parameter  $2|\delta|$  (see Figure 10.2), notwithstanding the fact that  $\bar{D}^-$  is the natural estimator of  $2\delta$ . Since the gain in power which can be achieved when using such curiously shaped critical regions instead of the triangular region corresponding to (10.10) seems not really relevant from a practical point of view, it is not surprising that neither

of the modifications to the interval inclusion test mentioned above made its way in the routine of BE assessment.

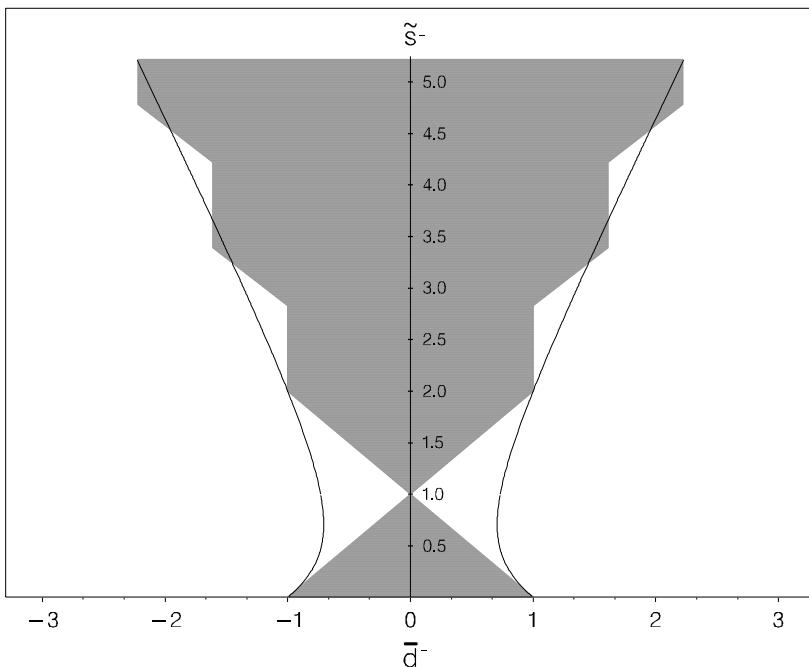


Figure 10.2 *Critical region of the interval inclusion test (lower triangular part of shaded region), the level-corrected Anderson-Hauck test (—) and the unbiased test (shaded region as a whole) for (10.4) with  $2\delta_0$  replaced by unity and  $N - 2 = 1$ . [Exact size of all three regions:  $\alpha = .25$ .]* (Redrawn from Brown, Hwang and Munk, 1997, with kind permission by the International Mathematical Institute.)

As pointed out by Hauschke et al. (1996), the problem of testing for average BE on the basis of the inter-period (log) differences  $X_i^-$  and  $Y_j^-$  differs fundamentally from an ordinary two-sample  $t$ -test setting with regard to the possibility of justifying the assumption of homoskedasticity. In fact, in the vast majority of real applications, it is reasonable to suppose that the within-subject variability, if different from period to period at all, depends only on treatment as such and not on treatment order, which implies by (10.7) that we have  $Var(X_i^-) = Var(Y_j^-)$  even for grossly different treatment-specific variances  $\sigma_{eT}^2$  and  $\sigma_{eR}^2$ . If one thinks it necessary to allow for heteroskedasticity of both distributions under comparison anyway, then it suggests itself to apply to (10.10) and (10.14) the same modifications which lead from the classical to

a heteroskedasticity-corrected two-sample  $t$ -test. Adopting Welch's (1938) approximate solution to the ordinary Behrens-Fisher problem, this means that we have to replace  $\tilde{S}^-$  and  $F_{N-2}^T(\cdot)$  with  $\tilde{S}_*^- \equiv [\frac{1}{m}S_{X^-}^2 + \frac{1}{n}S_{Y^-}^2]^{1/2}$  and  $F_{\hat{\nu}}^T(\cdot)$ , respectively, where  $S_{X^-}^2$ ,  $S_{Y^-}^2$  denote the two sample variances, and  $\hat{\nu}$  stands for the corrected (random) number of degrees of freedom to be determined from

$$\frac{1}{\hat{\nu}} = \left( \frac{S_{X^-}^2}{S_{X^-}^2 + (m/n)S_{Y^-}^2} \right)^2 \frac{1}{m-1} + \left( \frac{(m/n)S_{Y^-}^2}{S_{X^-}^2 + (m/n)S_{Y^-}^2} \right)^2 \cdot \frac{1}{n-1} .$$

The resulting modification of the critical region proposed by Anderson and Hauck has been studied by Dannenberg et al. (1994) with respect to size and power. The simulation results obtained by these authors show in particular that the Welch-corrected Anderson-Hauck test exhibits a similar (slight) tendency towards anticonservatism as the original version (10.14) in the homoskedastic case. Of course, the simplest means of reducing or even eliminating this tendency is to apply Welch's correction to the double  $t$ -test.

A problem still easier to handle than heteroskedasticity but much more important for real applications, is nonnormality of the distributions of the inter-period differences  $X_i^-, Y_j^-$ . Clearly, the classical parametric confidence limits  $(\bar{X}^- - \bar{Y}^-) \mp \tilde{S}^- t_{N-2; 1-\alpha}$  which the standard test (10.10) for average BE is based upon, can be replaced with any other pair of confidence bounds of one-sided level  $1 - \alpha$  each for the shift in location of two continuous distributions of identical shape. As is well known from the nonparametrics literature (see Lehmann, 1975, § 2.6), confidence bounds which are fully distribution-free in the two-sample shift model are given by suitable order statistics for the  $mn$ -dimensional vector of all differences  $X_i^- - Y_j^-$  between observations from different samples. As was already mentioned in § 6.2, this distribution-free variant of the interval inclusion rule (10.10) has been recommended for assessment of average BE in several contributions to pharmaceutical and clinical pharmacology journals (among others, see Hauschke et al., 1990; Hauck et al., 1997).

### 10.2.2 Testing for scaled average bioequivalence

For a correct interpretation of the meaning of a positive result of the standard test for average BE as well as any of its variants, it is indispensable to note the following basic fact: According to the alternative hypothesis to be established in the parametric model, it makes no difference whether some sufficiently small value, say .10, of the difference  $E(X_i^-) - E(Y_j^-)$  between the true means of the normal distributions under consideration, goes with a common (theoretical) standard deviation  $\sigma$  of .001 or 1000. Persumably, not many statisticians will disagree with the conclusion we came to in our general discussion [→ § 1.6] of sensible and less reasonable parametrizations of equivalence testing problems, that this conception of similarity between

normal distributions is hardly tenable. If one tries to trace back the literature on statistical methods of assessing what is nowadays called average BE, one finds that except for notational variations, hypotheses formulation (10.4) was taken as definitively given already in the earliest pertinent contributions. So there is strong indication that it was not the idea of statisticians to measure the distance between two homoskedastic Gaussian distributions on the basis of the absolute means only. In the early seventies when biostatisticians first started research on methods for analyzing BE studies (Westlake, 1972; Metzler, 1974), questioning the adequacy of that measure would have amounted to challenging a well-established tradition within the pharmaceutical sciences including clinical pharmacology.

Reformulating the problem of testing for average BE as

$$\tilde{H} : 2|\phi_T - \phi_R|/\sigma \geq \varepsilon \text{ versus } \tilde{K} : 2|\phi_T - \phi_R|/\sigma < \varepsilon \quad (10.15)$$

gives several important advantages over the traditional formulation (10.4). First of all, in view of (10.7) and (10.8), the alternative hypothesis now specifies that the ratio of the shift in location of the two Gaussian distributions under comparison over their common standard deviation be sufficiently small. Thus, the underlying measure of distance is scale-free and accommodates to the obvious fact that two normal distributions with common variance  $\sigma^2$  and given absolute difference .10 of their expected values are indistinguishable for all practical purposes if  $\sigma = 1000$ , whereas for  $\sigma = .001$ , the regions bounded by their density curves are almost perfectly disjoint. Furthermore, the arguments supporting a certain numerical specification of the tolerance  $\varepsilon$  do not depend on the special context of bioequivalence assessment, and for (10.15) there exists an exact unbiased test satisfying a strong optimality criterion and admitting power calculations in terms of a function which is predefined in advanced statistical packages and can thus be taken as giving explicit expressions. All tools required for carrying out the corresponding alternative test for (scaled) average BE as well as for planning a trial to be analyzed by means of it, have been provided in § 6.1. Thus, except for calculating the actual numerical values of the test statistic and the critical constant, nothing new comes into play if we base the analysis of a standard bioequivalence study like that considered in Example 10.1, on the two-sample  $t$ -test for equivalence.

### *Example 10.1 (continued)*

Suppose the entries in the bottom rows of Table 10.1 are the realized values of  $N = 12 + 13 = 25$  mutually independent random variables distributed as in (10.6), and we want to perform the UMPI test of § 6.1 at level  $\alpha = .05$  in order to prove statistically that the underlying normal distributions are equivalent in the sense of satisfying  $|\mu_- - \nu_-| = 2|\phi_T - \phi_R| < \varepsilon$  with  $\varepsilon = .74$  [ $\leftarrow$  Table 1.1, (v)]. Then, by (6.6), we have to compare the absolute value of the two-sample  $t$ -statistic, say  $T^-$ , computed from the  $X_i^-$  and  $Y_j^-$ , to a critical upper bound obtained as

$$\begin{aligned}
 \tilde{C}_{.05;12,13}(.74) &= \left[ F_{1,23; .05}((12 \cdot 13/25) \cdot .74^2) \right]^{1/2} \\
 &= \left[ \text{finv}(.05, 1, 23, 3.417024) \right]^{1/2} \\
 &= \sqrt{.111742} = .334278,
 \end{aligned}$$

using the SAS function for quantiles of noncentral  $F$ -distributions. On the other hand, we had  $\bar{X}^- = -0.06058$ ,  $\bar{Y}^- = -0.04515$  and  $\tilde{S}^- = 0.070306$ , so that the observed value of the  $t$ -statistic is  $T^- = -0.015430/0.070306 = -0.219469$ . Thus, the data shown in Table 10.1 lead to a positive decision in the test for scaled average bioequivalence as well [at the 5% level, and with equivalence range  $(-.74/2, .74/2)$  for  $\phi_T - \phi_R$ ].

### *Discussion*

- (i) Generally, there is no reason to expect that assessing scaled average BE will lead to the same qualitative decision as any of the procedures being available for testing of equivalence with respect to the absolute means. In particular, it is not unusual that with a given data set, the test for scaled average BE has to accept inequivalence although raw average BE can be established even by means of the simple interval inclusion rule (10.10). At first sight, one might suspect that this is a contradiction to the asserted optimality of the  $t$ -test for equivalence. Actually, this would be true only if for some  $\varepsilon > 0$ , the null hypothesis of (10.15) implied that of (10.4) which is obviously not the case. By the same reason, power comparisons between tests for average BE in the ordinary sense and scaled average BE are pointless in any case.
- (ii) Researchers in the pharmaceutical sciences often base their reservation about applying the scaled criterion of average BE on the argument that the exact value of the population variance  $\sigma^2$  of the inter-period (log-) differences is unknown to the experimenter. This would be a serious point only if transforming a given equivalence range for the absolute mean difference was the only way of coming to a consensus about what can be considered an adequate numerical specification of the equivalence limit to the standardized difference of means of two normal distributions with the same variance. Actually, as was explained in Chapter 1, nothing prevents from treating  $|\mu_- - \nu_-|/\sigma$  as the primary distance measure admitting direct specification of an equivalence limit without taking into consideration any other function of the three unknown parameters  $\mu_-$ ,  $\nu_-$  and  $\sigma^2$  involved in the problem. In addition, as is the case with any well-designed study to be eventually analyzed by means of hypotheses testing methods, planning of a BE trial also involves calculation of minimally required sample sizes, and it follows from equation (10.11) that this presupposes knowledge of  $\sigma^2$  even if the standard test for unscaled average

BE shall be used. Long-term experience shows that despite a high variability of the bioavailabilities measured in the individual periods of a BE trial, the standard deviation of the inter-period differences between the bioavailabilities observed in the same subject is typically smaller than the equivalence limit  $2\delta_0 = .4462$  to be specified for  $E(X_i^-) - E(Y_j^-)$  according to the guidelines of the regulatory authorities. Accordingly, a positive result of a test for unscaled average BE in the majority of cases ensures at best that the shift in location of the two normal distributions under comparison does not substantially exceed their common standard deviation. Thus, it seems not unfair to state that from a statistical point of view, the condition satisfied by two drug formulations proven bioequivalent with respect unscaled averages is often remarkably weak.

- (iii) Another objection which frequently has been raised against the use of  $|\mu^- - \nu^-|/\sigma = 2|\phi_T - \phi_R|/\sigma$  as the basic distance measure for assessment of average BE is that an experimenter who deliberately “inflates” the variability of the observed bioavailabilities by careless handling of the measurement procedure, will be “rewarded” by enabling him to meet the criterion of BE even for a comparatively large value of  $|\phi_T - \phi_R|$ . A good deal of skepticism seems in order about the soundness of this argument as well. First of all, it is not clear how an imagined experimenter who adopts the strategy in mind, can protect himself from increasing the shift in location of the distributions at the same time. Even if he succeeds in this latter respect, one cannot but admit that virtually all inferential procedures are liable to biases induced by manipulation of the conditions under which the primary observations are taken. To put it differently, methods of statistical inference are made for extracting the maximum information from data collected by people who are willing to adhere to the principles of good scientific practice, rather than for providing protection against fraud and doctoring of data.

## 10.3 Individual bioequivalence: Criteria and testing procedures

### 10.3.1 Introduction

The introduction of the concept of individual bioequivalence by Anderson and Hauck (1990) and Wellek (1990, 1993a) was the beginning of a process which, although overdue already at that time, is still going on and aims at emancipating bioequivalence assessment from taking into account only the first two moments of the distributions of the inter-period (log-) differences  $X_i^-$  and

$Y_j^-$ . The idea behind the concept in its original version is simple enough, and its logical basis is easily grasped even by people exclusively interested in the pharmacological and clinical issues connected with bioequivalence testing: Even perfect coincidence of  $\mu_- = E(X_i^-)$  and  $\nu_- = E(Y_j^-)$  by no means admits the conclusion that the proportion of individuals whose responses to both drug formulations exhibit the pharmacologically desirable degree of similarity, is sufficiently large. But precisely this has to be guaranteed in order to justify declaring the two formulations of the drug equivalent in the sense of allowing to switch from one to the other in the same individual without altering the response to a relevant extent (cf. Chen, 1997, p. 7).

The question of how this notion can be translated into a statistical hypothesis admits a straightforward answer if we restrict the model for the  $2 \times 2$  crossover design we have to consider when analyzing a standard BE trial, by assuming that the period effects also coincide and can thus be dropped from the expressions for the four cell means. Actually, this additional assumption is not as oversimplistic as it might seem at first sight. First of all, the proportion of real BE trials leading to a rejection of the null hypothesis  $\pi_1 = \pi_2$  in a suitable test of significance, is extremely low (confirming a general statement of Chow and Liu, 2008, p. 56). Moreover, there is no problem to positively establish approximate equality of both period effects by means of a preliminary  $t$ -test for equivalence of the distribution of the  $X_i^-$  to that of the  $-Y_j^-$  with respect to their standardized means. Thus, for practical purposes, it entails only a minor loss in generality to assume that through defining

$$Z_i = X_i^-, \quad i = 1, \dots, m; \quad Z_{i+j} = -Y_j^-, \quad j = 1, \dots, n \quad (10.16)$$

and

$$\zeta = \phi_T - \phi_R, \quad (10.17)$$

$Z_1, \dots, Z_N$  becomes a single homogeneous sample of size  $N = m + n$  with

$$Z_l \sim \mathcal{N}(\zeta, \sigma^2) \quad \forall l = 1, \dots, N, \quad \zeta \in \mathbb{R}, \quad \sigma^2 > 0. \quad (10.18)$$

In view of (10.16), (10.5) and the definition of the  $X_{ki}$ ,  $Y_{kj}$  as logarithms of the selected measure of bioavailability, nonexistence of period effects clearly ensures that  $|\exp\{Z_l\} - 1|$  measures the dissimilarity of the responses to both drug formulations observed in the same subject making up the  $l$ th element of the pooled sample. Now, according to basic standards of clinical pharmacology, we can expect that switching over from one formulation of the drug to the other will not change its potential therapeutic effect induced in the  $l$ th individual as long as the intraindividual bioavailability ratio  $\exp\{Z_l\}$  remains within the limits set by the 80–125 % rule. More generally, i.e., without restricting attention to a particular specification of the range considered acceptable for an individual bioavailability ratio, a natural formalization of the basic concept of individual BE is obtained by requiring that

$$P[(1 + \varepsilon)^{-1} < \exp\{Z_l\} < (1 + \varepsilon)] > \pi^* \quad (10.19)$$

where  $\varepsilon$  denotes a positive constant determining the tolerance for intraindividual differences in response levels, and  $\pi^*$  stands for the smallest value acceptable for the probability of observing bioequivalent responses in a randomly selected individual. The term “probability-based individual bioequivalence” (abbreviated to PBIBE in the sequel) has been used in the literature since the mid-nineties in order to distinguish (10.19) from an alternative condition proposed by Sheiner (1992) and Schall and Luus (1993) as a criterion of individual BE. This so-called moment-based criterion of individual BE is not discussed here because it presupposes that replicate crossover designs be used (for more details see FDA, 2001).

In the parametric model given by (10.18), the probability of observing bioequivalent responses at the individual level can obviously be written

$$P[(1 + \varepsilon)^{-1} < \exp\{Z_l\} < (1 + \varepsilon)] = \pi_\varepsilon(\zeta, \sigma), \quad (10.20)$$

provided we define

$$\pi_\varepsilon(\zeta, \sigma) = \Phi\left(\frac{\log(1 + \varepsilon) - \zeta}{\sigma}\right) - \Phi\left(\frac{-\log(1 + \varepsilon) - \zeta}{\sigma}\right). \quad (10.21)$$

In § 10.3.2, we will treat  $P[(1 + \varepsilon)^{-1} < \exp\{Z_l\} < (1 + \varepsilon)]$  as a functional allowed to depend on the cdf of the  $Z_l$  in an arbitrary way. The corresponding distribution-free test for PBIBE is based on a simple counting statistic and can thus be expected to leave a large margin for improvements in power when we assume that the expression on the left-hand side of (10.19) admits the representation (10.21). Although the problem of testing for a sufficiently large value of the parametric function  $\pi_\varepsilon(\zeta, \sigma)$  looks fairly simple, a satisfactory solution is difficult to find by means of classical methods. In § 10.3.3, a Bayesian construction is proposed instead which can be shown by numerical methods to maintain the ordinary frequentist significance level without being overly conservative and to yield gains in power of up to 30 % as compared to the distribution-free procedure.

### 10.3.2 Distribution-free approach to testing for probability-based individual bioequivalence

If establishing the validity of (10.19) is accepted as a reasonable objective of bioequivalence assessment, we need a procedure for testing

$$H_\varepsilon^{(1)} : p_\varepsilon \leq \pi^* \quad \text{versus} \quad K_\varepsilon^{(1)} : p_\varepsilon > \pi^*, \quad (10.22)$$

where  $p_\varepsilon$  symbolizes the probability of the event  $\{(1 + \varepsilon)^{-1} < \exp\{Z_l\} < (1 + \varepsilon)\} = \{|Z_l| < \log(1 + \varepsilon)\}$  under an arbitrary cdf  $F_Z(\cdot)$ , say, of the observed individual log-bioavailability ratios. Provided  $F_Z(\cdot)$  is continuous,

the (one-sided) testing problem (10.22) admits an easy solution exhibiting attractive theoretical properties. To see this, let us introduce the statistic

$$N_+ \equiv \#\left\{ l \in \{1, \dots, N\} \mid |Z_l| < \log(1 + \varepsilon) \right\} \quad (10.23)$$

which simply counts the number of subjects in the sample who exhibit a relative bioavailability lying within the tolerated range. Of course, the distribution of  $N_+$  is  $\mathcal{B}(N, p_\varepsilon)$  [binomial with parameters  $N, p_\varepsilon$ ]. Hence, it is natural to decide between the hypotheses of (10.22) by carrying out in terms of  $N_+$  the UMP level- $\alpha$  test of the null hypothesis  $p \leq \pi^*$  about the unknown parameter of a binomial distribution generated by repeating the same Bernoulli experiment independently  $N$  times. Clearly, this leads to the decision rule

$$\begin{cases} \text{Rejection of } H_\varepsilon^{(1)} & N_+ > k_N^*(\alpha) \\ \text{Rejection of } H_\varepsilon^{(1)} \text{ with probability } \gamma_N^*(\alpha) & \text{if } N_+ = k_N^*(\alpha), \\ \text{Acceptance of } H_\varepsilon^{(1)} & N_+ < k_N^*(\alpha) \end{cases} \quad (10.24)$$

where the critical constants have to be determined from

$$k_N^*(\alpha) = \min \left\{ k \in \mathbb{N}_0 \mid \sum_{j=k+1}^N b(j; N, \pi^*) \leq \alpha \right\}, \quad (10.25)$$

$$\gamma_N^*(\alpha) = \left( \alpha - \sum_{j=k_N^*(\alpha)+1}^N b(j; N, \pi^*) \right) / b(k_N^*(\alpha); N, \pi^*), \quad (10.26)$$

with  $b(\cdot; N, \pi^*)$  as the probability mass function of  $\mathcal{B}(N, \pi^*)$ . Obviously, (10.24) is simply a variant of the one-sided sign test which implies in particular that it is completely distribution-free. Since the hypotheses  $H_\varepsilon^{(1)}$  and  $K_\varepsilon^{(1)}$  refer to the class of *all* continuous distributions of the individual observations  $Z_l$ , mimicking Lehmann and Romano's (2005, pp. 291) proof of the UMP property of the ordinary sign test, (10.24) can be shown to be uniformly most powerful among all level- $\alpha$  tests of the nonparametric version of the null hypothesis of individual bio-inequivalence in the probability-based sense.

Under the acronym TIER (for Testing Individual Equivalence Ratios) coined by Anderson and Hauck (1990), the conservative, nonrandomized version of the procedure (10.24) gained some popularity for a time. Clearly, its most conspicuous practical advantage is extreme simplicity. In fact, neither for determining the critical constants nor the exact power against some specific alternative  $p_\varepsilon \in (\pi^*, 1)$ , other computational devices than an extensive table or a program for the binomial distribution function are required. Therefore, presentation of an extra example illustrating the practical use of the method is dispensable here.

With regard to its conceptual basis, the simple criterion (10.19) of PBIBE has rightly been criticized by several authors for potentially using a lower bound to the probability of observing equivalent bioavailabilities in a randomly selected subject which may be unattainable even in a study comparing the reference formulation to itself. A reasonable way around this potential difficulty recommended in the literature (see Schall, 1995, (2.3)) consists of replacing  $\pi^*$  by a bound scaled by relating it to a reliable estimate of the probability of finding the condition  $|Z_l| < \log(1 + \varepsilon)$  satisfied in a pseudo-trial of two identical drug products. In the parametric submodel (10.18), this probability is clearly given by  $\pi_\varepsilon(0, \sigma) = 2\Phi(\log(1+\varepsilon)/\sigma) - 1$  so that it seems natural to scale the bound  $\pi^*$  of (10.19) by replacing it with the number

$$\pi_{sc}^* \equiv \pi^* \cdot (2\Phi(\log(1+\varepsilon)/\sigma^*) - 1), \quad (10.27)$$

where  $\sigma^*$  denotes some realistic value of the population standard deviation  $\sigma$  of log-bioavailability ratios measured for the drug product under investigation. In contrast to the basic measurements taken in the individual period of a BE trial,  $\sigma^2$  does not depend on the between-subject variance component  $\sigma_S^2$  [recall (9.29)] but reflects the technical and analytical precision of the measurement process as such. Thus,  $\sigma^2$  is a substance-specific rather than biological constant, and a comparatively stable estimate might be obtained by pooling the point estimates over a variety of trials of the same or chemically related agents. Experience shows that in a BE trial performed in strict adherence to the established technical standards, the observed standard deviation of the  $Z_l$  often remains below .20. Using this as  $\sigma^*$  and .75 as the unscaled bound to  $P[|Z_l| < \log(1 + \varepsilon)]$  yields for  $\varepsilon = .25$ :

$$\pi_{sc}^* = .75 \cdot (2\Phi(.2231/.2000) - 1) = .75 \cdot .7353 \approx .55.$$

A more sophisticated approach to scaling the criterion of PBIBE is available in the Bayesian framework developed in the next subsection.

### 10.3.3 An improved parametric test for probability-based individual bioequivalence

In this subsection, we consider the parametric version of the problem of testing for PBIBE which arises from replacing in (10.22) the functional  $p_\varepsilon$  with  $\pi_\varepsilon(\zeta, \sigma)$  as defined in (10.21). In other words, we now put forward the problem of testing

$$\tilde{H}_\varepsilon^{(1)} : \pi_\varepsilon(\zeta, \sigma) \leq \pi^* \quad \text{versus} \quad \tilde{K}_\varepsilon^{(1)} : \pi_\varepsilon(\zeta, \sigma) > \pi^* \quad (10.28)$$

by means of a sample  $Z_1, \dots, Z_N$  of independent observations from  $\mathcal{N}(\zeta, \sigma^2)$ .

Our construction of a Bayesian test for (10.28) will be based on the usual (see Box and Tiao, 1973, p. 92) reference prior for the expectation  $\zeta$  and the standard deviation  $\sigma$  of the underlying Gaussian distribution. This means

that *a priori* we treat the true value of the population parameter  $(\zeta, \sigma)$  as the realization of a random variable  $(\zeta, \sigma)$ , say, inducing an improper probability distribution defined by the following density with respect to Lebesgue measure on  $\mathbb{R}^2$ :

$$\varrho(\zeta, \sigma) = \begin{cases} \sigma^{-1} & \text{for } -\infty < \zeta < \infty, \sigma > 0 \\ 0 & \text{for } -\infty < \zeta < \infty, \sigma \leq 0 \end{cases}. \quad (10.29)$$

The corresponding posterior density of  $(\zeta, \sigma)$  given the realized values  $(\bar{z}, s)$  of the sufficient statistics  $\bar{Z} = \sum_{i=1}^N Z_i/N$ ,  $S = [(N-1)^{-1} \sum_{i=1}^N (Z_i - \bar{Z})^2]^{1/2}$ , is well known (cf Box and Tiao, 1973, pp. 93-6) to be given as

$$\varrho_{(\bar{z}, s)}(\zeta, \sigma) = (\sqrt{N}/\sigma) \varphi\left(\sqrt{N}(\zeta - \bar{z})/\sigma\right) \sqrt{N-1} (s/\sigma^2) f_{N-1}^\chi\left(\sqrt{N-1}s/\sigma\right) \quad (10.30)$$

where  $\varphi(\cdot)$  denotes the standard normal density and  $f_{N-1}^\chi(\cdot)$  the density of a  $\chi$ -distribution with  $N - 1$  degrees of freedom [ $\rightarrow$  (10.13) with  $N$  augmented by 1].

Now, an algorithm for computing the posterior probability of the (alternative) hypothesis of (10.28) is obtained by exploiting the following basic facts:

- (i) For arbitrarily fixed  $\sigma > 0$ , the set of values of  $\zeta$  satisfying  $\tilde{K}_\varepsilon^{(1)}$  admits the representation

$$\{\zeta \in \mathbb{R} \mid \pi_\varepsilon(\zeta, \sigma) > \pi^*\} = \left( -\sigma q(\tilde{\varepsilon}/\sigma, \pi^*), \sigma q(\tilde{\varepsilon}/\sigma, \pi^*) \right), \quad (10.31)$$

where  $\tilde{\varepsilon} = \log(1 + \varepsilon)$  and  $q(\tilde{\varepsilon}/\sigma, \pi^*)$  has to be determined by solving the equation

$$\Phi(\tilde{\varepsilon}/\sigma - \psi) - \Phi(-\tilde{\varepsilon}/\sigma - \psi) = \pi^*, \quad \psi > 0. \quad (10.32)$$

- (ii) The solution of (10.32) exists and is uniquely determined for all  $\sigma$  falling below an upper bound  $\sigma_\varepsilon^*$ , say, defined by

$$\sigma_\varepsilon^* = \tilde{\varepsilon}/\Phi^{-1}((1 + \pi^*)/2). \quad (10.33)$$

[In view of Theorem A.1.5 (iii), this follows immediately from the fact that the left-hand side of (10.32) coincides with the power function of a test with critical region  $\{-\tilde{\varepsilon}/\sigma < U < \tilde{\varepsilon}/\sigma\}$ , where  $U \sim \mathcal{N}(\psi, 1)$ .]

- (iii) Combining (i) and (ii) with (10.30) shows that the posterior probability of the alternative hypothesis  $\{\pi_\varepsilon(\zeta, \sigma) > \pi^*\}$  of (10.28) can be written as

$$\begin{aligned}
P_{(\bar{z}, s)} \left[ \pi_\varepsilon(\boldsymbol{\zeta}, \boldsymbol{\sigma}) > \pi^* \right] &= \int_0^{\sigma^*} \left( \left[ \Phi \left( \sqrt{N} \left( q(\tilde{\varepsilon}/\sigma, \pi^*) - \bar{z}/\sigma \right) \right) - \Phi \left( \sqrt{N} \left( -q(\tilde{\varepsilon}/\sigma, \pi^*) - \bar{z}/\sigma \right) \right) \right] \right. \\
&\quad \left. \sqrt{N-1} (s/\sigma^2) f_{N-1}^\chi \left( \sqrt{N-1} s/\sigma \right) \right) d\sigma. \quad (10.34)
\end{aligned}$$

The integral on the right-hand side of (10.34) can be evaluated with very high numerical accuracy by means of Gauss-Legendre quadrature. Since computation of individual values of the function  $\sigma \mapsto q(\tilde{\varepsilon}/\sigma, \pi^*)$  on the basis of (10.32) is very fast and the cumulative standard normal distribution function  $\Phi(\cdot)$  is predefined in any programming environment to be taken into consideration for the present purpose, the degree of the polynomial determining the abscissas can once more be chosen as large as 96.

Implementations in SAS and R of the algorithm described above for the computation of the posterior probability of  $\tilde{K}_\varepsilon^{(1)}$  given an arbitrary realization  $(\bar{z}, s)$  of the sufficient statistic  $(\bar{Z}, S)$ , can be found in the **WKTSEQ2 Source Code Package** under the program name **po\_pbbe**. Using this computational tool, it is no problem to perform the Bayesian test with respect to the reference prior (10.29). For definiteness, its rejection region reads

$$\left\{ (\bar{z}, s) \in \mathbb{R} \times \mathbb{R}_+ \mid P_{(\bar{z}, s)} \left[ \pi_\varepsilon(\boldsymbol{\zeta}, \boldsymbol{\sigma}) > \pi^* \right] > 1 - \alpha \right\}. \quad (10.35)$$

For the specific case  $\alpha = .05$ ,  $\varepsilon = .25$ ,  $\pi^* = .75$ , and sample size  $N = 20$ , the corresponding set in the sample space of  $(\bar{Z}, S)$  is displayed as the filled portion of the graph shown in Figure 10.3. Roughly speaking, it looks like a vertically compressed image of the region bounded by the upper solid curve which is the geometric representation of the alternative hypothesis under consideration. Since  $(\bar{z}, s)$  is the observed value of the natural estimator for  $(\boldsymbol{\zeta}, \boldsymbol{\sigma})$ , it is clear that no test which is valid with respect to the significance level, can have a rejection region being larger than the subset of the parameter space specified by the alternative hypothesis one wishes to establish. Hence, the form of the Bayesian rejection region seems very reasonable.

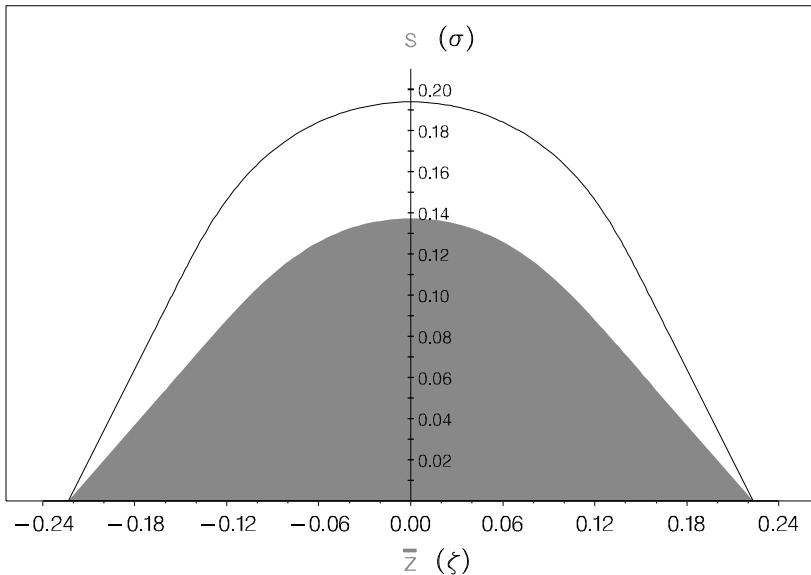


Figure 10.3 *Rejection region of the Bayesian test at (nominal) level  $\alpha = .05$  for PBIBE, for  $\varepsilon = .25$ ,  $\pi^* = .75$  and  $N = 20$ . [Upper solid line: Boundary of the parameter subspace specified by the alternative hypothesis.]* (From Wellek, 2000a, with kind permission by Wiley-VCH.)

#### *Numerical study of size and power of the Bayesian critical region*

The following result (rigorously proven in Wellek, 2000a) forms the basis of an algorithm for computing exact rejection probabilities of the Bayesian test with rejection region (10.35).

- (a) For any fixed point  $s_o$  in the sample space of the statistic  $S$ , there exists a nonnegative real number denoted  $\bar{z}_{\alpha\varepsilon}^*(s_o)$  in the sequel, such that

$$\left\{ \bar{z} \in \mathbb{R} \mid P_{(\bar{z}, s_o)} [\pi_\varepsilon(\zeta, \sigma) > \pi^*] > 1 - \alpha \right\} = \left( -\bar{z}_{\alpha, \varepsilon}^*(s_o), \bar{z}_{\alpha, \varepsilon}^*(s_o) \right). \quad (10.36)$$

Thus, all horizontal sections of the rejection region (10.35) are symmetric intervals on the real line.

- (b) The number  $\bar{z}_{\alpha\varepsilon}^*(s_0)$  is either zero [so that the corresponding section of (10.35) is empty], or it is uniquely determined by the equation

$$P_{(\bar{z}, s_0)} \left[ \pi_\varepsilon(\zeta, \sigma) > \pi^* \right] = 1 - \alpha, \quad \bar{z} \in \mathbb{R}_+. \quad (10.37)$$

- (c) There exists a positive real number  $s_{\alpha,\varepsilon}^*$  [depending, like  $\bar{z}_{\alpha\varepsilon}^*(s_0)$ , not only on  $\alpha$  and  $\varepsilon$ , but also on  $N$  and  $\pi^*$ ] such that we have

$$\{s \in \mathbb{R}_+ | \bar{z}_{\alpha,\varepsilon}^*(s) > 0\} = [0, s_{\alpha,\varepsilon}^*). \quad (10.38)$$

Let now  $\beta_{\alpha,\varepsilon}^*(\zeta, \sigma)$  denote the rejection probability of the Bayesian test under any parameter constellation  $(\zeta, \sigma) \in \mathbb{R} \times \mathbb{R}_+$ . In view of the independence of the two components of the sufficient statistic  $(\bar{Z}, S)$ , applying (a)–(c) yields the following integral representation of  $\beta_{\alpha,\varepsilon}^*(\zeta, \sigma)$ :

$$\begin{aligned} \beta_{\alpha,\varepsilon}^*(\zeta, \sigma) = & \int_0^{s_{\alpha,\varepsilon}^*} \left[ \Phi\left(\sqrt{N}(\bar{z}_\alpha^*(s) - \zeta)/\sigma\right) - \right. \\ & \left. \Phi\left(\sqrt{N}(-\bar{z}_\alpha^*(s) - \zeta)/\sigma\right) \right] (\sqrt{N-1}/\sigma) f_{N-1}^\chi(s\sqrt{N-1}/\sigma) ds \end{aligned} \quad (10.39)$$

where  $f_{N-1}^\chi(\cdot)$  again stands for the density function of the square root of a variable following a central  $\chi^2$ -distribution with  $N - 1$  degrees of freedom. The formal structure of the integrand appearing on the right-hand side of this equation is similar to that of the function to be integrated in order to compute the posterior probability of the alternative hypothesis of (10.28) by means of (10.34). Hence, in evaluating (10.39), 96 point Gauss-Legendre integration works as well as it does in computing the posterior probability of  $\tilde{K}_\varepsilon^{(1)}$  given an arbitrary realization of  $(\bar{Z}, S)$ .

Table 10.2 shows the results of applying formula (10.39) to various parameter constellations  $(\zeta, \sigma)$  lying on the common boundary of the hypotheses (10.28) for  $\varepsilon = .25, \pi^* = .75$ , the nominal significance level  $\alpha = .05$ , and three different sample sizes whose order of magnitude ranges from “small” ( $N = 20$ ) through “moderate” ( $N = 50$ ) to “large” ( $N = 100$ ). For true standard deviations  $\sigma$  being smaller than or equal to .10, the differences between  $\beta_{\alpha,\varepsilon}^*(\zeta, \sigma)$  and the nominal  $\alpha$  are clearly negligible, and even for the largest values of  $\sigma$  to be taken into consideration for purposes of assessing the level properties of the Bayesian test, the differences nowhere go in the wrong, e.g., anticonservative direction.

Table 10.2 *Exact rejection probabilities of the Bayesian test for PBIBE on the common boundary of the hypotheses [ $\alpha = .05$ ,  $\pi^* = .75$ ,  $\varepsilon = .25$  ( $\leftrightarrow \tilde{\varepsilon} = .2231$ )]. (From Wellek, 2000a, with kind permission by Wiley-VCH.)*

$\zeta$	$\sigma$	$N$	$\beta_{\alpha,\varepsilon}^*(\zeta, \sigma)$
.2164	.01	20	.050 000
.2164	.01	50	.050 009
.2164	.01	100	.049 758
.1894	.05	20	.050 000
.1894	.05	50	.050 000
.1894	.05	100	.050 000
.1557	.10	20	.048 438
.1557	.10	50	.049 401
.1557	.10	100	.049 659
.1163	.15	20	.035 674
.1163	.15	50	.040 441
.1163	.15	100	.043 053
.0709	.18	20	.028 230
.0709	.18	50	.034 270
.0709	.18	100	.038 209

Of course, (10.39) allows also numerically exact computation of the power of the Bayesian test against arbitrary specific alternatives  $(\zeta, \sigma) \in \tilde{K}_\varepsilon^{(1)}$ . Since computing the rejection probabilities of the distribution-free procedure of testing for PBIBE under arbitrary values of  $p_\varepsilon \in (0, 1)$  is an elementary matter, we are able to systematically assess the gain in efficiency achievable from using the Bayesian test based on (10.35) instead of the so-called TIER procedure (10.24). Although the randomized version of the latter is of theoretical interest only, Table 10.3 compares the rejection probabilities of both the nonrandomized and the optimal version with that of the Bayesian test for all specific alternatives selected. Otherwise, we would not be in a position to discriminate true gains in power achieved by using the parametric test, from “spurious” improvements vanishing if the conventional 5% level is replaced with a level “natural” for the counting statistic  $N_\varepsilon = \# \left\{ l \in \{1, \dots, N\} \mid |Z_l| < \log(1 + \varepsilon) \right\}$ .

Table 10.3 *Exact rejection probabilities of the Bayesian test and both versions of TIER against various specific alternatives, for the case  $\alpha = .05$ ,  $\pi^* = .75$ , and  $\varepsilon = .25$ .* (From Wellek, 2000a, with kind permission by Wiley-VCH.)

$\zeta$	$\sigma$	$\pi_\varepsilon(\zeta, \sigma)$	$N$	Bayes	TIER nonrandom	TIER random
.16423	.07	.80	20	.14423	.06918	.12171
.16423	.07	.80	50	.25380	.19041	.20238
.16423	.07	.80	100	.41369	.27119	.31550
.12127	.12	.80	20	.12532	.06918	.12171
.12127	.12	.80	50	.23833	.19041	.20238
.12127	.12	.80	100	.40096	.27119	.31550
.03745	.17	.80	20	.08508	.06918	.12171
.03745	.17	.80	50	.19513	.19041	.20238
.03745	.17	.80	100	.36831	.27119	.31550
.17132	.05	.85	20	.34955	.17556	.26355
.17132	.05	.85	50	.68261	.51875	.53404
.17132	.05	.85	100	.92453	.76328	.79930
.11937	.10	.85	20	.33178	.17556	.26355
.11937	.10	.85	50	.67423	.51875	.53404
.11937	.10	.85	100	.92231	.76328	.79930
.03891	.15	.85	20	.23753	.17556	.26355
.03891	.15	.85	50	.61062	.51875	.53404
.03891	.15	.85	100	.91057	.76328	.79930
.15907	.05	.90	20	.67436	.39175	.50117
.15907	.05	.90	50	.97097	.87785	.88444
.15907	.05	.90	100	.99977	.98999	.99264
.09456	.10	.90	20	.63842	.39175	.50117
.09456	.10	.90	50	.96741	.87785	.88444
.09456	.10	.90	100	.99974	.98999	.99264
.01338	.135	.90	20	.52538	.39175	.50117
.01338	.135	.90	50	.95347	.87785	.88444
.01338	.135	.90	100	.99970	.98999	.99264

Overall the results displayed in Table 10.3 admit the following conclusions:

- The Bayesian test dominates the nonrandomized TIER uniformly on the selected grid of points in  $\tilde{K}_\varepsilon^{(1)}$ . In many cases, the difference between the two power functions has an order of magnitude making further comments on the practical relevance of the improvement dispensable.
- Even as compared to the randomized version of TIER, the Bayesian test comes out as clearly superior because for the few constellations with a

negative sign of the difference, the distance between the power functions is practically negligible.

### *Generalizations*

*1. Scaling of the criterion of PBIBE.* The Bayesian framework admits the possibility of letting the factor to be used for scaling the bound  $\pi^*$  to the probability of obtaining a bioequivalent individual response, depend on the actual value of the standard deviation  $\sigma$  of the log-bioavailability ratios  $Z_l$ . As explained on p. 329, the natural candidate for a function of  $\sigma$  whose values give suitable weight factors to  $\pi^*$ , is  $\sigma \rightarrow \pi_\varepsilon(0, \sigma)$ , so that we propose to modify the condition  $\pi_\varepsilon(\zeta, \sigma) > \pi^*$  for PBIBE by requiring that  $\pi_\varepsilon(\zeta, \sigma) > \pi^* \cdot \pi_\varepsilon(0, \sigma)$ . This leads to replacing (10.28) with

$$\tilde{H}_\varepsilon^{(1)} : \frac{\pi_\varepsilon(\zeta, \sigma)}{\pi_\varepsilon(0, \sigma)} \leq \pi^* \quad \text{versus} \quad \tilde{K}_\varepsilon^{(1)} : \frac{\pi_\varepsilon(\zeta, \sigma)}{\pi_\varepsilon(0, \sigma)} > \pi^*. \quad (10.40)$$

As shown in the graphical representation of Figure 10.4, incorporation of that kind of scaling induces a radical change to the shape of the equivalence region specified by the alternative hypothesis to be established by means of the data. In contrast to  $\tilde{K}_\varepsilon^{(1)}$  of (10.28) [→ outer contour plotted in Fig. 10.3],  $\tilde{K}_\varepsilon^{(1)}$  corresponds to a region in the  $(\zeta, \sigma)$ -plane being unbounded in upward direction such that the width of the horizontal sections increases to infinity as  $\sigma \rightarrow \infty$ . This implies in particular, that for whatever large a shift in the mean away from zero, scaled PBIBE in the sense of (10.40) is satisfied for any sufficiently extreme degree of variability. If the latter property is considered undesirable, it might make sense to combine the Bayesian test for  $\tilde{K}_\varepsilon^{(1)}$  whose rejection region is shown in Figure 10.4 as well [again for  $\alpha = .05$ ,  $\varepsilon = .25$ ,  $N = 20$ , and a lower bound to the target parametric function of .75], with a test for reasonably low variability, as described in the following subsection in connection with a disaggregate test for population bioequivalence.

*2. Accommodating nonnegligible period effects.* If the period effects cannot be neglected (which is, as mentioned earlier, quite exceptional in bioequivalence trials), the difference (10.17) between the direct formulation effects can be written  $\zeta = \zeta_1 - \zeta_2$  where  $\zeta_1$  and  $\zeta_2$  denote the expectations of two normal distributions with common (unknown) variance  $\sigma_*^2$  from which independent samples  $Z_1^{[1]}, \dots, Z_m^{[1]}$  and  $Z_1^{[2]}, \dots, Z_n^{[2]}$  have been taken. Clearly, we have just to put  $Z_i^{[1]} = X_i^-/2$ ,  $Z_j^{[2]} = Y_j^-/2$ , and observe that the standard deviation  $\sigma$  we are interested in according to (10.21) is now given by  $\sigma = 2\sigma_*$ .

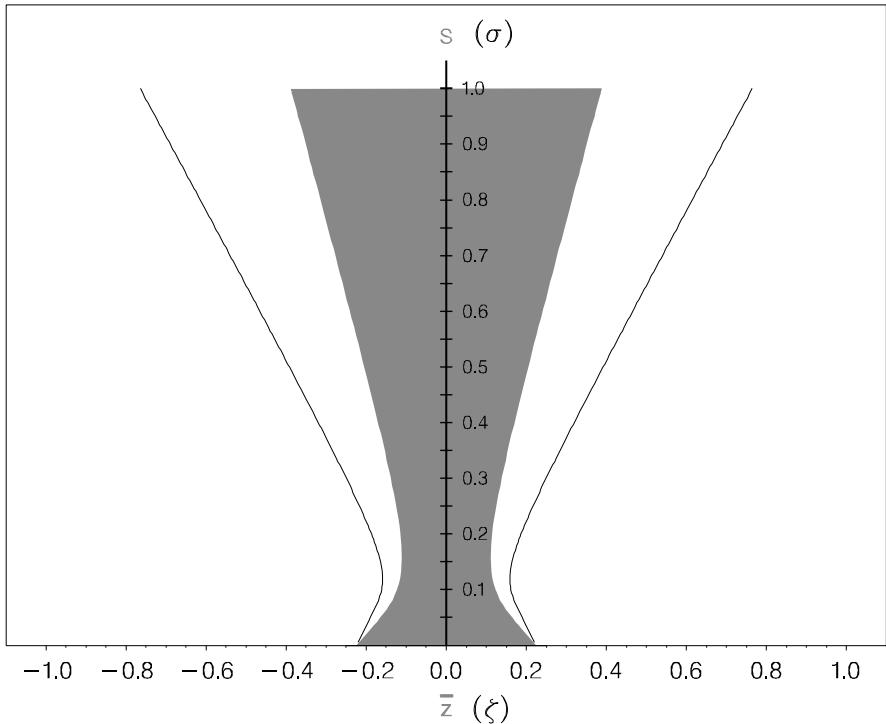


Figure 10.4 Rejection region of the Bayesian test at (nominal) level  $\alpha = .05$  for scaled PBIBE in the sense of (10.40), for  $\varepsilon = .25$ ,  $\pi^* = .75$  and  $N = 20$ . [Outer solid line: Boundary of the parameter subspace specified by the alternative hypothesis. (From Wellek, 2000a, with kind permission by Wiley-VCH.)]

Now, the posterior distribution of  $(\zeta, \sigma_*)$  given the observed values of  $\bar{Z}^{[1]} - \bar{Z}^{[2]}$  and  $S_* = \left[ (m + n - 2)^{-1} (SQ_Z^{[1]} + SQ_Z^{[2]}) \right]^{1/2}$  [with  $SQ_Z^{[1]}$  and  $SQ_Z^{[2]}$  as the corrected sum of squares for the  $Z_i^{[1]}$  and the  $Z_j^{[2]}$ , respectively] has the same form as in the paired  $t$ -test setting considered in the body of this paper. In fact, as shown in § 2.5.1 of Box and Tiao (1973) the posterior density of  $(\zeta, \sigma_*)$  with respect to the reference prior with (improper) density  $(\zeta_1, \zeta_2, \sigma_*) \mapsto \sigma_*^{-1}$  on  $\mathbb{R}^2 \times \mathbb{R}_+$ , is given by:

$$\varrho_{(\hat{\zeta}, s_*)}(\zeta, \sigma_*) = \sqrt{mn/N} (1/\sigma_*) \varphi\left(\sqrt{mn/N}(\zeta - \hat{\zeta})/\sigma_*\right) \cdot \sqrt{N-2} (s_*/\sigma_*^2) f_{N-2}^\chi\left(\sqrt{N-2} s_*/\sigma_*\right) \quad (10.41)$$

where  $f_{N-2}^\chi(\cdot)$  has to be defined as in (10.13). Hence, except for minor mod-

ifications, the algorithms described above can also be used for computing posterior probabilities of the alternative hypotheses of (10.28) and (10.40), as well as exact rejection probabilities of the associated Bayesian tests for (scaled) PBIBE with data from bioequivalence trials involving nonnegligible period effects.

### *Example 10.2*

For illustration, we use the data set analyzed by means of a nonparametric procedure for the assessment of scaled average BE in Wellek (1996). It has been obtained from a standard comparative bioavailability study of a generic ( $T$ ) and the original manufacturer's formulation ( $R$ ) of the calcium blocking agent nifedipine. As is so often the case in practice of bioequivalence assessment, the total sample size has been rather small ( $N = 20$ ). The complete listing of the raw data is given in Table 10.4.

Although the primary analysis performed in the reference cited above used an approach allowing for unequal period effects  $\pi_1 \neq \pi_2$ , the natural estimator  $(1/2)(\bar{X}^- + \bar{Y}^-)$  of  $\pi_1 - \pi_2$  (see, e.g., Jones and Kenward, 2003, § 2.3) turns out to be rather small [-.0028] as compared to its standard error [.0357]. Hence there seems little harm in assuming that switching from the first period to the second per se does not change the average bioavailability of the active agent under study. This leads to treating the values of  $\log(AUC_1/AUC_2)$  from the subjects in sequence group  $T/R$  and those of  $\log(AUC_2/AUC_1)$  from the subjects randomized to  $R/T$ , as a single sample of size  $N = 20$  from the same Gaussian distribution  $\mathcal{N}(\zeta, \sigma^2)$ . The values of the sufficient statistics are readily computed to be  $\bar{Z} = -.05168$  and  $S = .15559$ , respectively.

Setting the nominal significance level  $\alpha$  equal to the conventional value 5%, the tolerance  $\varepsilon$  determining the equivalence range for an intra-subject bioavailability ratio to .25, and the lower bound for ordinary and scaled PBIBE to .75, we are able to apply both versions of the Bayesian decision rule simply by means of the graphical devices provided by Figure 10.3 and 10.4. On the one hand, the observed point  $(-.05168, .15559)$  fails to be included in the rejection region displayed in Figure 10.3 whose vertex is located below  $s = .14$ . Accordingly, in the Bayesian test for nonscaled PBIBE as specified by the alternative hypothesis  $\tilde{K}_\varepsilon^{(1)}$  of (10.28), the data

Table 10.4 Results of a bioequivalence trial of two formulations of the calcium-blocking agent nifedipine. (From Wellek, 1996, with kind permission by Wiley-VCH.)

Subject ID ( $l$ )	$AUC_1$	$AUC_2$	SEQ	$Z_l$
1	106.9	112.9	$T/R$	-0.05461
2	131.3	124.4	$T/R$	0.05398
3	81.4	89.5	$T/R$	-0.09486
4	154.7	134.9	$T/R$	0.13695
5	111.2	108.3	$T/R$	0.02643
6	85.8	94.0	$T/R$	-0.09128
7	295.2	418.6	$T/R$	-0.34926
8	217.0	207.0	$T/R$	0.04718
9	252.3	239.3	$T/R$	0.05290
10	157.9	207.3	$T/R$	-0.27221
11	217.3	195.2	$R/T$	-0.10725
12	174.4	122.7	$R/T$	-0.35161
13	155.8	188.2	$R/T$	0.18893
14	299.5	309.2	$R/T$	0.03187
15	157.6	153.5	$R/T$	-0.02636
16	121.4	104.7	$R/T$	-0.14799
17	143.9	119.3	$R/T$	-0.18748
18	157.0	146.8	$R/T$	-0.06717
19	114.5	138.2	$R/T$	0.18813
20	71.0	70.3	$R/T$	-0.00991

shown in Table 10.4 do not allow a positive decision. On the other hand,  $(-.05168, .15559)$  is obviously an inner point of the left half of the grey filled region shown in Figure 10.4 so that the Bayesian test for scaled PBIBE decides in favor of the alternative  $\tilde{K}_\varepsilon^{(1)}$  [ $\rightarrow (10.40)$ ] with the same data. Interestingly, such a qualitative discrepancy between the final decisions to be taken also occurs if, in assessing the above data set, the conventional test (10.10) for average bioequivalence is contrasted with the  $t$ -test for scaled average bioequivalence (Wellek, 1991): At the 5% level, average bioequivalence in the sense of the well-known 80–125% criterion can be established, but scaled average bioequivalence cannot, even if a shift by up to a whole common standard deviation is defined compatible with equivalence of two homoskedastic Gaussian distributions.

## 10.4 Approaches to defining and establishing population bioequivalence

### 10.4.1 Introduction

With regard to coherence of terminology, the concept of population bioequivalence which has attained considerable popularity in the pertinent literature during the last 10 years, is hardly a fortunate choice. As a matter of fact, every inferential method of BE assessment aims at establishing a certain statement about the populations from which the subjects participating in the actual trial have been recruited by means of random selection. So the term population bioequivalence, which was originally coined by Anderson and Hauck (1990) and technically elaborated by Schall and Luus (1993), reveals almost nothing about the specific content of the concept it denotes. Nevertheless, after the preparations made in §10.1, its technical content can readily be made precise. The idea is to require of bioequivalent drug formulations that the marginal distributions associated with the treatments in the two sequence groups be sufficiently similar both with respect to the treatment-related component of the mean, and total variability. Under the standard model for the  $2 \times 2$  crossover design, the variances of the  $X_{ki}$  and the  $Y_{kj}$  depend only on treatment and are given by

$$\text{Var}(X_{1i}) = \text{Var}(Y_{2j}) = \sigma_S^2 + \sigma_{eT}^2 \equiv \sigma_{TT}^2, \quad (10.42\text{a})$$

$$\text{Var}(X_{2i}) = \text{Var}(Y_{1j}) = \sigma_S^2 + \sigma_{eR}^2 \equiv \sigma_{TR}^2. \quad (10.42\text{b})$$

In the existing literature (see, e.g., Schall and Luus, 1993; Holder and Hsuan, 1993), as well as in the guidelines of the regulatory agencies (FDA, 2001), so-called aggregate criteria of population BE have been preferred up to now. Instead of aiming to establish sufficient similarity both of  $\phi_T$  to  $\phi_R$  and of  $\sigma_{TT}^2$  to  $\sigma_{TR}^2$ , such an aggregate approach relies on a single real-valued function of all four parameters to be compared to some prespecified upper bound. As a so-called reference-scaled version of such an aggregate criterion, the most recent FDA guidance for BE trials recommends

$$\theta_P = \frac{(\phi_T - \phi_R)^2 + (\sigma_{TT}^2 - \sigma_{TR}^2)}{\sigma_{TR}^2}. \quad (10.43)$$

There are several reasons to question whether it really makes sense to aggregate the moments under comparison exactly in the way leading to (10.43).

- (i) Except for the scale-factor  $1/\sigma_{TR}^2$ , the parametric function to be assessed according to (10.43) combines a Euclidean distance (viz., that between

the formulation effects) with an ordinary signed difference. The commonly referenced argument (see Schall and Luus, loc. cit.) for the use of this combination is based on the equation  $E[(Y_T - Y_R)^2 - (Y_R - Y_{R'})^2] = (\phi_T - \phi_R)^2 + \sigma_{TT}^2 - \sigma_{TR}^2$ . However, the latter presupposes that all three observed (log)-bioavailabilities  $Y_T$ ,  $Y_R$  and  $Y_{R'}$  are taken from different subjects, whereas in a real bioequivalence trial the drug formulations are administered consecutively to each individual.

- (ii) In principle, scaling the primary distance measure has much to recommend it, provided there exists an intrinsic relationship between the parametric function to be scaled and the parameter with respect to which standardization is done. However, the incommensurability of the two terms making up the numerator of (10.43) raises a question as to whether a single parameter can give an appropriate reference for both and hence for their sum. In fact, under the basic parametric model for two-period crossover BE studies [→ (9.28), (9.29), (10.1), (10.2)], the optimal estimator for  $\phi_T - \phi_R$  has a sampling variance which is proportional to the sum of the two intra-subject variances  $\sigma_{eT}^2$  and  $\sigma_{eR}^2$ . Hence, a suitably scaled measure for the distance between the formulations with respect to average log-bioavailability is given by

$$\eta^2 = \frac{(\phi_T - \phi_R)^2}{\sigma^2} \quad (10.44)$$

where

$$\sigma^2 = \sigma_{eT}^2 + \sigma_{eR}^2 . \quad (10.45)$$

On the other hand, the adequacy of  $1/\sigma_{TR}^2$  as a scaling factor for the second term in the numerator of the parametric function introduced in (10.43) is obvious. Hence, it is natural to require of population-bioequivalent drug formulations that both  $\eta^2$  and

$$\Lambda = \sigma_{TT}^2 / \sigma_{TR}^2 - 1 \quad (10.46)$$

should be sufficiently small.

In order to circumvent the problems pointed out above, Wellek (2000b) proposed a disaggregate scaled criterion of population BE which requires that the two inequalities

$$(\phi_T - \phi_R)^2 / (\sigma_{WT}^2 + \sigma_{WR}^2) < \varepsilon_\phi^* \quad (10.47a)$$

and

$$\sigma_{TT}^2 / \sigma_{TR}^2 < 1 + \varepsilon_\sigma^* \quad (10.47b)$$

must be satisfied *simultaneously*, for suitably chosen constants  $\varepsilon_\phi^*, \varepsilon_\sigma^* > 0$ . A formal inferential procedure for assessing population BE in this latter sense consists of performing a valid test of

$$H_U : H_\phi \vee H_\sigma \quad \text{versus} \quad K_\cap : K_\phi \wedge K_\sigma \quad (10.48)$$

where the elementary testing problems involved read in explicit formulation

$$\begin{aligned} H_\phi : (\phi_T - \phi_R)^2 / (\sigma_{WT}^2 + \sigma_{WR}^2) &\geq \varepsilon_\phi^* \quad \text{vs.} \\ K_\phi : (\phi_T - \phi_R)^2 / (\sigma_{WT}^2 + \sigma_{WR}^2) &< \varepsilon_\phi^* \end{aligned} \quad (10.48a)$$

and

$$H_\sigma : \sigma_{TT}^2 / \sigma_{TR}^2 \geq 1 + \varepsilon_\sigma^* \quad \text{vs.} \quad K_\sigma : \sigma_{TT}^2 / \sigma_{TR}^2 < 1 + \varepsilon_\sigma^*. \quad (10.48b)$$

Exploiting the intersection-union principle formulated and proven in § 7.1, testing of (10.48) reduces to performing a valid test for each of the elementary problems. But in view of (10.45), except for renaming the equivalence bound to  $|\phi_T - \phi_R|/\sigma$ , (10.48a) is the same as the problem of testing for scaled average BE in the sense of § 10.2.2. Hence, all that is really left to do for constructing a test for establishing the proposed disaggregate criterion of population BE, is to derive a solution for the one-sided testing problem (10.48b). This will be the topic of § 10.4.2. In § 10.4.3, the rejection region of the complete disaggregate testing procedure will be presented, and its use illustrated by reanalyzing the BE trial considered in Example 10.1 of § 10.2.1. The section will be concluded by presenting simulation results on the power of the disaggregate test, and the sample sizes required for attaining given power against selected specific alternatives, respectively.

#### 10.4.2 A testing procedure for establishing one-sided bioequivalence with respect to total variability

The basic idea behind the construction of an exact level- $\alpha$  test for (10.48b) to be described in this subsection, goes back as far as the 1930s (Pitman, 1939; Morgan, 1939) and was exploited in a slightly modified version by Liu and Chow (1992) as well as Guilbaud (1993) for deriving tests for one-sided equivalence with respect to intra-subject rather than total variability of the pharmacokinetic measure of interest.

As before [recall p. 313], let us denote the log-bioavailabilities observed during the  $k$ th period in a subject randomized to sequence group  $T/R$  and  $R/T$  by  $X_{ki}$  and  $Y_{kj}$ , respectively. Furthermore, for any fixed positive number  $\omega$  let

$$U_{1i}^\omega = X_{1i} + \omega X_{2i}, \quad U_{2i}^\omega = (1/\omega)X_{1i} - X_{2i}, \quad i = 1, \dots, m, \quad (10.49a)$$

and

$$V_{1j}^\omega = Y_{2j} + \omega Y_{1j}, \quad V_{2j}^\omega = (1/\omega)Y_{2j} - Y_{1j}, \quad j = 1, \dots, n. \quad (10.49b)$$

Then, it follows from (9.28), (9.29), (10.1) and (10.2) that the pairs  $(U_{1i}^\omega, U_{2i}^\omega)$  and  $(V_{1j}^\omega, V_{2j}^\omega)$  form samples from two bivariate normal distributions with possibly different mean vectors but the same covariance matrix. Hence, in particular the correlation within any pair of  $U^\omega$ 's is the same as within any pair

of  $V^\omega$ 's, and the common correlation coefficient  $\rho_\omega$ , say, is easily computed to be

$$\rho_\omega = \frac{\sigma_{TT}^2 - \omega^2 \sigma_{TR}^2}{\sqrt{(\sigma_{TT}^2 + \omega^2 \sigma_{TR}^2)^2 - 4\omega^2 \sigma_S^4}}. \quad (10.50)$$

In view of the possible heterogeneity of the distributions of the  $(U_{1i}^\omega, U_{2i}^\omega)$  and the  $(V_{1j}^\omega, V_{2j}^\omega)$  with respect to the means,  $\rho_\omega$  cannot be estimated by the empirical correlation coefficient for the pooled sample consisting of all  $(U_{1i}^\omega, U_{2i}^\omega)$  and  $(V_{1j}^\omega, V_{2j}^\omega)$  together. Instead, pooling has to be done for the sample moments which have to be computed first separately for the  $(U_{1i}^\omega, U_{2i}^\omega)$  and the  $(V_{1j}^\omega, V_{2j}^\omega)$  which leads to the estimator

$$R_\omega = \left[ \left( SSX^{(1)} + SSY^{(2)} \right) - \omega^2 \left( SSX^{(2)} + SSY^{(1)} \right) \right] / \left[ \left( (SSX^{(1)} + SSY^{(2)}) + \omega^2 (SSX^{(2)} + SSY^{(1)}) \right)^2 - 4\omega^2 (SPX + SPY)^2 \right]^{1/2}. \quad (10.51)$$

In this formula,  $SSX^{(k)}$  and  $SSY^{(k)}$  stands for the (corrected) sum of squares of the  $X_{ki}$  and  $Y_{kj}$  ( $k = 1, 2$ ), respectively. Similarly,  $SPX$  and  $SPY$  have to be defined as  $\sum_{i=1}^m (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)$  and  $\sum_{j=1}^n (Y_{1j} - \bar{Y}_1)(Y_{2j} - \bar{Y}_2)$  where  $\bar{X}_k = \sum_{i=1}^m X_{ki}/m$  and  $\bar{Y}_k = \sum_{j=1}^n Y_{kj}/n$  ( $k = 1, 2$ ), respectively.

The appropriateness of  $R_\omega$  with  $\omega = \sqrt{1 + \varepsilon_\sigma^*}$  as a test statistic for (10.48b) follows from the following properties:

- (i) If  $\omega$  equals the true value of the ratio  $\sigma_{TT}/\sigma_{TR}$  of total standard deviations, then  $R_\omega$  has the same distribution as the empirical correlation coefficient for a single sample of size  $m + n - 1$  from a bivariate normal distribution with diagonal covariance matrix. Equivalently, this means that for  $\omega = \sigma_{TT}/\sigma_{TR}$ , the distribution of

$$T_\omega = \sqrt{m + n - 3} R_\omega / \sqrt{1 - R_\omega^2} \quad (10.52)$$

is central  $t$  with  $m + n - 3$  degrees freedom.

- (ii) For fixed sums of squares  $SSX^{(k)}, SSY^{(k)}$  ( $k = 1, 2$ ) and sums of products  $SPX, SPY$ , and arbitrary  $\omega_1, \omega_2 > 0$ , there holds the relation

$$\omega_1 \leq \omega_2 \Rightarrow T_{\omega_1} \geq T_{\omega_2}. \quad (10.53)$$

Property (i) is plausible from the obvious fact that  $\rho_\omega$  vanishes for  $\omega = \sigma_{TT}/\sigma_{TR}$  and that  $R_\omega$  is by definition the empirical correlation coefficient pooled over two independent samples from bivariate normal distributions with the same theoretical correlation  $\rho_\omega$  (a rigorous proof can be given by means of Theorems 7.3.2 and 7.6.1 of Anderson, 1984). The other property (ii) can

be established by examining the derivative with respect to  $\omega$  of the expression on the right-hand side of (10.51). Combining both properties shows that the decision rule

$$\text{Reject } H_\sigma \text{ if and only if } T_{\sqrt{1+\varepsilon_\sigma^*}} < t_{m+n-3; \alpha} \quad (10.54)$$

defines an unbiased test of exact level  $\alpha$  for (10.48b). Of course, this statement holds under the customary notational convention for central  $t$ -quantiles according to which  $t_{\nu; \gamma}$  symbolizes the lower  $100\gamma$ th percentile of a central  $t$ -distribution with  $\nu$  degrees of freedom (for any  $0 < \gamma < 1$ ).

### 10.4.3 Complete disaggregate testing procedure and illustrating example

Bearing equation (10.44) in mind, it becomes immediately clear that putting  $\varepsilon = 2\sqrt{\varepsilon_\phi^*}$  makes the alternative hypothesis  $\tilde{K}$  of (10.15) coincide with the first condition required by our aggregate criterion of population BE. Thus, the testing sub-problem (10.48a) admits an optimal solution which is obtained by applying the critical region (6.6) of the two-sample  $t$ -test for equivalence with  $\varepsilon = 2\sqrt{\varepsilon_\phi^*}$  and the within subject differences  $X_i^-$  and  $Y_j^-$  as the primary observations. This gives the rejection region

$$\left\{ |\bar{X}^- - \bar{Y}^-| / \tilde{S}^- < \tilde{C}_{\alpha; m, n} \left( 2\sqrt{\varepsilon_\phi^*} \right) \right\} \quad (10.55)$$

where  $\bar{X}^-$ ,  $\bar{Y}^-$  and  $\tilde{S}^-$  have to be computed in the way explained on p. 316. Now, according to the intersection-union principle, we have to combine (10.55) and (10.54) into the aggregated decision rule:

Reject lack of population BE if and only if there holds both

$$\begin{aligned} |\bar{X}^- - \bar{Y}^-| / \tilde{S}^- &< \tilde{C}_{\alpha; m, n} \left( 2\sqrt{\varepsilon_\phi^*} \right) \quad \text{and} \\ \sqrt{m+n-3} R_{\sqrt{1+\varepsilon_\sigma^*}} / \sqrt{1 - R_{\sqrt{1+\varepsilon_\sigma^*}}^2} &< t_{m+n-3, \alpha}. \end{aligned} \quad (10.56)$$

*Example 10.1 (continued)*

We illustrate the use of the aggregate decision rule (10.56) by applying it to the same data set that has been analyzed in § 10.2 by means of the interval inclusion test for average BE and the two-sample  $t$ -test for scaled average BE. Table 10.1 [→ p. 319] displays the raw data in a format well suited to carrying out almost all necessary computational steps by means of a simple pocket calculator. To begin with the first component problem  $H_\phi$  vs.  $K_\phi$  [→ (10.48a)], we already know from § 10.2.2 that for  $\varepsilon_\phi^* = (.74/2)^2 = .1369$ , the UMPI test rejects at the 5%-level. Thus, there remains to extend the analysis

of the data set by carrying out the test for one-sided equivalence of the total variabilities derived in § 10.4.2.

The raw input data for computing the test statistic (10.52) are the entries in the middle rows of Table 10.1. Calculating the sums of squares and products appearing on the right-hand side of equation (10.51) with these values gives:

$$SSX^{(1)} = 0.5814, \quad SSX^{(2)} = 0.3503, \quad SPX = 0.3133;$$

$$SSY^{(1)} = 1.0280, \quad SSY^{(2)} = 1.3879, \quad SPY = 1.0059.$$

If we specify  $\varepsilon_\sigma^* = 1.00$  for the constant which determines the upper equivalence limit in the hypotheses (10.48b), then plugging this value into (10.51) yields  $R_{\sqrt{1+\varepsilon_\sigma^*}} = -2.715$ . The corresponding value of the  $t$ -statistic (10.52) is  $T_{\sqrt{1+\varepsilon_\sigma^*}} = -1.3229$ . Since the .05-quantile of a central  $t$ -distribution with  $m + n - 3 = 22$  degrees of freedom is  $-1.7171$ , (10.54) implies that the null hypothesis  $H_\sigma$  of a relevant increase of the total variability associated with the test formulation of the drug, cannot be rejected. Hence, relying on the specifications  $\varepsilon_\sigma = 1.00$  and  $\alpha = .05$ , one-sided equivalence with respect to total variability cannot be established, and as a final consequence it has to be stated that the data of Table 10.1 do not contain sufficient evidence in support of population BE in the sense of the disaggregate criterion underlying (10.48).

For comparison, we mention that an application of the aggregate criterion of population BE suggested in the current FDA guidance to the same data leads to an identical qualitative conclusion. In fact, the point estimate of the parametric function (10.43) is obtained to be  $\hat{\theta}_P = .4293$ , and taking  $B = 100,000$  bootstrap samples from the observed empirical distributions gave a (bias-corrected) upper 95%-confidence bound of 1.0609 for  $\theta_P$  which is actually quite large.

#### 10.4.4 Some results on power and sample sizes

If  $\Pi$  denotes the power of any statistical test obtained by means of the intersection-union principle against an arbitrarily selected specific alternative and each elementary test  $\phi_\nu$ , ( $\nu = 1, \dots, q$ ) [cf. (7.3)] has rejection probability  $\Pi_\nu$  under the same parameter constellation, then an application of the elementary Bonferroni inequality yields the relationship

$$\Pi \geq 1 - \sum_{\nu=1}^q (1 - \Pi_\nu). \quad (10.57)$$

Using the expression appearing on the right-hand side of this inequality as a first order approximation to the power of the compound test can be recommended only for settings in which for any  $\nu_1 \neq \nu_2$ , the acceptance regions of subtests  $\phi_{\nu_1}$  and  $\phi_{\nu_2}$  correspond to almost disjoint events. Furthermore, the approximation based on (10.57) is of real use for practice only if the power of

each component test is a quantity easily accessible by direct computational techniques.

Among the two individual tests making up our decision procedure for the assessment of disaggregate population BE, the first one admits exact power calculations as shown in § 6.1. In contrast to the two-sample  $t$ -statistic, the correlation coefficient  $R_{\sqrt{1+\varepsilon_\sigma^*}}$  used for assessing possible differences between drug formulations in total variability, has a rather complicated distribution, and exact computation of nonnull rejection probabilities of the corresponding test would require numerical integration in several dimensions. Hence, in the case being our concern here, even the right-hand side of (10.57) cannot be evaluated exactly without recourse to advanced and time-consuming numerical methods. In view of this, the power of the test for one-sided equivalence with respect to total variability as well as that of the complete disaggregate testing procedure was investigated by means of Monte Carlo simulation.

Table 10.5 *Rejection probabilities of the  $t$ -test for scaled average bioequivalence ( $\leftrightarrow \Pi_1$ ), the test for one-sided equivalence with respect to total variability ( $\leftrightarrow \Pi_2$ ), and the disaggregate test for PBE as a whole ( $\leftrightarrow \Pi$ ). [Except for  $\Pi_1$ , all probabilities are based on simulation, with 100,000 replications per experiment.] (Partially reproduced from Wellek, 2000b, with kind permission by John Wiley & Sons, Inc.)*

$m$	$n$	$\tilde{\varepsilon}_\phi$ †	$\varepsilon_\sigma^*$	$\sigma_{TR}^2/\sigma_S^2$	$\Pi_1$	$\Pi_2$	$\Pi$
12	12	0.74	1.00	2.00	.24453	.57116	.13881
12	12	1.00	1.00	2.00	.57214	.57015	.32245
12	12	1.00	2.00	3.00	.57214	.84551	.48139
12	12	1.40	2.25	3.00	.90558	.88555	.80123
12	12	1.00	2.00	5.00	.57214	.82286	.47036
12	12	1.00	1.50	5.00	.57214	.69319	.39415
18	18	0.74	1.00	2.00	.44688	.74091	.32827
18	18	1.00	1.00	2.00	.81196	.74109	.60117
18	18	1.00	1.00	3.00	.81196	.67928	.55122
18	18	1.00	2.00	3.00	.81196	.95286	.77391
18	18	1.00	2.00	5.00	.81196	.94106	.76486
24	24	0.74	1.00	2.00	.63693	.84767	.53882
24	24	1.00	1.00	2.00	.92162	.84957	.78173
24	24	1.00	1.00	3.00	.92162	.79226	.72912
24	24	1.00	2.00	3.00	.92162	.98714	.91064
24	24	1.00	2.00	5.00	.92162	.98215	.90421
24	24	1.00	0.50	2.00	.92162	.47040	.43184

† equivalence limit for  $2(\phi_T - \phi_R)/\sqrt{\sigma_{WT}^2 + \sigma_{WR}^2}$

All simulation results shown in Table 10.5 refer to the 5% level and to balanced designs with equal numbers of subjects assigned to both sequence

groups. As the only specific alternative of interest, perfect coincidence of the marginal distributions corresponding to the drug formulations, i.e.,  $\phi_T = \phi_R$  and  $\sigma_{TT}^2 = \sigma_{TR}^2$  is assumed. Inspecting the numerical material displayed in the table leads to the following major conclusions:

- The minimum sample size required to guarantee a reasonable overall power  $\Pi$  of the disaggregate testing procedure (10.56), is a complicated function both of the equivalence limits  $\tilde{\varepsilon}_\phi$ ,  $1 + \varepsilon_\sigma^*$ , and the ratio  $\sigma_{TR}^2/\sigma_S^2$  between the total and the inter-subject variance.
- For fixed sample sizes  $m, n$  and given upper equivalence limit  $1 + \varepsilon_\sigma^*$  for  $\sigma_{TT}^2/\sigma_{TR}^2$ , the power  $\Pi_2$  of the test for equivalence in total variability decreases in  $\sigma_{TR}^2/\sigma_S^2$ .
- The approximation to the overall power  $\Pi$  based on (10.57) is poor if the power of at least one of the two subtests is smaller than .70. However, if both  $\Pi_1$  and  $\Pi_2$  are larger than .80, the approximation is remarkably accurate.

#### 10.4.5 Discussion

Technically, there is no problem to obtain a left-sided analogue of the test (10.54) for noninferiority with respect to total variability and to combine it with the right-sided version by means of the intersection-union principle into a test for equivalence in the strict, i.e., two-sided sense (for an unbiased test for two-sided equivalence with respect to intra-subject variability which is uniformly more powerful than the corresponding double one-sided test see Wang, 1997). However, reducing bioequivalence assessment with respect to variability to a one-sided testing problem of the form (10.48b) seems conceptually much more adequate than working with the two-sided version of that testing problem. To be sure, it is obvious that “hyperavailability” of a drug formulation under assessment in the sense of an increased population average may cause serious problems in patients treated with the drug. But it is hard to see any reason why a relevantly reduced variability of the pharmacokinetic measure of interest should be considered an undesirable property of a drug formulation which its potential consumer must be protected against. Hence, for purposes of establishing population bioequivalence, it seems fully sufficient indeed to give convincing evidence for the truth of (10.48a) and (10.48b) whereas the double two-sided criterion  $(\phi_T - \phi_R)^2/(\sigma_{WT}^2 + \sigma_{WR}^2) < \varepsilon_\phi^*$ ,  $1 - \varepsilon_\sigma^* < \sigma_{TT}^2/\sigma_{TR}^2 < 1 + \varepsilon_\sigma^{**}$  with  $\varepsilon_\sigma^* < 1$  seems unnecessarily restrictive. This restrictiveness entails a rather drastic loss in the power of the associated testing procedure. For instance, replacing the equivalence region  $0 < \sigma_{TT}/\sigma_{TR} < 2.00$  by  $0.50 < \sigma_{TT}/\sigma_{TR} < 2.00$  in the situation covered by line 14 of Table 10.5 would imply that the power of the second subtest and of the overall testing procedure drops from .7923 and .7291 to .5866 and .5392, respectively.

Finally, it seems worth noticing that for both elementary tests we did combine to form our disaggregate procedure, suitable nonparametric or distribution-free analogues are available. A distribution-free procedure for testing a hypothesis which coincides with that of (10.48a) in the standard parametric model is obtained by applying the Mann-Whitney test for equivalence discussed in § 6.2 to the within-subject differences  $X_i^-$  and  $Y_j^-$ . Furthermore, nonparametric tests for one-sided equivalence with respect to intra-subject variability are provided in Section 2 of Liu and Chow (1992). The latter can readily be adopted for the problem of testing for one-sided equivalence with respect to total variability.

---

## 10.5 Bioequivalence assessment as a problem of comparing bivariate distributions

Although the same basic model is used for describing the *primary* observations, i.e., the log-bioavailabilities observed in the individual periods of the trial [recall p. 313], the view taken in this final subsection on the problem of BE assessment differs radically from all approaches treated up to now since no reduction of the pairs  $(X_{1i}, X_{2i}), (Y_{1j}, Y_{2j})$  to univariate quantities will be carried out. The reason for this change of perspective is so elementary that one may ask why it has been totally ignored in the existing literature on the topic: If both formulations of the drug under study are actually identical,  $(Y_{11}, Y_{21}), \dots, (Y_{1n}, Y_{2n})$  must be just a replicate sample from the *same* bivariate distribution from which the random sample  $(X_{11}, X_{21}), \dots, (X_{1m}, X_{2m})$  making up the data set for sequence group *T-R*, has been taken. Thus, for identical formulations, we must in particular have  $E(X_{1i}, X_{2i}) = E(Y_{1j}, Y_{2j})$  (for arbitrary  $i \in \{1, \dots, m\}$ ,  $j \in \{1, \dots, n\}$ ) which is incompatible with a nonzero difference both of the direct treatment and the carryover effects. In fact, if (10.2) is replaced with (9.31) [with  $(\phi_A, \phi_B)$  being renamed  $(\phi_T, \phi_R)$ ], then we have  $E(X_{1i}) - E(Y_{1j}) = \phi_T - \phi_R$ ,  $E(X_{2i}) - E(Y_{2j}) = (\phi_R - \phi_T) + (\lambda^{(1)} - \lambda^{(2)})$  so that  $\phi_T \neq \phi_R \Rightarrow E(X_{1i}) \neq E(Y_{1j})$  and  $(\phi_T = \phi_R, \lambda^{(1)} \neq \lambda^{(2)}) \Rightarrow E(X_{2i}) \neq E(Y_{2j})$ . As a clear conclusion of this simple argument, we can state that inequality of carryover effects is a specific form of inequivalence rather than a violation of the conditions for the validity of a suitable BE assessment procedure using the data from both periods of the trial. (Essentially the same fact has been pointed out from a nonequivalence perspective by Jones and Kenward, 2003, p. 44.) Thus, in what follows we adopt the view that assessment of BE requires comparison of the distributions underlying the pairs  $(X_{1i}, X_{2i})$  and  $(Y_{1j}, Y_{2j})$ .

Assuming the usual parametric model in the simplified version with treat-

ment-independent within-subject variances, we have

$$(X_{1i}, X_{2i}) \sim \mathcal{N}((\mu_1, \mu_2), \Sigma), \quad (Y_{1j}, Y_{2j}) \sim \mathcal{N}((\nu_1, \nu_2), \Sigma), \quad (10.58)$$

where the elements of the common covariance matrix  $\Sigma$  are given by

$$\sigma_1^2 = \sigma_S^2 + \sigma_e^2, \quad \sigma_{12} = \sigma_{21} = \sigma_S^2, \quad \sigma_2^2 = \sigma_S^2 + \sigma_e^2. \quad (10.59)$$

Clearly, testing for equivalence of the two bivariate normal distributions can be done by means of any of the procedures considered in § 8.2. We focus here on the equivalence version of Hotelling's two-sample  $T^2$ -test and start with observing that in the bivariate case, Mahalanobis distance between two Gaussian distributions with common covariance matrix can be written

$$\Delta^2(\boldsymbol{\mu}, \boldsymbol{\nu}; \Sigma) = \frac{(\mu_1 - \nu_1)^2 \sigma_2^2 - 2\sigma_{12}(\mu_1 - \nu_1)(\mu_2 - \nu_2) + (\mu_2 - \nu_2)^2 \sigma_1^2}{\sigma_1^2 \sigma_2^2 - \sigma_{12}^2}. \quad (10.60)$$

In the bivariate approach, we require of bioequivalent drug formulations that the distance between the bivariate Gaussian distributions underlying the nonreduced data set obtained in both sequence groups, be smaller than some specified positive bound, say  $\Delta_o^2$ . Accordingly, we propose to formulate the testing problem

$$H_b : \Delta^2(\boldsymbol{\mu}, \boldsymbol{\nu}; \Sigma) \geq \Delta_o^2 \text{ versus } K_b : \Delta^2(\boldsymbol{\mu}, \boldsymbol{\nu}; \Sigma) < \Delta_o^2, \quad (10.61)$$

where the subscript b is added in order to indicate that this is the bivariate formulation of the problem of assessing bioequivalence with data obtained from a standard comparative bioavailability study.

It is interesting to have a closer look at the geometric shape of the region in the parameter space which corresponds to the alternative hypothesis of (10.61). Under the restricted covariance structure (10.59) arising from the standard model for BE trials,  $K_b$  can easily be shown to hold true if and only if we have

$$\left[ \left( \frac{(\phi_T - \phi_R) + (\lambda^{(2)} - \lambda^{(1)})/2}{\sigma_e/\sqrt{2}} \right)^2 + \left( \frac{\lambda^{(2)} - \lambda^{(1)}}{2\sqrt{\sigma_S^2 + \sigma_e^2/2}} \right)^2 \right]^{1/2} < \Delta_o. \quad (10.62)$$

Obviously, (10.62) defines a circular disk of radius  $\Delta_o$  around zero in a suitable two-dimensional space. The most important feature of this space is that it refers to a two-dimensional parameter whose Euclidean distance from the origin increases on  $\{(\phi_T, \phi_R, \lambda^{(1)}, \lambda^{(2)}) \mid (\phi_T - \phi_R)(\lambda^{(1)} - \lambda^{(2)}) > 0\}$  both in  $|\phi_T - \phi_R|$  and  $|\lambda^{(1)} - \lambda^{(2)}|$ . Moreover, the circular disk corresponding to (10.62)

provides a good approximation to the equivalence region specified by  $K_b$  even in cases where two different within-subject variances  $\sigma_{eT}^2$  and  $\sigma_{eR}^2$  have to be taken into account. In order to see this, it suffices to note that (10.62) reflects the independence of the natural estimators of  $(\phi_T - \phi_R) + (\lambda^{(2)} - \lambda^{(1)})/2$  and  $(\lambda^{(2)} - \lambda^{(1)})/2$  implied by  $\sigma_{eT}^2 = \sigma_{eR}^2$ . Moreover, in the general case, the square of the correlation coefficient between these estimators is easily computed to be  $\frac{(\sigma_{eT} - \sigma_{eR})^2(\sigma_{eT} + \sigma_{eR})^2}{(\sigma_{eT}^2 + \sigma_{eR}^2)(4\sigma_S^2 + \sigma_{eT}^2 + \sigma_{eR}^2)}$  which is typically still close to zero since the within-subject variances will be much smaller than the between-subject variance.

In the bivariate case, the two-sample  $T^2$ -statistic admits the explicit representation

$$T^2 = \frac{mn}{N(S_1^2 S_2^2 - S_{12}^2)} \left[ (\bar{X}_1 - \bar{Y}_1)^2 S_2^2 - 2S_{12}^2 (\bar{X}_1 - \bar{Y}_1)(\bar{X}_2 - \bar{Y}_2) + (\bar{X}_2 - \bar{Y}_2)^2 S_1^2 \right], \quad (10.63)$$

with  $S_1^2$ ,  $S_2^2$  and  $S_{12}$  defined as in (8.34). Specializing (8.37) and (8.38), the critical region of an exact UMPI level- $\alpha$  test for (10.61) can be written

$$\left\{ T^2 < 2((N-2)/(N-3)) F_{2,N-3;\alpha}(mn\Delta_o^2/N) \right\}, \quad (10.64)$$

and the power against any specific alternative  $(\mu, \nu, \Sigma)$  with  $\Delta^2(\mu, \nu, \Sigma) = \Delta_a^2 \in [0, \Delta_o^2]$  is given by

$$\beta(\Delta_a^2) = P[F_{2,N-3}(mn\Delta_a^2/N) \leq F_{2,N-3;\alpha}(mn\Delta_o^2/N)]. \quad (10.65)$$

Except for the change referring to the degrees of freedom, the notation in (10.64) and (10.65) is the same as in §8.2.1. I.e.,  $F_{2,N-3}(\psi^2)$  stands for a random variable following a noncentral  $F$ -distribution with 2,  $N-3$  degrees of freedom and noncentrality parameter  $\psi^2 \geq 0$ , whereas  $F_{2,N-3;\alpha}(\psi^2)$  denotes the lower  $100\alpha$  percentage point of this distribution.

Of course, without the assumption of homoskedasticity and bivariate normality of the distributions under comparison, it cannot even be taken for granted that the rejection region (10.64) has size  $\alpha$ , which raises the question of robustness against departures from one or the other of these assumptions. For higher-dimensional data, the sensitivity of the  $T^2$ -test for equivalence against deviations from homoskedasticity has been investigated in Chapter 8. For the bivariate case, we conducted separate simulation studies to assess the impact of nonnormality and heteroskedasticity both on size and power to detect exact coincidence of both mean vectors. Table 10.6 shows the simulated rejection probabilities under constellations with non-Gaussian forms of the distributions, and first and second moments defining a point in the parameter space lying on the common boundary of the hypotheses and at the most extreme specific alternative, respectively.

Table 10.6 *Simulation results on the sensitivity of the  $T^2$ -test for (bio-)equivalence against nonnormality of distributions.*

$m$	$n$	$\sigma_S^2$	$\sigma_e^2$	Distribu- tional Shape	Reject. Prob. at $\Delta^2 = \Delta_o^2$	Power at $\Delta^2 = 0$
10	10	*	*	Gaussian	.0500 <sup>†)</sup>	.3581 <sup>†)</sup>
"	"	.165	.020	Logistic	.0487	.3546
"	"	.20	.10	" "	.0416	.3535
"	"	.165	.020	Exponential	.0428	.3375
"	"	.20	.10	" "	.0416	.3349
15	15	*	*	Gaussian	.0500 <sup>†)</sup>	.5840 <sup>†)</sup>
"	"	.165	.020	Logistic	.0496	.5793
"	"	.20	.10	" "	.0480	.5793
"	"	.165	.020	Exponential	.0451	.5741
"	"	.20	.10	" "	.0431	.5737

<sup>†)</sup> Exact values, depending on all parameters only through  $\Delta^2$

The distributional forms appearing in the fifth column refer to the marginal distributions of linearly transformed pairs  $(D_i^{(1)}, Z_i^{(1)})$  [ $\leftrightarrow$  group  $T/R$ ],  $(D_j^{(2)}, Z_j^{(2)})$  [ $\leftrightarrow$  group  $R/T$ ] with uncorrelated components. In the homoskedastic case  $\sigma_{eT}^2 = \sigma_{eR}^2$ , such transforms are readily obtained by setting  $D_i^{(1)} = X_{1i} - X_{2i}$ ,  $Z_i^{(1)} = X_{1i} + X_{2i}$  and  $D_j^{(2)} = Y_{1j} - Y_{2j}$ ,  $Z_j^{(2)} = Y_{1j} + Y_{2j}$ , respectively. Clearly, the invariance properties of the  $T^2$ -test imply that it makes no difference whether the  $(X_{1i}, X_{2i})$ ,  $(Y_{1j}, Y_{2j})$  or the  $(D_i^{(1)}, Z_i^{(1)})$ ,  $(D_j^{(2)}, Z_j^{(2)})$  are treated as the primary data. Table 10.7 shows analogous simulation results for bivariate Behrens-Fisher situations. For generating the data for the Monte Carlo experiments performed for the purpose of studying the influence of heteroskedasticity on the size of the critical region (10.64), Mahalanobis distance was computed with respect to the average  $\bar{\Sigma} = (1/2)(\Sigma_1 + \Sigma_2)$  of both group-specific covariance matrices, and the mean vectors  $\mu$ ,  $\nu$  were specified in a way ensuring that  $\Delta^2(\mu, \nu; \bar{\Sigma})$  attains the bound  $\Delta_o^2$  set to equivalent homoskedastic Gaussian distributions under the original alternative hypothesis  $K_b$ .

Both series of simulation experiments refer to a nominal significance level of 5% and an equivalence bound  $\Delta_o^2$  of unity. All rejection probabilities shown in the tables are estimates based on 40,000 replications of the respective Monte Carlo experiment. The insights provided by the results into both aspects of the robustness problem arising in the present context are as clear-cut as encouraging:

- Both kinds of deviations from the original model affect mainly the size of the test.

- The differences found between the nominal level and the true rejection probabilities on the boundary of the region of nonequivalence point to the overconservative direction.
- Given the size of both samples, overconservatism is more marked under heteroskedasticity than deviations from the Gaussian form of the marginal distributions.

Table 10.7 *Simulation results on the sensitivity of the  $T^2$ -test for (bio-) equivalence against heteroskedasticity.*

$m$	$n$	$\sigma_S^2$	$\sigma_{eT}^2$	$\sigma_{eR}^2$	Reject. Prob. under $\Delta^2(\mu, \nu; \bar{\Sigma}^{\dagger}) = \Delta_o^2$	Power against $\mu = \nu$
10	10	.165	.005	.020	.0406	.3578
"	"	.20	.05	.20	.0325	.3588
15	15	.165	.005	.020	.0361	.5810
"	"	.20	.05	.20	.0277	.5823
20	20	.165	.005	.020	.0344	.7539
"	"	.20	.05	.20	.0264	.7532

${}^{\dagger})\bar{\Sigma} \equiv (1/2)(\Sigma_1 + \Sigma_2)$

### Example 10.3

As an illustration of the bivariate approach to testing for scaled average BE, we apply it to the log-bioavailabilities shown in Table 10.8. With these data, the sample means and pooled estimates of the covariance and both variances are calculated

$$\begin{aligned}\bar{X}_1 &= 4.8819 & , \quad \bar{X}_2 &= 4.8623 ; \\ \bar{Y}_1 &= 5.0921 & , \quad \bar{Y}_2 &= 5.0697 ; \\ S_1^2 &= 0.1839 & , \quad S_2^2 &= 0.1490 ; \quad S_{12} = 0.1405\end{aligned}$$

Plugging in these values in Equation (10.63) yields

$$\begin{aligned}T^2 &= \frac{100}{20(0.1839 \cdot 0.1490 - 0.1405^2)} \left[ (4.8819 - 5.0921)^2 \cdot \right. \\ &\quad .1490 - 2 \cdot 0.1405 \cdot (4.8819 - 5.0921) \cdot (4.8623 - 5.0697) \\ &\quad \left. + (4.8623 - 5.0697)^2 \cdot 0.1839 \right] = 1.4643 .\end{aligned}$$

Table 10.8 Logarithmically transformed AUC's measured in both periods of a standard bioequivalence trial of two other formulations of nifedipine [different from those referred to in Example 10.2].

Sequence Group  $T/R$

$i$	1	2	3	4	5	6
$X_{1i}$	5.32833	4.60833	4.61290	4.85399	4.82933	4.61250
$X_{2i}$	4.91018	4.98835	4.53084	5.00148	4.88278	4.68550
	7	8	9	10		
	4.89988	4.48080	6.00856	4.58419		
	4.83594	4.43013	5.72340	4.63451		

Sequence Group  $R/T$

$j$	1	2	3	4	5	6
$Y_{1j}$	4.97670	5.79190	5.37921	5.11251	5.29075	4.64138
$Y_{2j}$	5.32219	5.51966	5.17163	4.98796	5.31198	4.34402
	7	8	9	10		
	4.52988	4.70318	5.35740	5.13770		
	4.79501	4.50227	5.56355	5.17898		

On the other hand, specifying  $\Delta_0 = 1$  and using the SAS intrinsic function `finv` for computing the noncentral  $F$ -quantile required for determining the critical upper bound to  $T^2$ , we obtain the latter as

$$2 \frac{N-2}{N-3} F_{2,N-3; \alpha}(mn\Delta_0^2/N) = 2 \cdot \frac{18}{17} \text{finv}(.05, 2, 17, 5.00) \\ = 2.1176 \cdot 0.4551 = 0.9637,$$

provided of course, that the test has to be carried out at the usual level  $\alpha = 5\%$ . Since the observed value of  $T^2$  clearly fails to remain under this bound, we have to accept the null hypothesis  $H_b$  of (10.61) with the present data, despite the fact that  $\Delta_0$  has been chosen to be much larger than recommended earlier [ $\rightarrow$  § 1.7] for the univariate version of the same test.

It is of considerable interest to note that this negative result is almost exclusively due to a comparatively large estimated difference between the carryover effects, whereas estimating the difference of treatment effects gives a standardized value being as small as .0272. Thus, the data set shown in Table 10.7 is an instance of the possibility that the bivariate and the conventional

univariate approach to testing for (scaled) average BE may lead to opposite conclusions. With regard to contributions to the existing literature dealing with the problem of BE assessment from a point of view declared explicitly to be multivariate in nature (see Chinchilli and Elswick Jr., 1997), it is important to note that a different kind of multivariateness is intended there. In fact, these authors propose to avoid reduction of the serum-concentration profile to simple univariate measures of bioavailability like  $AUC$  and  $C_{max}$  by using a measure of distance between the profiles obtained from the same subject which integrates the distances at all relevant points on the time axis.

# 11

---

## *Tests for relevant differences between treatments*

---

### 11.1 Introduction

Formally speaking, the most straightforward way of explaining what we mean by a test for relevant differences, is to state that the testing problem to which such a procedure is tailored, arises from interchanging the two hypotheses making up an equivalence problem without modifying them otherwise, except for replacing the interval corresponding to the former alternative with its closure. Thus, in a generic formulation, all problems to be considered in this chapter read

$$H_* : \theta \in [\theta_0 - \varepsilon_1, \theta_0 + \varepsilon_2] \quad \text{versus} \quad K_* : \theta < \theta_0 - \varepsilon_1 \vee \theta > \theta_0 + \varepsilon_1. \quad (11.1)$$

All symbols appearing in (11.1) have the same meaning as before [cf. in particular § 1.5] and need not to be defined again. The reason for including the endpoints of the interval of irrelevant deviations from  $\theta_0$  in  $H_*$  is purely technical in nature: Generally, in a hypothesis testing problem of any form, it is thought to be desirable that the power function of any unbiased level- $\alpha$  test takes on  $\alpha$  as its maximum value under the null hypothesis  $H_0$ . Given the power function is continuous, this condition is satisfied whenever the subset in the parameter space corresponding to  $H_0$ , is compact.

From a conceptual perspective, problems of testing for relevant differences are much more directly related to traditional two-sided than equivalence problems. Actually, the null hypothesis  $H_*$  of (11.1) is simply a generalization of the ordinary point null hypothesis specifying  $\theta = \theta_0$  and reduces to the latter by choosing both relevance margins to equal zero. Replacing under the null hypothesis the single point  $\theta_0$  in the parameter space of  $\theta$  by some nondegenerate interval, is a promising step toward finding a way around the problems connected with the plain fact that considering an effect being significant in terms of the result of a traditional two-sided test as substantial or scientifically relevant can be grossly misleading.

Although the scope of potential applications of tests for problems of the form (11.1) is certainly not less broad than that of equivalence testing procedures, the role that inferential procedures of that type have played up to now in the practice of statistical data analysis has been marginal at best. This

holds true notwithstanding the fact that in the medical statistics literature it was pointed out already decades ago (cf. Victor, 1987) that analyzing studies run to demonstrate superiority of one treatment over the other by means of tests for relevant difference rather than tests of traditional null hypotheses should be the rule rather than the exception.

Perhaps, one of the major reasons explaining this state of the matter has to do with complications involved in establishing sample size criteria for tests for relevant differences. Actually, as will become apparent from the results to be presented in the sections to follow, the power function of each reasonable test of (11.1) with  $\theta$  as the only parameter upon which the family of distributions under consideration depends, is convex from below with a unique global minimum taken on at, or in a neighborhood of, the target value  $\theta_0$ . Thus, determining the sample size for a trial to be analyzed by means of a test of this form, requires specification of a point in the parameter space whose distance from  $\theta_0$  exceeds the relevance margins  $\varepsilon_\nu$  ( $\nu = 1, 2$ ) by some additional relevance threshold defining the alternative against which the desired level of power has to be attained. Accordingly, even in a one-parameter setting, complete statistical planning of a study whose confirmatory analysis is to be based on a test for a problem of the form (11.1), entails specification of three extra constants in addition to the significance level  $\alpha$ .

A pragmatic proposal we will follow in the remainder of this chapter for choosing the numerical values of both the margins  $\varepsilon_\nu$ , and the specific alternative to be detected with given power, is to make use of the specifications proposed in Table 1.1 for equivalence testing in the following way: The interval  $[\theta_0 - \varepsilon_1, \theta_0 + \varepsilon_2]$  of nonrelevant deviations will be set equal to what was declared there to be a strict specification of the equivalence range, and the specific alternative to one of the endpoints of the equivalence interval in its “liberal” version. Taking the paired  $t$ -test setting as a specific example, this amounts to setting under  $H_*$  the limits  $\pm .25$  to  $\delta/\sigma$ , whereas the prespecified power has to be attained against  $\delta/\sigma = .5$ , and so on for the other settings covered by Table 1.1.

## 11.2 Exploiting the duality between testing for two-sided equivalence and existence of relevant differences

Essentially, given the margins, the only difference between a problem of testing for relevant differences in the sense of (11.1) and the corresponding two-sided equivalence problem is that null and alternative hypotheses switch their roles. Hence, it seems reasonable to generate tests for the former by reinterpreting an equivalence testing procedure available for the same setting in the following way: Decision has to be taken in favor of  $K_*$  if and only if the equivalence

test cannot reject its null hypothesis, implying that what was formerly a type-II error becomes now an error of the first kind, and the other way round. Representing any test by its critical function (which, except for randomized decisions, is simply the indicator of its rejection region) and indicating whether the test is to be used for establishing equivalence or existence of relevant differences by appending the subscript  $eq$  or  $rd$  to the symbol  $\phi$  commonly used for critical functions of statistical tests, the relationship we have in mind can be made precise by writing:

$$\phi_{rd} = 1 - \phi_{eq}. \quad (11.2)$$

Since the rejection probability under any  $\theta$  of both tests is given by  $E_\theta(\phi_{rd})$  and  $E_\theta(\phi_{eq})$  where expectation has to be taken with respect to the joint distribution of the random sample under analysis, it is clear that a minimum requirement for obtaining from (11.2) a valid level- $\alpha$  test of (11.1) is that we have

$$E_\theta(\phi_{eq}) \geq 1 - \alpha \quad \text{for all } \theta \in [\theta_0 - \varepsilon_1, \theta_0 + \varepsilon_2]. \quad (11.3)$$

In other words, the equivalence test  $\phi_{eq}$  to be started from, must satisfy the following twofold condition:

- (i) As a test of the null hypothesis  $H : \theta \leq \theta_0 - \varepsilon_1$  or  $\theta \geq \theta_0 + \varepsilon_2$ ,  $\phi_{eq}$  must be performed at (nominal) "significance" level  $\alpha' = 1 - \alpha$  rather than  $\alpha$ ;
- (ii)  $\phi_{eq}$  has to be unbiased at this level  $\alpha'$ .

It is important to note that, due to their lack of unbiasedness, confidence interval inclusion tests are generally *not* suitable for being inverted in tests for relevant differences through applying the duality principle (11.2). In order to see this, we consider the one-sample problem with normally distributed data of unit variance, symmetric choice  $\varepsilon_1 = \varepsilon_2 = \varepsilon$  of the margins, and  $\theta_0 = 0$  as the target value of the parameter  $\theta = E(X)$  of interest. As shown in § 4.4 [cf. p. 64], the confidence interval inclusion test at level  $\alpha' = 1 - \alpha$  of  $H : |\theta| \geq \varepsilon$  versus  $K : |\theta| < \varepsilon$  rejects if and only if the observed mean computed from  $n$  mutually independent observations satisfies  $\sqrt{n}|\bar{X}_n| < \sqrt{n}\varepsilon - u_\alpha$ . Applying (11.2), this is converted to the decision rule of rejecting  $H_*$  if and only if it turns out that

$$\begin{aligned} \bar{X}_n &\geq \varepsilon - u_\alpha/\sqrt{n} \quad \text{or} \quad \bar{X}_n \leq -\varepsilon + u_\alpha/\sqrt{n} \\ \Leftrightarrow \quad \bar{X}_n - u_{1-\alpha}/\sqrt{n} &\geq \varepsilon \quad \text{or} \quad \bar{X}_n + u_{1-\alpha}/\sqrt{n} \leq -\varepsilon. \end{aligned}$$

The second of these conditions for rejecting  $H_*$  can be rephrased by saying that the equal-tails  $(1 - 2\alpha)$ -confidence interval for  $\theta$  must not have a nonempty intersection with the indifference zone  $(-\varepsilon, \varepsilon)$  so that applying the duality principle (11.2) transforms the interval inclusion test for equivalence into an *interval exclusion test for relevant differences*. From the first formulation of the rejection rule of this interval exclusion test, it readily follows that

its rejection probability under any  $\theta \in \mathbb{R}$  is given by

$$1 - \Phi(\sqrt{n}(\varepsilon - \theta) + u_{1-\alpha}) + \Phi(-\sqrt{n}(\varepsilon + \theta) - u_{1-\alpha}).$$

For  $\theta = \varepsilon$ , the probability of rejecting  $H_*$  is thus

$$1 - \Phi(u_{1-\alpha}) + \Phi(-\sqrt{n}2\varepsilon - u_{1-\alpha}).$$

In view of  $1 - \Phi(u_{1-\alpha}) = \alpha$  and  $\Phi(\cdot) > 0$ , this shows that the rejection probability of the interval exclusion test at the common boundary of the hypotheses (11.1) exceeds  $\alpha$ . Accordingly, the test obtained by means of the duality principle fails to maintain the significance level in the present case.

Whenever the above conditions (i) and (ii) for an application of the duality principle are satisfied, the practical implementation of a test for relevant differences is straightforward. In a first step, one determines critical constants  $C_1^*$ ,  $C_2^*$ ,  $\gamma_1^*$ ,  $\gamma_2^*$  (which, in multiparameter settings, might also depend on the value of some conditioning statistic) by solving the system (3.9) [see p. 42] with  $\theta_\nu = \theta_\circ + (-1)^\nu \varepsilon_\nu$ ,  $\nu = 1, 2$ , and  $1 - \alpha$  instead of  $\alpha$ . Subsequently, decision is taken in favor of the alternative hypothesis  $K_*$  of relevant differences, if the observed value of the test statistic  $T(\mathbf{X})$  falls *outside* the interval  $[C_1^*, C_2^*]$ . For  $C_1^* < T(\mathbf{X}) < C_2^*$ , the null hypothesis  $H_*$  of nonexistence of relevant differences is accepted, and at the left- (right)-hand boundary of the critical region  $\mathcal{T} \cap (-\infty, C_1^*) \cup (C_2^*, \infty)$ ,  $H_*$  is rejected with probability  $\gamma_1^*$  ( $\gamma_2^*$ ). [As before, the symbol  $\mathcal{T}$  stands for the range of the test statistic  $T(\mathbf{X})$ .]

The fact that it is natural to consider any testing problem of the form (11.1) as a generalized two-sided problem in the usual sense, is corroborated by observing that even in the case of a one-parameter family being STP<sub>3</sub>, there are no UMP tests for relevant differences. Actually, the results proven in §§3.7 and 4.2 of Lehmann and Romano (2005) imply that inverting a UMP level-( $1 - \alpha$ ) test for equivalence by means of (11.2) yields a test of (11.1) which is uniformly most powerful only among all unbiased tests for the same problem. As is well known, the classical two-sided problem with a one-point null hypothesis about the parameter of a STP<sub>3</sub>-family of distributions, likewise admits only a UMPU solution whereas a test which is uniformly most powerful among all valid tests, does not exist.

Of course, exact UMPU solutions to problems of testing for relevant differences exist only for a comparably small number of settings. Often, one will have recourse to asymptotic methods to be obtained by applying the duality principle to an equivalence test of the form described in general terms in § 3.4. All that has to be done for implementing such a testing procedure is that in the corresponding asymptotic test for equivalence, the critical constant to which the suitably centered and normalized test statistic is compared must be determined by computing the square root of the *upper* 100( $1 - \alpha$ )-percentage point of a  $\chi^2$ -distribution with a single degree of freedom and noncentrality parameter  $\psi^2 = (\varepsilon_1 + \varepsilon_2)^2 / 4 \hat{\tau}_N^2$ , with  $\hat{\tau}_N^2$  denoting a weakly consistent estimator for the asymptotic variance of the test statistic. Of course, rejection of the null

hypothesis of absence of relevant differences has again to be done if the test statistic falls to the left or the right of the critical interval. Finally, another promising option for addressing problems of testing for relevant differences for which no satisfactory solutions by means of exact finite-sample constructions are available, is to adapt the objective Bayesian approach to the hypotheses formulation now under consideration. Technically, this just requires replacing in the general decision rule (2.7) of a Bayesian test for noninferiority the set  $(\theta_0 - \varepsilon, \infty)$  with  $(-\infty, \theta_0 - \varepsilon_1) \cup (\theta_0 + \varepsilon_2, \infty)$ .

---

## 11.3 Solutions to some special problems of testing for relevant differences

### 11.3.1 One-sample problem with normally distributed data of known variance

The setting to be considered briefly here is largely the same as in § 4.1 and need not be described in detail again. It suffices to recall that any testing problem involving interval hypotheses about the expected value  $\theta$ , say, of a Gaussian normal distribution with known variance  $\sigma_0^2 > 0$  from which a sample  $(X_1, \dots, X_n)$  of arbitrary size  $n$  is given, can be reduced without loss of statistical information to a pair of statements specifying that the expected value  $\tilde{\theta}$  of some normally distributed random variable  $Z$  with unit variance lies in- or outside some symmetric interval with endpoints  $\pm\tilde{\varepsilon}$ . If we denote the critical function of the optimal (UMPU) level- $\alpha$  test for

$$\tilde{H}_* : -\tilde{\varepsilon} \leq \tilde{\theta} \leq \tilde{\varepsilon} \quad \text{versus} \quad \tilde{K}_* : \tilde{\theta} < -\tilde{\varepsilon} \text{ or } \tilde{\theta} > \tilde{\varepsilon} \quad (11.4)$$

based on this  $Z$  by  $\tilde{\psi}_*(\cdot)$ , then it follows from the duality principle that

$$\tilde{\psi}_*(z) = \begin{cases} 1 & \text{for } |z| > C_{\alpha; \tilde{\varepsilon}}^* \\ 0 & \text{for } |z| \leq C_{\alpha; \tilde{\varepsilon}}^* \end{cases}, \quad (11.5)$$

where  $C_{\alpha; \tilde{\varepsilon}}^*$  denotes the unique solution of

$$\Phi(C - \tilde{\varepsilon}) - \Phi(-C - \tilde{\varepsilon}) = 1 - \alpha, \quad C > 0. \quad (11.6)$$

For computational purposes, it is more convenient to make use of the possibility of expressing  $C_{\alpha; \tilde{\varepsilon}}^*$  explicitly through the quantile function of a noncentral  $\chi^2$ -distribution with a single degree of freedom. The required expression is obtained from Equation (4.6) by replacing  $\alpha$  with its complement  $\alpha' = 1 - \alpha$  so that we can write

$$C_{\alpha; \tilde{\varepsilon}}^* = \sqrt{\chi_{1; \alpha'}^2(\tilde{\varepsilon}^2)}. \quad (11.7)$$

By means of (11.7), it is easy to determine the entries in Table 11.1 which is the relevant-differences analogue of Table 4.1. Substituting  $\alpha'$  for  $\alpha$  in the approximation formula (4.26) yields

$$C_{\alpha; \tilde{\varepsilon}}^* \approx \sqrt{n}\varepsilon - u_{1-\alpha'} = \sqrt{n}\varepsilon + u_{1-\alpha}, \quad (11.8)$$

and not surprisingly, it becomes obvious from the exact values shown in Table 11.1, that the accuracy of this approximation is almost perfect even for considerably smaller values of  $\tilde{\varepsilon}$  than required in the equivalence case.

Table 11.1 *Optimal critical constant  $C_{\alpha; \tilde{\varepsilon}}^*$  [ $\rightarrow$  (11.6), (11.7)] for testing for relevant differences at the 5% level in the one-sample case with normally distributed data of known variance.*

$\tilde{\varepsilon}$	0.	1.	2.	3.	4.
.0	1.95996	2.64615	3.64485	4.64485	5.64485
.1	1.96973	2.74544	3.74485	4.74485	5.74485
.2	1.99855	2.84511	3.84485	4.84485	5.84485
.3	2.04505	2.94496	3.94485	4.94485	5.94485
.4	2.10699	3.04490	4.04485	5.04485	6.04485
.5	2.18148	3.14487	4.14485	5.14485	6.14485
.6	2.26539	3.24486	4.24485	5.24485	6.24485
.7	2.35583	3.34486	4.34485	5.34485	6.34485
.8	2.45047	3.44485	4.44485	5.44485	6.44485
.9	2.54760	3.54485	4.54485	5.54485	6.54485

In order to illustrate the use of the critical constants computed from (11.6) or, equivalently, (11.7), let us suppose that we want to test for relevant differences of the expected value of a normal distribution with unit variance from zero specifying the indifference zone to range from  $-.25$  to  $+.25$ . If the size of the sample having been made available for that purpose is  $n = 100$  and the test is to be carried out at the usual significance level  $\alpha = .05$ , then we read from the third column of the above table that the observed value of the sample mean  $\bar{X}$  has to be compared with the critical bound  $4.14485/\sqrt{100}$ . Thus, the existence of relevant differences can be declared established at the 5%-level if and only if it turns out that  $|\bar{X}|$  exceeds the bound  $0.414485$  which is more than twice as large as the critical bound to be used in a traditional two-sided test for the same setting. The implications of this fact for the power of both tests are clear enough: As shown in Figure 11.1, introducing a relevance margin of  $\varepsilon = .25$  considerably reduces the rejection probability of the test under any true parameter value  $\theta \in \mathbb{R}$ , of course except those specific alternatives which are so extreme that even the power of the test for relevant

differences is arbitrarily close to unity. Generally, when  $\bar{X}$  relates to a sample of  $n$  observations from  $\mathcal{N}(\theta, 1)$ , the rejection probability of any test with critical region  $\{-\infty < \sqrt{n}\bar{X} < -C\} \cup \{C < \sqrt{n}\bar{X} < \infty\}$  is easily computed exactly using the formula

$$POW = \Phi(-C - \sqrt{n}\theta) + \Phi(-C + \sqrt{n}\theta) \quad (11.9)$$

The values behind both power curves displayed in Figure 11.1 were obtained in this way.

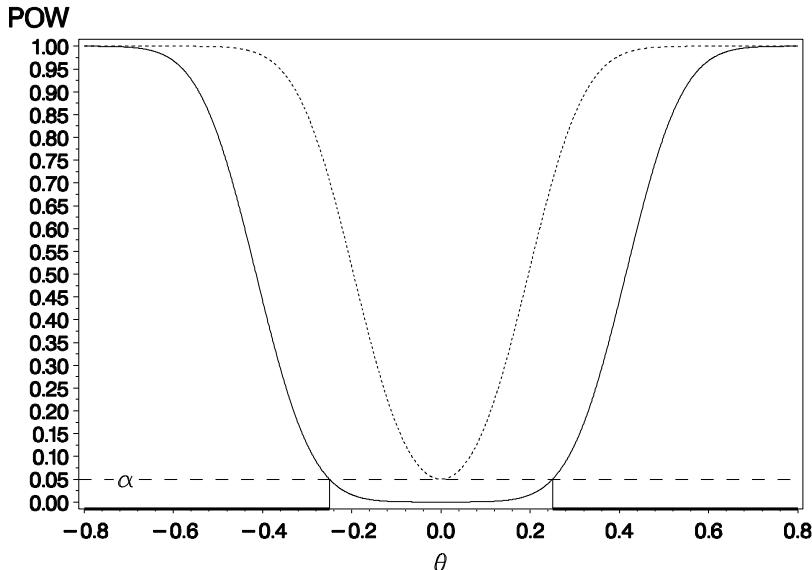


Figure 11.1 *Power function of the UMPU test at level  $\alpha = .05$  for  $|\theta| \leq .25$  vs.  $|\theta| > .25$  [solid line] and the conventional two-sided null hypothesis  $\theta = 0$  [dotted line] based on  $n = 100$  observations from  $\mathcal{N}(\theta, 1)$ . [Bold-drawn bars on horizontal coordinate axis demarcate the region of relevant differences.]*

### 11.3.2 Two-sample $t$ -test for relevant differences

The relevant-differences analogue of the equivalence testing problem considered in § 6.1 reads:

$$\begin{aligned} H_* : -\varepsilon_1 &\leq (\xi - \eta)/\sigma \leq \varepsilon_2 \quad \text{versus} \\ K_* : (\xi - \eta)/\sigma &< -\varepsilon_1 \text{ or } (\xi - \eta)/\sigma > \varepsilon_2, \end{aligned} \quad (11.10)$$

where  $\xi$  and  $\eta$  are the expected values of two normal distributions with common unknown variance  $\sigma^2 > 0$  from which independent samples  $X_1, \dots, X_m$

and  $Y_1 \dots, Y_n$  are given. The sizes  $m$  and  $n$  of both samples need not to be equal, and similarly, no restrictions on the two relevance margins  $\varepsilon_1, \varepsilon_2$  have to be made. Although the setting of the two-sample  $t$ -test involves a multiparameter family of distributions, like its equivalence counterpart, the problem (11.10) can be reduced by invariance to a pair of hypotheses relating to a single STP<sub>3</sub> family. In this reduced version, it admits a UMPU solution which is readily obtained by applying the principle of duality to the critical region given by (6.3) and (6.4). Adopting the arguments used in § 6.6 of Lehmann and Romano (2005) in the one-sample analogue of problem (11.10) with  $\varepsilon_1 = \varepsilon_2 = 0$ , it can be shown that the test obtained by composing the critical function of this UMPU test with the mapping  $(x_1, \dots, y_n) \mapsto \sqrt{mn(N-2)/N}(\bar{x} - \bar{y}) / \left\{ \sum_{i=1}^m (x_i - \bar{x})^2 + \sum_{j=1}^n (y_j - \bar{y})^2 \right\}^{1/2} \equiv t(x_1, \dots, y_n)$ , is UMP among all level- $\alpha$  tests for (11.10) which are both invariant and unbiased.

The practical implementation of this  $t$ -test for relevant differences between two homoskedastic Gaussian distributions can be accomplished with the same computational tools made available before for the two-sample  $t$ -test for equivalence. Actually, the rejection region can be written

$$\left\{ T < \tilde{C}_{\alpha'; m, n}^1(-\varepsilon_1, \varepsilon_2) \right\} \cup \left\{ T > \tilde{C}_{\alpha'; m, n}^2(-\varepsilon_1, \varepsilon_2) \right\}, \quad (11.11)$$

where the critical constants  $\tilde{C}_{\alpha'; m, n}^\nu$ ,  $\nu = 1, 2$ , are determined from Equations (6.4) with substituting  $1 - \alpha$  for  $\alpha$ . If the margins are chosen symmetrically setting  $\varepsilon_1 = \varepsilon_2 = \varepsilon$  for some  $\varepsilon > 0$ , (11.11) simplifies to

$$\left\{ |T| > \tilde{C}_{\alpha'; m, n}(\varepsilon) \right\}, \quad (11.12)$$

where  $\tilde{C}_{\alpha'; m, n}(\varepsilon)$  can be computed by determining the square root of the upper  $100\alpha$  percentage point of an  $F$ -distribution with  $1, N - 2$  degrees of freedom and noncentrality parameter  $\lambda_{nc}^2 = mn\varepsilon^2/N$ .

Clearly, the power of the test with rejection region (11.11) against any alternative  $(\xi, \eta, \sigma^2)$  with  $(\xi - \eta)/\sigma \equiv \theta \in (-\infty, -\varepsilon_1) \cup (\varepsilon_2, \infty)$  is given by

$$\begin{aligned} \beta_*(\theta) = F_{N-2; \theta\sqrt{mn/N}}^T(\tilde{C}_{\alpha'; m, n}^1(-\varepsilon_1, \varepsilon_2)) + \\ \left( 1 - F_{N-2; \theta\sqrt{mn/N}}^T(\tilde{C}_{\alpha'; m, n}^2(-\varepsilon_1, \varepsilon_2)) \right), \end{aligned} \quad (11.13)$$

where  $F_{N-2; \lambda_{nc}}^T(\cdot)$  symbolizes the cumulative distribution function of a random variable which is noncentral  $t$  with  $df = N - 2$  and noncentrality parameter  $\lambda_{nc} \in \mathbb{R}$ .

In view of the above formula, it is an easy exercise to recalculate the entries in Tables 6.1 and 6.2 for the  $t$ -test for relevant differences. Of course, the specific alternative of interest must now be chosen as some point  $\theta$  with  $|\theta| > \varepsilon$ . The entries in Table 11.3 hold under the assumption that we specify  $\theta = 2\varepsilon$ .

Table 11.2 *Critical constant  $\tilde{C}_{.95; n,n}(\varepsilon)$  of the two-sample t-test for relevant differences at level  $\alpha = 5\%$  in the case of a balanced design with  $n = 10(5)75$  and relevance margin  $\varepsilon = .10(.10).50$ .*

$n$	$\varepsilon =$				
	.10	.20	.30	.40	.50
10	2.15247	2.29572	2.50159	2.73959	2.99103
15	2.12304	2.32217	2.58970	2.88182	3.18130
20	2.12175	2.37146	2.68894	3.02417	3.36387
25	2.13028	2.42582	2.78532	3.15753	3.53289
30	2.14322	2.48048	2.87655	3.28184	3.68971
35	2.15838	2.53387	2.96262	3.39825	3.83619
40	2.17473	2.58545	3.04402	3.50791	3.97396
45	2.19172	2.63512	3.12130	3.61176	4.10433
50	2.20902	2.68291	3.19497	3.71060	4.22831
55	2.22645	2.72893	3.26545	3.80506	4.34674
60	2.24387	2.77330	3.33309	3.89566	4.46027
65	2.26122	2.81615	3.39821	3.98282	4.56946

Table 11.3 *Power against the alternative  $\theta = 2\varepsilon$  attained in the two-sample t-test for relevant differences for the sample sizes and margin specifications underlying Table 11.2.*

$n$	$\varepsilon =$				
	.10	.20	.30	.40	.50
10	.06467	.10159	.14937	.20304	.26168
15	.07233	.12512	.19137	.26744	.35179
20	.07963	.14640	.22979	.32710	.43381
25	.08663	.16616	.26609	.38314	.50815
30	.09335	.18485	.30085	.43579	.57493
35	.09983	.20276	.33431	.48513	.63441
40	.10609	.22008	.36659	.53120	.68694
45	.11216	.23691	.39774	.57404	.73301
50	.11805	.25335	.42778	.61374	.77315
55	.12378	.26945	.45672	.65041	.80792
60	.12936	.28524	.48457	.68416	.83788
65	.13482	.30075	.51133	.71513	.86358

### Example 11.1

In 2005, the German Competence Network on Dementias started an actively controlled clinical trial to evaluate the efficacy and safety of an antidementive

combination therapy (galantamine plus memantine) in subjects with mild-to-moderate stage of probable Alzheimer's disease. In the control arm of the study, galantamine was administered as monotherapy, and random assignment of patients to the two arms was done in a 1:1 manner. As primary outcome criterion for assessing the antidementive effects, the change of the ADAS-cog score after 12 months of treatment was used. As is true for any rating scale based on a set of questionnaire items, the distribution of this variable is in principle discrete. However, in view of the comparatively large number of possible values taken on by the ADAS-cog score (range:  $\{x \in \mathbb{N}_0 | x \leq 70\}$ ), when planning the trial one felt justified to rely on the standard parametric model assumptions behind the two-sample  $t$ -test. Sample-size calculation was based on the one-sided version of the test at the usual level  $\alpha = .05$  and the requirement that a true effect of size  $\theta = (\xi - \eta)/\sigma = .40$  should be detected with a power of 80% at least. Accordingly, the recruitment target was set to  $n = 78$  per group.

Suppose that the same trial is to be planned with a view to testing for relevant differences in terms of  $(\xi - \eta)/\sigma$  between the combination of both drugs and monotherapy fixing the margin at  $\varepsilon = .20$  and leaving unchanged both the alternative of interest and the requirement on the minimum power. Then it becomes obvious from the entries in the second column of Table 11.3 that the size of both samples would have to be much larger than 78. By iterating formula (11.13) on  $m = n$ , the precise value of  $n$  turns out to be as large as 310. Mainly, this increase is due to introducing a nonzero relevance margin rather than replacing a one- with a two-sided test.

### 11.3.3 Exact Fisher type test for relevant differences between two binomial distributions

Another standard setting of considerable practical importance where applying the duality principle yields an exactly valid optimal test for relevant differences is the binomial two-sample problem with the odds ratio as the parameter of interest. Denoting the critical function of the equivalence test considered in § 6.6.4 for any fixed value  $s \in \{0, 1, \dots, m+n\}$  of the column total  $S$  belonging to the contingency table summarizing the data set under analysis [see Table 6.20] by  $\phi(\cdot, s)$  and observing that this test has to be performed at (nominal) level  $\alpha' = 1 - \alpha$  in the present context, we can write:

$$1 - \phi(x, s) = \begin{cases} 1 & \text{if } X < C_{1; \alpha'}^{m,n}(s; \varrho_1, \varrho_2) \\ & \quad \text{or } X > C_{2; \alpha'}^{m,n}(s; \varrho_1, \varrho_2) \\ 1 - \gamma_{1; \alpha'}^{m,n}(s; \varrho_1, \varrho_2) & \text{if } X = C_{1; \alpha'}^{m,n}(s; \varrho_1, \varrho_2) \\ 1 - \gamma_{2; \alpha'}^{m,n}(s; \varrho_1, \varrho_2) & \text{if } X = C_{2; \alpha'}^{m,n}(s; \varrho_1, \varrho_2) \\ 0 & \text{if } C_{1; \alpha'}^{m,n}(s; \varrho_1, \varrho_2) < \\ & \quad X < C_{2; \alpha'}^{m,n}(s; \varrho_1, \varrho_2) \end{cases}. \quad (11.14)$$

All critical constants appearing in the definition of this decision rule have to be determined by solving the equations (6.98) with  $\alpha'$  being substituted for  $\alpha$ . The algorithm provided in Chapter 6 for that purpose works for any choice of the nominal level so that the program **bi2st** can also be used for testing for relevant differences from unity of the odds ratio associated with the two binomial distributions under comparison.

Since the construction carried out in § 6.6.4 ensures in particular that  $\phi(X, S)$  is unbiased at level  $\alpha'$  for  $H : \varrho \notin (\varrho_1, \varrho_2)$  versus  $K : \varrho \in (\varrho_1, \varrho_2)$ ,  $1 - \phi(X, S)$  gives a test of

$$H_* : \varrho \in [\varrho_1, \varrho_2] \quad \text{vs.} \quad K_* : \varrho < \varrho_1 \text{ or } \varrho > \varrho_2 \quad (11.15)$$

whose rejection probability under  $H_*$  does not exceed  $\alpha$ . From the results proven in Lehmann and Romano (2005, § 4.4), it follows that it is unbiased in turn and maximizes the power uniformly among all unbiased level- $\alpha$  tests for (11.15). The practical application of the decision rule given by (11.14) is best illustrated by another concrete example.

### *Example 11.2*

One of the polymorphisms which were investigated in an association study on the genetic background of psoriasis was a SNP denoted TNFA-238 which belongs to a gene related to the tumor necrosis factor alpha. The proportion of carriers of at least one G-allele at the corresponding locus was found to differ considerably between cases and controls. The pertinent part of the data published by Reich et al. (2002) on the results of this study yield the following  $2 \times 2$  contingency table:

Table 11.4 *Frequencies of carriers of at least one G-allele at a SNP located in a gene encoding the tumor necrosis factor TNF- $\alpha$  obtained in a genetic association study on psoriasis. (Data from Reich et al., 2002.)*

Group	Presence of G-allele		
	+	-	$\Sigma$
Control	29 (8.41%)	316 (91.59%)	345 (100.0%)
Psoriasis	53 (22.94%)	178 (77.06%)	231 (100.0%)
$\Sigma$	82	494	576

Considering the association between the SNP and the disease status as relevant only if the true odds ratio satisfies either  $\varrho < 2/3$  or  $\varrho > 3/2$ , the basic step to be taken for determining the critical constants for the test (11.14) at level  $\alpha = .05$  consists of running program `bi2st` with .95, 345, 231, 82, 2/3, and 3/2 as the values assigned to its five arguments. The results it then returns are:

$$C_1 = 35, C_2 = 62, \gamma_1 = .2891, \gamma_2 = .6703.$$

According to (11.14), this means that for the values of the marginal frequencies shown in Table 11.4, the null hypothesis of irrelevant deviations of the odds ratio from unity has to be rejected if the observed count  $X$  of carriers of a G-allele among the controls is less than 35 or larger than 62. In the exact UMPU test at level  $\alpha = .05$ , randomized decisions in favor of  $K_*$  have to be taken with probabilities  $1 - \gamma_1 = .7109$  (for  $X = 35$ ) and  $1 - \gamma_2 = .3297$  (for  $X = 62$ ), and for  $36 \leq X \leq 61$ , the null hypothesis  $H_*$  must be accepted. Since the realized value of  $X$  is much smaller than the lower critical bound, existence of a relevant association between TNFA-238 and the psoriasis condition can be declared established with the data presented by Reich et al. (2002).

Computing the power of the exact randomized UMPU test (11.14) against arbitrary alternatives  $(p_1, p_2)$  requires no more than running the program made available in § 6.6.4 for calculating the power of the exact Fisher type test for equivalence. Of course, both the significance level and the resulting power has to be replaced by its complement with respect to 1. The power of the nonrandomized version of (11.14) cannot be obtained in exactly the same way since before inverting the corresponding equivalence test, the boundaries  $C_{\nu; \alpha'}^{m, n}(s; \varrho_1, \varrho_2)$ ,  $\nu = 1, 2$ , must be included in rather than excluded from its critical interval. In order to accommodate this slight modification, a separate program named `bi2rlv1` is made available in the **WKTSEQ2 Source Code Package**. The appropriate input to the latter is the target value of  $\alpha$  itself rather than  $\alpha'$ , and the power of both versions of the test for relevant differences is output directly rather than as the complement of the correct value. Using this tool for computing the power which the test for relevant deviations of the odds ratio from 1 provides under the specifications made in the above example with the observed proportions .0841 and .2294 as the actual values of  $p_1$  and  $p_2$ , yields  $POW = .9352$  and  $POWNR = .9194$  for the exact and the nonrandomized version of (11.14), respectively.

Adopting the computational tools made available in Chapter 6 for sample size determination with the exact Fisher type test for equivalence, for the case of relevant differences entails more far reaching modifications as compared with the computation of power. Actually, applying the algorithm behind the program `bi2aeq2` with a  $(p_1, p_2)$  falling outside the indifference zone specified by the null hypothesis of (11.15) would not work since, in the equivalence framework, the rejection probability of the test under such a parameter configuration has the meaning of a type-I error risk and is thus decreasing rather than increasing in the sample size(s). The result of adapting the algorithm with a view to this basic fact is the source code to be found in the **WKTSEQ2**

**Source Code Package** under the base filename `bi2rlv2`. Table 11.5 shows a selection of results obtained by means of this procedure for the exact Fisher type test for relevant differences at level  $\alpha = .05$ , relevance limits  $\varrho_1 = 2/3$ ,  $\varrho_2 = 3/2$ , and target power 80%. All parameter points  $(p_1, p_2)$  it covers lie on the same level curve defined by  $\varrho(p_1, p_2) = 7/3$  in the parameter subspace corresponding to  $K_*$ .

Table 11.5 *Sample sizes required in the exact Fisher type test for relevant differences at level  $\alpha = .05$  to maintain power 80% against selected alternatives  $p_1 = p_2 = p_*$  on the contour  $p_1/(1 - p_1) = \varrho^* p_2/(1 - p_2)$  with  $\varrho^* = 7/3$  for  $\lambda \equiv m/n \in \{.50, 1.0, 2.0\}$  and relevance limits  $\varrho_1 = 2/3$ ,  $\varrho_2 = 3/2$ . [For the motivation behind this choice and the specification of  $\varrho^*$  see § 1.7.]*

$p_1$	$p_2$	$\lambda$	$m$	$n$	$N = m + n$
.205882	.10	0.5	373	746	1119
"	"	1.0	525	525	1050
"	"	2.0	828	414	1242
.437500	.25	0.5	214	428	642
"	"	1.0	292	292	584
"	"	2.0	448	224	672
.700000	.50	0.5	208	416	624
"	"	1.0	273	273	546
"	"	2.0	404	202	604
.875000	.75	0.5	348	696	1044
"	"	1.0	443	443	886
"	"	2.0	634	317	951
.9296875	.85	0.5	556	1112	1668
"	"	1.0	701	701	1402
"	"	2.0	990	495	1395

As can be expected by intuition, the exact Fisher test for relevant differences shares with its equivalence counterpart the property that the required sample sizes are the larger the extremer the baseline responder rate  $p_2$  happens to be. Furthermore, the results shown in the above table confirm once more the general rule that, in terms of sample sizes, unbalanced designs are less cost-effective than sampling schemes with 1:1 allocation of the observational units to the treatment groups under comparison.

Finally, it is worth mentioning that the technique used in § 6.6.5 to construct an improved nonrandomized version of the optimal conditional test works in

the present context as well. The steps to be taken for adapting the algorithm required for determining maximally increased nominal significance levels for the nonrandomized counterpart of the UMPU level- $\alpha$  test for the problem (11.15) are largely self-explanatory so that no further details are given here.

# Appendix A

---

## *Basic theoretical results*

---

### A.1 UMP tests for equivalence problems in STP<sub>3</sub> families

**A.1.1 Definition.** Let  $(p_\theta(\cdot))_{\theta \in \Theta}$  be a family of real-valued functions with common domain  $\mathcal{X}$ . Suppose that both  $\mathcal{X}$  and the parameter space  $\Theta$  is some simply ordered set. Moreover, for arbitrary  $n = 1, 2, \dots$ , let  $\mathcal{X}^{(n)}$  and  $\Theta^{(n)}$  denote the set of all increasingly ordered  $n$ -tuples of pairwise different points in  $\mathcal{X}$  and  $\Theta$ , respectively, and define

$$\Delta_n \begin{pmatrix} x_1, \dots, x_n \\ \theta_1, \dots, \theta_n \end{pmatrix} \equiv \det \begin{pmatrix} p_{\theta_1}(x_1) & \dots & p_{\theta_1}(x_n) \\ \vdots & \ddots & \vdots \\ p_{\theta_n}(x_1) & \dots & p_{\theta_n}(x_n) \end{pmatrix} \quad (\text{A.1})$$

for arbitrary  $(x_1, \dots, x_n) \in \mathcal{X}^{(n)}$ ,  $(\theta_1, \dots, \theta_n) \in \Theta^{(n)}$ . Then, the family  $(p_\theta(\cdot))_{\theta \in \Theta}$  is said to be strictly totally positive of order  $r \geq 1$  (abbreviated to STP<sub>r</sub>) if for each  $n = 1, \dots, r$  we have that

$$\Delta_n \begin{pmatrix} x_1, \dots, x_n \\ \theta_1, \dots, \theta_n \end{pmatrix} > 0 \quad \forall ((x_1, \dots, x_n), (\theta_1, \dots, \theta_n)) \in \mathcal{X}^{(n)} \times \Theta^{(n)}. \quad (\text{A.2})$$

If (A.2) holds true for every  $n \in \mathbb{N}$ , the family  $(p_\theta(\cdot))_{\theta \in \Theta}$  is called strictly totally positive of order  $\infty$  (STP <sub>$\infty$</sub> ).

### Important special STP <sub>$\infty$</sub> families of probability densities on $\mathbb{R}$

**A.1.2 Lemma.** Let  $\mathcal{X}$  be a Borelian set in  $\mathbb{R}^1$ ,  $\mu(\cdot)$  a  $\sigma$ -finite measure on  $\mathcal{X}$ , and  $(P_\theta)_{\theta \in \Theta}$  a family of probability distributions on  $\mathcal{X}$  such that  $\Theta$  is a nondegenerate interval on the real line and for each  $\theta \in \Theta$ , a  $\mu$ -density of  $P_\theta$  is given by

$$p_\theta(x) = c(\theta) \exp\{\theta x\} h(x), \quad x \in \mathcal{X}, \quad (\text{A.3})$$

with  $c(\cdot) : \Theta \rightarrow \mathbb{R}_+$  and  $h$  as a Borel-measurable transformation from  $\mathcal{X}$  to  $\mathbb{R}_+$  [ $\leftrightarrow$  one-parameter exponential family in  $x$  and  $\theta$ ]. Then,  $(p_\theta(\cdot))_{\theta \in \Theta}$  is STP <sub>$\infty$</sub> .

*Proof.* → Karlin (1968, p. 19). ■

**A.1.3 Lemma.** For arbitrary  $\nu \in \mathbb{N}$  and  $\theta \in \mathbb{R}$  let  $p_\theta(\cdot; \nu)$  be the density (with respect to Lebesgue measure) of a noncentral  $t$ -distribution with  $\nu$  degrees of freedom and noncentrality parameter  $\theta$ . Then it follows that  $(p_\theta(\cdot; \nu))_{\theta \in \mathbb{R}}$  is STP $_\infty$ .

*Proof.* → Karlin (1968, § 4(ii)). ■

**A.1.4 Lemma.** For any  $(\nu_1, \nu_2) \in \mathbb{N}^2$ , let  $V$  denote a random variable whose distribution is central  $F$  with  $\nu_1$  and  $\nu_2$  degrees of freedom. Furthermore, let  $h_\varrho(\cdot)$  be the density function of  $\varrho \cdot V$  where  $\varrho$  denotes any positive real number. Then, the family  $(h_\varrho(\cdot))_{\varrho > 0}$  (occasionally called family of stretched  $F$ -densities with  $\nu_1, \nu_2$  degrees of freedom – see, e.g., Witting, 1985, p. 217) is STP $_\infty$ .

*Proof.* Let  $(U_1, U_2)$  be an independent pair of random variables such that  $\nu_1 U_1$  and  $\nu_2 U_2$  has a central  $\chi^2$ -distribution with  $\nu_1$  and  $\nu_2$  degrees of freedom, respectively. Moreover, define for any  $(a, b) \in \mathbb{R}_+^2$   $\tilde{g}_{\nu_1}(a; b) \equiv b g_{\nu_1}(ab)$ ,  $\tilde{g}_{\nu_2}(a; b) \equiv b g_{\nu_2}(ab)$  with  $g_{\nu_1}(\cdot)$  and  $g_{\nu_2}(\cdot)$  as the density function of  $U_1$  and  $U_2$ , respectively. Clearly, for any  $b > 0$ , the density function of  $b^{-1} \cdot U_1$  and  $b^{-1} \cdot U_2$  is then given by  $\tilde{g}_{\nu_1}(\cdot; b)$  and  $\tilde{g}_{\nu_2}(\cdot; b)$  respectively.

By definition of  $V$ , we have  $\hat{\varrho} V \stackrel{d}{=} U_1 / (\varrho^{-1} \cdot U_2)$ . Hence, applying the well-known formula for the density of the product of two independent random variables of the absolutely continuous type (see, e.g., Kingman and Taylor, 1973, p. 299, Ex. 11.5.3), we can write for any  $z \in \mathbb{R}_+$ :

$$h_\varrho(z) = \int_0^\infty y g_{\nu_1}(yz) \tilde{g}_{\nu_2}(y; \varrho) dy = \int_0^\infty \tilde{g}_{\nu_1}(z; y) \tilde{g}_{\nu_2}(y; \varrho) dy. \quad (\text{A.4})$$

As shown by Witting (1985, Theorem 2.33 a) the families  $(\tilde{g}_{\nu_1}(\cdot; y))_{y > 0}$  and  $(\tilde{g}_{\nu_2}(\cdot; \varrho))_{\varrho > 0}$  both exhibit the structure assumed in A.1.2 and are thus STP $_\infty$ . Hence, we can apply the composition theorem for pairs of strictly (totally) positive families (Karlin, 1968, p. 17) with  $X = Z = Y = (0, \infty)$ ,  $r = s = 1$ ,  $K(z, y) = \tilde{g}_{\nu_1}(z; y)$ ,  $L(y, \varrho) = \tilde{g}_{\nu_2}(y; \varrho)$ ,  $\sigma = \lambda|_{(0, \infty) \times (0, \infty)}$  yielding the proposed result. ■

## Existence and basic properties of UMP tests for equivalence in $STP_3$ families of distributions

**A.1.5 Theorem.** Let  $\alpha$  be an arbitrary real number in  $(0, 1)$ ,  $\mathcal{X}$  a nondegenerate interval in  $\mathbb{R}$ ,  $\mathcal{B}_{\mathcal{X}}$  the Borelian  $\sigma$ -algebra on  $\mathcal{X}$ ,  $\mu$  a  $\sigma$ -finite measure on  $\mathcal{B}_{\mathcal{X}}$  whose support contains at least two different points, and  $(p_{\theta}(\cdot))_{\theta \in \Theta}$  an  $STP_3$  family of  $\mu$ -densities on  $\mathcal{X}$  with  $\Theta$  exhibiting the form of some nondegenerate interval on the real line as well. Assume further that the function  $(x, \theta) \mapsto p_{\theta}(x)$  is continuous in both of its arguments, and that some fixed pair  $(\theta_1, \theta_2)$  of points in  $\Theta$  with  $\theta_1 < \theta_2$  is given. Then, the following statements hold true:

- (i) For the problem  $H : \theta \in \Theta \setminus (\theta_1, \theta_2)$  versus  $K : \theta \in (\theta_1, \theta_2)$  there exists a UMP level- $\alpha$  test  $\phi : \mathcal{X} \rightarrow [0, 1]$  given by

$$\phi(x) = \begin{cases} 1 & \text{for } x \in (C_1, C_2) \\ \gamma_i & \text{for } x = C_i, i = 1, 2 \\ 0 & \text{for } x \in \mathcal{X} \setminus [C_1, C_2] \end{cases} \quad (\text{A.5})$$

where  $C_i \in \mathcal{X}, i = 1, 2, C_1 \leq C_2$  and

$$E_{\theta_i} \phi(X) \equiv \int \phi p_{\theta_i} d\mu = \alpha \text{ for } i = 1, 2. \quad (\text{A.6})$$

- (ii) This  $\phi$  minimizes the rejection probability  $E_{\theta} \phi(X)$  uniformly in  $\theta \in \Theta \setminus [\theta_1, \theta_2]$  among all tests which satisfy (A.6).
- (iii) If the cardinality of the set which the dominating measure  $\mu$  is concentrated upon, is greater than 2, then there exists a point  $\theta_o \in (\theta_1, \theta_2)$  such that the power function  $\theta \mapsto E_{\theta} \phi(X)$  of the UMP test is monotonically increasing and decreasing on  $(-\infty, \theta_o] \cap \Theta$  and  $[\theta_o, \infty) \cap \Theta$ , respectively.

*Proof.* A proof a weaker version of the theorem restricted to the special case that the underlying family of distributions exhibits the form (A.3), can be found in Lehmann and Romano (2005, Ch. 3.7). Proving the same result in full generality as repeatedly used in the present book, is left to the reader of Lehmann's book as an exercise worked out by Kallenberg (1984, pp. 54–58). ■

### Simplified computation of the UMP test for equivalence under symmetry restrictions

**A.1.6 Lemma.** Under the assumptions and notational conventions of Theorem A.1.5, let the limits of the equivalence range for  $\theta$  be chosen as  $\theta_1 = -\varepsilon$ ,  $\theta_2 = \varepsilon$ , with arbitrarily fixed  $\varepsilon > 0$ . Suppose further that under  $\theta = \varepsilon$  the random variable which the hypotheses  $H$  and  $K$  refer to, has the same distribution as  $-X$  under  $\theta = -\varepsilon$ . Then, a UMP level- $\alpha$  test for  $H : \theta \in \Theta \setminus (-\varepsilon, \varepsilon)$  versus  $K : \theta \in (-\varepsilon, \varepsilon)$  is given by

$$\phi(x) = \begin{cases} 1 & \text{if } |x| < C \\ \gamma & \text{if } |x| = C \\ 0 & \text{if } |x| > C \end{cases} \quad (\text{A.7})$$

where

$$C = \max \left\{ x \in [0, \infty) \mid P_\varepsilon[|X| < x] \leq \alpha \right\} \quad (\text{A.8})$$

and

$$\gamma = \begin{cases} \frac{\alpha - P_\varepsilon[|X| < C]}{P_\varepsilon[|X| = C]} & \text{for } P_\varepsilon[|X| = C] > 0 \\ 0 & \text{for } P_\varepsilon[|X| = C] = 0 \end{cases}. \quad (\text{A.9})$$

*Proof.* In view of (A.7) and (A.6), it suffices to show that

$$P_{-\varepsilon}[|X| < C] + \gamma P_{-\varepsilon}[|X| = C] = \alpha = P_\varepsilon[|X| < C] + \gamma P_\varepsilon[|X| = C].$$

The validity of the second of the above equalities is obvious since (A.8) and (A.9) simply give the solution to the equation  $P_\varepsilon[|X| < x] + q P_\varepsilon[|X| = x] = \alpha$ ,  $0 \leq x < \infty$ ,  $0 \leq q < 1$ . Furthermore, the assumption of the equality of the distribution of  $-X$  with respect to  $P_{-\varepsilon}$  and that of  $X$  with respect to  $P_\varepsilon$  trivially implies that we have  $P_{-\varepsilon}[|X| \in B_+] = P_\varepsilon[|X| \in B_+] \forall B_+ \in \mathcal{B}_{[0, \infty)}$ . Putting  $B_+ = [0, C)$  and  $B_+ = \{C\}$ , respectively, this yields  $P_{-\varepsilon}[|X| < C] + \gamma P_{-\varepsilon}[|X| = C] = P_\varepsilon[|X| < C] + \gamma P_\varepsilon[|X| = C]$ , as required. ■

**A.1.7 Corollary.** Suppose the assumptions of Theorem A.1.5 are satisfied, and there exist isotonic continuous mappings  $T : \mathcal{X} \rightarrow \mathbb{R}$ ,  $h : \Theta \rightarrow \mathbb{R}$  such that  $Z \equiv h(\Theta)$  is an interval symmetric about zero and the distribution of  $-T(X)$  under  $h^{-1}(-\zeta)$  is the same as that of  $T(X)$  under  $h^{-1}(\zeta)$  for each  $\zeta \in Z$ . Furthermore, let  $\theta_1 = h^{-1}(-\zeta_0)$ ,  $\theta_2 = h^{-1}(\zeta_0)$  for some  $\zeta_0 \in Z \cap \mathbb{R}_+$ , and define

$$t^* = \max \left\{ t \in [0, \infty) \mid P_{\theta_2} [|T(X)| < t] \leq \alpha \right\}, \quad (\text{A.10})$$

$$C_1 = T^{-1}(-t^*), \quad C_2 = T^{-1}(t^*), \quad (\text{A.11})$$

$$\gamma^* = \begin{cases} \frac{\alpha - P_{\theta_2} [|T(X)| < t^*]}{P_{\theta_2} [|T(X)| = t^*]} & \text{for } P_{\theta_2} [|T(X)| = t^*] > 0 \\ 0 & \text{for } P_{\theta_2} [|T(X)| = t^*] = 0 \end{cases}, \quad (\text{A.12})$$

$$\phi(x) = \begin{cases} 1 & C_1 < x < C_2 \\ \gamma^* & \text{if } x \in \{C_1, C_2\} \\ 0 & x \in \mathcal{X} \setminus [C_1, C_2] \end{cases}. \quad (\text{A.13})$$

Then,  $\phi$  is a UMP level- $\alpha$  test for  $H : \theta \in \Theta \setminus (\theta_1, \theta_2)$  versus  $K : \theta \in (\theta_1, \theta_2)$ . If we consider the rejection probability  $\phi$  as a function of  $\zeta = h(\theta)$ , the latter is symmetric about  $\zeta = 0$  and increases (decreases) on the left (right) of zero.

*Proof.* Continuity and isotony of the mapping  $h$  and injectivity of the transformation  $T$  imply that the problem  $H$  versus  $K$  originally put forward is the same as that of testing  $H^* : \zeta \in \mathbb{Z} \setminus (-\zeta_0, \zeta_0)$  versus  $K^* : \zeta \in (-\zeta_0, \zeta_0)$  in the family of distributions of  $T(X)$ .

For brevity, let us denote  $P_{h^{-1}(\zeta)}^{T(X)}$ , i.e., the distribution of  $T(X)$  under  $\theta = h^{-1}(\zeta)$ , by  $P_\zeta^*$ , and the image  $T(\mu)$  of the basic measure  $\mu$  under the transformation  $T$  by  $\mu^*$ . Then, it is readily verified by means of the transformation theorem for  $\mu$ -integrals (see, e.g., Halmos, 1974, p. 163) that a density of  $P_\zeta^*$  with respect to  $\mu^*$  is given by  $p_\zeta^* = p_{h^{-1}(\zeta)} \circ T^{-1}$ . Since, by assumption, both  $T$  and  $h$  are continuous and isotonic, it is clear that  $(p_\zeta^*)_{\zeta \in \mathbb{Z}}$  shares with  $(p_\theta)_{\theta \in \Theta}$  the property of being an STP<sub>3</sub> family satisfying the continuity conditions of Theorem A.1.5. Thus, all prerequisites are fulfilled for an application of the above Lemma A.1.6 with  $(T(X), \zeta, \zeta_0)$  instead of  $(X, \theta, \varepsilon)$ . Accordingly, we can conclude that a UMP level- $\alpha$  test  $\phi^*$  for  $H^*$  versus  $K^*$  is obtained by computing the critical constants  $t^*, \gamma^*$  from

$$t^* = \max \left\{ t \in [0, \infty) \mid P_{\zeta_0}^* [|T(X)| < t] \leq \alpha \right\}, \quad (\text{A.8}^*)$$

$$\gamma^* = \begin{cases} \frac{\alpha - P_{\zeta_0}^* [|T(X)| < t^*]}{P_{\zeta_0}^* [|T(X)| = t^*]} & \text{for } P_{\zeta_0}^* [|T(X)| = t^*] > 0 \\ 0 & \text{for } P_{\zeta_0}^* [|T(X)| = t^*] = 0 \end{cases}, \quad (\text{A.9}^*)$$

and defining  $\phi^*$  by

$$\phi^*(t) = \begin{cases} 1 & |t| < t^* \\ \gamma^* & \text{for } |t| = t^* \\ 0 & |t| > t^* \end{cases}. \quad (\text{A.7}^*)$$

In view of  $P_{\zeta_0}^* \equiv P_{h^{-1}(\zeta_0)}^{T(X)} = P_{\theta_2}^{T(X)}$ , we obviously have (A.8\*)  $\Leftrightarrow$  (A.10) and (A.9\*)  $\Leftrightarrow$  (A.12). Moreover, the assumptions made on the transformation  $T$  ensure that there holds  $\phi(x) = \phi^*(T(x)) \ \forall x \in \mathcal{X}$  with  $\phi$  and  $(C_1, C_2)$  defined as in (A.13) and (A.11), respectively. Hence  $\phi$  is obtained by representing the UMP level- $\alpha$  test  $\phi^*$  for  $H^*$  vs.  $K^*$  as a function of  $x$  rather than  $t$  which in view of the equivalence of the testing problems  $(H, K)$  and  $(H^*, K^*)$  implies that  $\phi$  is indeed UMP among all level- $\alpha$  tests for the former.

Thus, it only remains to show that the function  $\zeta \mapsto E_{h^{-1}(\zeta)}\phi(X)$  is symmetric about zero and exhibits the monotonicity properties stated in the second part of the corollary. By  $\phi = \phi^* \circ T$ ,  $P_\zeta^* = T(P_{h^{-1}(\zeta)})$ , we may write for arbitrary  $\zeta \in \mathbf{Z}$ :

$$E_{h^{-1}(\zeta)}\phi(X) \equiv \int \phi dP_{h^{-1}(\zeta)} = \int \phi^* \circ T dP_{h^{-1}(\zeta)} = \int \phi^* dP_\zeta^*. \quad (\text{A.14})$$

By construction,  $\phi^*$  is a UMP level- $\alpha$  test for  $H^* : \zeta \in \mathbf{Z} \setminus (-\zeta_0, \zeta_0)$  versus  $K^* : \zeta \in (-\zeta_0, \zeta_0)$  with  $\zeta$  as the parameter of a family of distributions which satisfies the conditions of Theorem A.1.5. By part (iii) of A.1.5, it thus follows that the interval  $(-\zeta_0, \zeta_0)$  contains some point  $\zeta_0^*$  such that the function  $\zeta \mapsto \int \phi^* dP_\zeta^*$  increases strictly on  $\mathbf{Z} \cap (-\infty, \zeta_0^*]$  and  $\mathbf{Z} \cap [\zeta_0^*, \infty)$ , respectively.

From (A.7\*), it is obvious that we may write  $\phi^*(T(x)) = \psi^*(|T(x)|) \ \forall x \in \mathcal{X}$  with  $\psi^* \equiv I_{[0, C^*]} + \gamma^* I_{\{C^*\}}$ . Hence, we can replace the middle of the equations (A.14) with

$$E_{h^{-1}(\zeta)}\phi(X) = \int \psi^* d(|T|(P_{h^{-1}(\zeta)})) . \quad (\text{A.15})$$

The assumption  $(-T)(P_{h^{-1}(-\zeta)}) = T(P_{h^{-1}(\zeta)})$  clearly implies that the measure with respect to which the integral on the right-hand side of (A.15) has to be taken, coincides with  $|T|(P_{h^{-1}(-\zeta)})$ , so that we have  $E_{h^{-1}(-\zeta)}\phi(X) = E_{h^{-1}(\zeta)}\phi(X) \ \forall \zeta \in \mathbf{Z}$  which establishes the asserted symmetry of the power function  $\zeta \mapsto E_{h^{-1}(\zeta)}\phi(X)$ . But a function which is symmetric about zero can only attain a strict global maximum, if the latter is attained at 0. ■

---

## A.2 UMPU equivalence tests in multiparameter exponential families

**A.2.1 Definition.** Let  $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$  be a measurable space,  $\mu$  a  $\sigma$ -finite measure on  $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ ,  $\Theta \subseteq \mathbb{R}$ ,  $\Omega \subseteq \mathbb{R}^k$ ,  $k \in \mathbb{N}$ . Suppose further that for each  $(\theta, \vartheta) = (\theta, \vartheta_1, \dots, \vartheta_k) \in \Theta \times \Omega$ ,  $P_{\theta, \vartheta}$  is a probability distribution on  $\mathbf{X}$  having a  $\mu$ -density of the form

$$p_{\theta, \vartheta}(x) = c(\theta, \vartheta) \exp \left\{ \theta T(x) + \sum_{j=1}^k \vartheta_j S_j(x) \right\}, \quad x \in \mathcal{X}, \quad (\text{A.16})$$

with  $c(\theta, \vartheta) \in \mathbb{R}_+$   $\forall (\theta, \vartheta)$  and  $T(\cdot), S_1(\cdot), \dots, S_k(\cdot)$  as  $\mathcal{B}_{\mathcal{X}}$ -measurable transformations from  $\mathcal{X}$  to  $\mathbb{R}$ . Then,  $(P_{\theta, \vartheta})_{(\theta, \vartheta) \in \Theta \times \Omega}$  is called a  $(k+1)$ -parameter exponential family in  $(\theta, \vartheta_1, \dots, \vartheta_k)$  and  $(T, S_1, \dots, S_k)$ .

**A.2.2 Theorem.** In the situation described by the above definition, let  $\Theta$  and  $\Omega$  be a nondegenerate interval in  $\mathbb{R}^1$  and  $\mathbb{R}^k$ , respectively, and  $(\theta_1, \theta_2)$  a pair of points in  $\Theta$  satisfying  $\theta_1 < \theta_2$ . Then, at any level  $\alpha \in (0, 1)$ , there exists a uniformly most powerful unbiased (UMPU) test  $x \mapsto \phi(T(x), \mathbf{S}(x))$  for

$$H : (\theta, \vartheta) \in (\Theta \setminus (\theta_1, \theta_2)) \times \Omega \quad \text{versus} \quad K : (\theta, \vartheta) \in (\theta_1, \theta_2) \times \Omega. \quad (\text{A.17})$$

Given the value  $\mathbf{s}$  of the statistic  $\mathbf{S} \equiv (S_1, \dots, S_k) : \mathcal{X} \rightarrow \mathbb{R}^k$ ,  $\phi(\cdot, \mathbf{s}) : T(\mathcal{X}) \rightarrow [0, 1]$  is obtained by constructing by means of Theorem A.1.5 a UMP level- $\alpha$  test for the problem  $H_{\mathbf{s}} : \theta \in \Theta \setminus (\theta_1, \theta_2)$  versus  $K_{\mathbf{s}} : \theta \in (\theta_1, \theta_2)$  concerning the family  $(P_{\theta}^{T(X)|\mathbf{s}})_{\theta \in \Theta}$  of conditional distributions of  $T(X)$  given  $\{\mathbf{S} = \mathbf{s}\}$ .

*Proof.* See Lehmann and Romano (2005, Ch. 4.4). ■

### A.3 A sufficient condition for the asymptotic validity of tests for equivalence

**A.3.1 Definition.** For each  $N \in \mathbb{N}$ , let  $X^{(N)} = (X_1, \dots, X_N)$  be a vector of  $p$ -dimensional ( $p \in \mathbb{N}$ ) random variables on a common probability space  $(\Omega, \mathcal{A}, P)$ . Suppose the distribution of  $X^{(N)}$  is determined by a fixed number  $k \geq 1$  (independent of  $N$ ) of  $p$ -dimensional distribution functions  $F_1, \dots, F_k$  in the following way:

$$\begin{aligned} X_1, \dots, X_{n_1(N)} &\sim F_1 \\ X_{n_1(N)+1}, \dots, X_{n_2(N)} &\sim F_2 \\ &\vdots && \vdots && \vdots \\ X_{n_{k-1}(N)+1}, \dots, X_{n_k(N)} &\sim F_k, \end{aligned}$$

where  $n_1(N) + \dots + n_k(N) = N$ . Furthermore, let  $\theta(\cdot)$  denote some functional on the underlying space of cdf's,  $\mathcal{H}$  the class of all vectors  $\mathbf{F} = (F_1, \dots, F_k)$  of cdf's satisfying the null hypotheses  $H : \theta(\mathbf{F}) \in \Theta \setminus (\theta_1, \theta_2)$ , and  $P_{\mathbf{F}}^{(N)}$  the common distribution of  $(X_1, \dots, X_N)$  under  $\mathbf{F}$ .

A test  $\phi_N : \mathbb{R}^{pN} \rightarrow [0, 1]$  is said to be asymptotically valid at level  $\alpha \in (0, 1)$  for  $H$  if we have

$$\limsup_{N \rightarrow \infty} E_{\mathbf{F}}(\phi_N) \leq \alpha \quad \forall \mathbf{F} \in \mathcal{H} \quad (\text{A.18})$$

where

$$E_{\mathbf{F}}^{(N)}(\phi_N) \equiv \int \phi_N(x_1, \dots, x_N) dP_{\mathbf{F}}^{(N)}(x_1, \dots, x_N). \quad (\text{A.19})$$

**A.3.2 Lemma.** For any  $s > 0$  denote

$$c_{\alpha}^{\theta_1, \theta_2}(s) = C_{\alpha}(s(\theta_2 - \theta_1)/2) \quad (\text{A.20})$$

with  $C_{\alpha}(\psi)$  as the square root of the  $\alpha$ -quantile of a  $\chi^2$ -distribution with a single degree of freedom and noncentrality parameter  $\psi^2 > 0$ . Then

$$c_{\alpha}^{\theta_1, \theta_2}(s) - s(\theta_2 - \theta_1)/2 \rightarrow u_{\alpha} \quad \text{as } s \rightarrow \infty \quad (\text{A.21})$$

provided that  $u_{\alpha} = \Phi^{-1}(\alpha)$ .

*Proof.* → Wellek (1996, p. 707). ■

**A.3.3 Lemma.** For each natural number  $N$  let  $S_N$  be a positive random

variable on some probability space  $(\Omega, \mathcal{A}, P_N)$  such that  $S_N \xrightarrow{P_N} c$  as  $N \rightarrow \infty$  for some  $c \in (0, \infty)$ ,  $q(\cdot)$  a function from  $(0, \infty)$  into  $\mathbb{R}$  converging to some  $y^* \in \mathbb{R}$  as  $s \rightarrow \infty$ , and  $(a_N)$  a sequence of positive real numbers with  $a_N \rightarrow +\infty$  as  $N \rightarrow \infty$ . Then, we can conclude that

$$q(a_N S_N) \xrightarrow{P_N} y^* \quad \text{as } N \rightarrow \infty . \quad (\text{A.22})$$

*Proof.* → Wellek (1996, p. 707). ■

**A.3.4 Theorem.** Let  $(X^{(N)})_{N \in \mathbb{N}}$  be as specified in A.3.1 with  $\mathcal{F}$  as the underlying class of  $k$ -tuples of distribution functions, and assume that for each  $N \in \mathbb{N}$ , a test statistic  $T_N$  on the sample space of  $X^{(N)}$  is given such that we have

$$\frac{\sqrt{N}(T_N - \theta(\mathbf{F}))}{\sigma(\mathbf{F})} \xrightarrow{\mathcal{L}} Z \sim \mathcal{N}(0, 1) \quad \text{as } N \rightarrow \infty \quad \forall \mathbf{F} \in \mathcal{F} \quad (\text{A.23})$$

for suitable functionals  $\theta(\cdot) : \mathcal{F} \rightarrow \mathbb{R}$  and  $\sigma(\cdot) : \mathcal{F} \rightarrow \mathbb{R}_+$ . Suppose further, that  $(\hat{\sigma}_N)_{N \in \mathbb{N}}$  is a consistent estimator of  $\sigma(\mathbf{F})$  for each  $\mathbf{F} \in \mathcal{F}$ , and  $c_\alpha^{\theta_1, \theta_2}(\cdot)$  is as in (A.20). If we then define

$$\phi_N(X^{(N)}) = \begin{cases} 1 & \text{for } \sqrt{N}|T_N - \theta_\circ|/\hat{\sigma}_N < c_\alpha^{\theta_1, \theta_2}(\sqrt{N}/\hat{\sigma}_N) \\ 0 & \text{for } \sqrt{N}|T_N - \theta_\circ|/\hat{\sigma}_N \geq c_\alpha^{\theta_1, \theta_2}(\sqrt{N}/\hat{\sigma}_N) \end{cases} \quad (\text{A.24})$$

with  $\theta_\circ \equiv (\theta_1 + \theta_2)/2$ , we can be sure that  $(\phi_N)_{N \in \mathbb{N}}$  is asymptotically valid at level  $\alpha \in (0, 1)$  for  $H : \theta(\mathbf{F}) \in \Theta \setminus (\theta_1, \theta_2)$  versus  $K : \theta(\mathbf{F}) \in (\theta_1, \theta_2)$ .

*Proof.* First we consider an arbitrarily fixed  $\mathbf{F} \in \mathcal{H}$  such that  $\theta(\mathbf{F}) \geq \theta_2$  and write for brevity  $\theta(\mathbf{F}) = \theta$ ,  $\sigma(\mathbf{F})/\sqrt{N} = \tau_N$ ,  $\hat{\sigma}_N/\sqrt{N} = \hat{\tau}_N$ . By definition of  $\phi_N(\cdot)$ , we have

$$\begin{aligned} E_{\mathbf{F}}^{(N)}(\phi_N) &= P_{\mathbf{F}}^{(N)} \left[ (T_N - \theta_\circ)/\hat{\tau}_N < c_\alpha^{\theta_1, \theta_2}(1/\hat{\tau}_N) \right] - \\ &\quad P_{\mathbf{F}}^{(N)} \left[ (T_N - \theta_\circ)/\hat{\tau}_N \leq -c_\alpha^{\theta_1, \theta_2}(1/\hat{\tau}_N) \right] \\ &\leq P_{\mathbf{F}}^{(N)} \left[ (T_N - \theta)/\hat{\tau}_N + (\theta - \theta_\circ)/\hat{\tau}_N - c_\alpha^{\theta_1, \theta_2}(1/\hat{\tau}_N) < 0 \right] \\ &\leq P_{\mathbf{F}}^{(N)} \left[ (T_N - \theta)/\hat{\tau}_N + (\theta_2 - \theta_1)/2\hat{\tau}_N - c_\alpha^{\theta_1, \theta_2}(1/\hat{\tau}_N) < 0 \right] \quad (\text{A.25}) \\ &\quad \text{by } \theta \geq \theta_2, \theta_\circ = (\theta_1 + \theta_2)/2 . \end{aligned}$$

By assumption (A.23) and consistency of  $(\hat{\sigma}_N)_{N \in \mathbb{N}}$  for  $\sigma(\mathbf{F})$  we can write:

$$P_{\mathbf{F}}^{(N)} \left[ (T_N - \theta)/\hat{\tau}_N < w \right] \rightarrow \Phi(w) \quad \text{as } N \rightarrow \infty \quad \forall w \in \mathbb{R} \quad (\text{A.26})$$

and

$$N^{1/2} \hat{\tau}_N \xrightarrow{P_{\mathbf{F}}^{(N)}} \sigma^* \text{ as } N \rightarrow \infty \text{ for some } \sigma^* \in (0, \infty). \quad (\text{A.27})$$

In view of (A.20) and (A.21) all the conditions of Lemma A.3.3 are satisfied if we put  $P_N = P_{\mathbf{F}}^{(N)}$ ,  $S_N = (N^{1/2} \hat{\tau}_N)^{-1}$ ,  $c = 1/\sigma^*$ ,  $q(s) = c_{\alpha}^{\theta_1, \theta_2}(s) - s(\theta_2 - \theta_1)/2$ ,  $y^* = u_{\alpha}$ , and  $a_N = N^{1/2}$ . Hence, it follows that

$$(\theta_2 - \theta_1)/2\hat{\tau}_N - c_{\alpha}^{\theta_1, \theta_2}(1/\hat{\tau}_N) \xrightarrow{P_{\mathbf{F}}^{(N)}} -u_{\alpha} \text{ as } N \rightarrow \infty. \quad (\text{A.28})$$

Combining (A.28) with (A.26) gives

$$\begin{aligned} P_{\mathbf{F}}^{(N)} \left[ (T_N - \theta)/\hat{\tau}_N + (\theta_2 - \theta_1)/2\hat{\tau}_N - c_{\alpha}^{\theta_1, \theta_2}(1/\hat{\tau}_N) < w \right] \\ \rightarrow \Phi(w + u_{\alpha}) \text{ as } N \rightarrow \infty \quad \forall w \in \mathbb{R}. \end{aligned} \quad (\text{A.29})$$

Taking  $w = 0$  in (A.29) shows that the expression on the right-hand side of (A.25) converges to  $\alpha$  as  $N \rightarrow \infty$  which completes the  $\theta \geq \theta_2$  half of the proof.

Let us now choose  $\mathbf{F} \in \mathcal{H}$  such that  $\theta = \theta(\mathbf{F}) \leq \theta_1$ . Then  $E_{\mathbf{F}}^{(N)}(\phi_N)$  can be bounded above in the following way:

$$\begin{aligned} E_{\mathbf{F}}^{(N)}(\phi_N) &\leq 1 - P_{\mathbf{F}}^{(N)} \left[ (T_N - \theta_{\circ})/\hat{\tau}_N \leq -c_{\alpha}^{\theta_1, \theta_2}(1/\hat{\tau}_N) \right] \\ &= 1 - P_{\mathbf{F}}^{(N)} \left[ (T_N - \theta)/\hat{\tau}_N + (\theta - \theta_{\circ})/\hat{\tau}_N + c_{\alpha}^{\theta_1, \theta_2}(1/\hat{\tau}_N) \leq 0 \right] \\ &\leq 1 - P_{\mathbf{F}}^{(N)} \left[ (T_N - \theta)/\hat{\tau}_N - (\theta_2 - \theta_1)/2\hat{\tau}_N + c_{\alpha}^{\theta_1, \theta_2}(1/\hat{\tau}_N) \leq 0 \right] \quad (\text{A.30}) \\ &\text{by } \theta \leq \theta_1, \quad \theta_{\circ} = (\theta_1 + \theta_2)/2. \end{aligned}$$

Since relations (A.26) and (A.28) hold true for any  $\mathbf{F}$  with  $0 < \theta(\mathbf{F}) < 1$  we get in direct analogy to (A.29)

$$\begin{aligned} P_{\mathbf{F}}^{(N)} \left[ (T_N - \theta)/\hat{\tau}_N - (\theta_2 - \theta_1)/2\hat{\tau}_N + c_{\alpha}^{\theta_1, \theta_2}(1/\hat{\tau}_N) \leq w \right] \\ \rightarrow \Phi(w - u_{\alpha}) \text{ as } N \rightarrow \infty \quad \forall w \in \mathbb{R}. \end{aligned} \quad (\text{A.31})$$

Specifying  $w = 0$  in this latter relation shows that the right-hand side of (A.30) converges to  $1 - \Phi(-u_{\alpha})$  as  $N \rightarrow \infty$  so that the  $\theta \leq \theta_1$  part of the proof is complete as well.  $\blacksquare$

# Appendix B

---

## *List of special computer programs*

Program Name	Objective	Language	→ p.
bi2aeq1	Power of the exact Fisher type test for equivalence	Fortran, (R) <sup>†</sup>	191
bi2aeq2	Sample sizes for the exact Fisher type test for equivalence	Fortran, (R)	191
bi2aeq3	Increased nominal significance level for the nonrandomized version of the exact Fisher type test for equivalence	Fortran, (R)	193
bi2by_ni_del	Corrected nominal levels for the objective Bayesian test for one-sided equivalence of two binomial distributions with respect to the difference	SAS/IML	185
bi2by_ni_or	Corrected nominal levels for the objective Bayesian test for one-sided equivalence of two binomial distributions with respect to the odds ratio	SAS/IML	180
bi2diffac	Corrected nominal significance level for the asymptotic test for equivalence of two unrelated binomial proportions with respect to the difference of their population counterparts	SAS, R	196
bi2dipow	Exact rejection probability of the asymptotic test for equivalence of two unrelated binomial proportions with respect to the difference of their expectations at any nominal level under an arbitrary parameter configuration	SAS, R	196
bi2rlv1	Power of the exact Fisher type test for relevant differences	Fortran, (R)	366
bi2rlv2	Sample sizes for the exact Fisher type test for relevant differences	Fortran, (R)	367

<sup>†</sup>) (R) indicates that a shared object accessible within R is available.

Program Name	Objective	Language	→ p.
bi1st	Critical constants and power against the alternative $p = (p_1 + p_2)/2$ of the UMP test for equivalence of a single binomial proportion to some given reference value	SAS, R	60
bi2st	Critical constants for the exact Fisher type UMPU test for equivalence of two binomial distributions with respect to the odds ratio	SAS, R	190
bi2ste1	Power of the exact Fisher type test for noninferiority	Fortran, (R)	175
bi2ste2	Sample sizes for the exact Fisher type test for noninferiority	Fortran, (R)	175
bi2ste3	Increased nominal significance level for the nonrandomized version of the exact Fisher type test for noninferiority	Fortran, (R)	177
bi2wld_ni_del	Corrected nominal levels for the Wald-type (asymptotic) test for one-sided equivalence of two binomial distributions with respect to the difference	SAS/IML	184
cf_reh_exact	Exact confidence bounds to the relative excess heterozygosity (REH) exhibited by a SNP genotype distribution	SAS/IML	306
cf_reh_midp	Mid-p-value – based confidence bounds to the relative excess heterozygosity (REH) exhibited by a SNP genotype distribution	SAS/IML	307
exp1st	Critical constants and power against the null alternative of the UMP test for equivalence of the hazard rate of a single exponential distribution to some given reference value	SAS, R	57
fstretch	Critical constants and power against $\sigma^2 = \tau^2$ of the UMPI test for dispersion equivalence of two Gaussian distributions	SAS, R	168
gofhwex	Critical constants of the exact UMPU test for approximate compatibility of a SNP genotype distribution with HWE	SAS, R	296

Program Name	Objective	Lan-	guage	→ p.
<code>gofhwex_1s</code>	Critical constants of the exact UMPU test for absence of a substantial deficit of heterozygotes as compared with a HWE-conform SNP genotype distribution [ $\leftrightarrow$ non-inferiority version of the test implemented by means of <code>gofhwex</code> ]	SAS, R		301
<code>gofind_t</code>	Establishing approximate independence in a two-way contingency table: Test statistic and critical bound	SAS, R		274
<code>gofsimpt</code>	Establishing goodness of fit of an observed to a fully specified multinomial distribution: Test statistic and critical bound	SAS, R		268
<code>mawi</code>	Mann-Whitney test for equivalence of two continuous distributions of arbitrary shape: Test statistic and critical upper bound	SAS, R		128
<code>mcnasc_ni</code>	Corrected nominal levels for the asymptotic test for noninferiority in the McNemar setting	SAS, R		86
<code>mcnby_ni</code>	Analogue of <code>mcnasc_ni</code> for the objective Bayesian test for noninferiority in the McNemar setting	SAS/IML		89
<code>mcnby_ni_pp</code>	Posterior probability of the alternative hypothesis of noninferiority in the McNemar setting, given a specific point in the sample space	SAS/IML		89
<code>mcnemasc</code>	Corrected nominal significance level for the asymptotic test for equivalence of two paired binomial proportions with respect to the difference of their expectations (McNemar setting)	SAS, R		83
<code>mcnempow</code>	Exact rejection probability of the asymptotic test for equivalence of two paired binomial proportions with respect to the difference of their expectations (McNemar setting)	SAS, R		84
<code>mwtie_fr</code>	Analogue of <code>mwtie_xy</code> for settings with grouped data	SAS, R		155
<code>mwtie_xy</code>	Distribution-free two-sample equivalence test for tied data: Test statistic and critical upper bound	SAS, R		154

Program Name	Objective	Language	$\rightarrow p$
po_pbibe	Bayesian posterior probability of the alternative hypothesis of individual bioequivalence	SAS, R	331
postmys	Bayesian posterior probability of the alternative hypothesis $\varepsilon_1 < \delta/\sigma_D < \varepsilon_2$ in the setting of the one-sample $t$ -test	SAS, R	38
pow_abe	interval inclusion test for average bioequivalence: exact power against an arbitrary specific alternative	SAS, R	318
powsign	Nonconditional power of the UMPU sign test for equivalence and its nonrandomized counterpart against the alternative $p_+ = p_-$	SAS, R	73
sgnrk	Signed rank test for equivalence of an arbitrary continuous distribution of the intraindividual differences to a distribution satisfying $q_+ \equiv P[D_i + D_j > 0] = 1/2$ : Test statistic and critical upper bound	SAS, R	101
srktie_d	Generalized signed rank test for equivalence for tied data: Test statistic and critical upper bound	SAS, R	111
srktie_m	Analogue of srktie_d for settings where the distribution of the $D_i$ is concentrated on a finite lattice	SAS, R	113
tt1st	Critical constants and power against the null alternative of the one-sample $t$ -test for equivalence with an arbitrary, maybe nonsymmetric choice of the limits of the equivalence range for $\delta/\sigma_D$	SAS, R	96
tt2st	Critical constants and power against the null alternative of the two-sample $t$ -test for equivalence with an arbitrary, maybe nonsymmetric choice of the limits of the equivalence range for $(\xi - \eta)/\sigma$	SAS, R	122

# Appendix C

---

## *Frequently used special symbols and abbreviations*

BE	bioequivalence
cdf	cumulative distribution function
iff	if and only if
$df$	number of degrees of freedom
$\lambda_{nc}^2$	noncentrality parameter of a $\chi^2$ or $F$ -distribution
$\text{STP}_r$	strictly totally positive of order $r$ , with $r \in \mathbb{N}$ or $r = \infty$
$\phi(\cdot)$	critical function of a test
UMP	uniformly most powerful
UMPU	uniformly most powerful unbiased
UMPI	uniformly most powerful invariant
$\mathbb{N}$	set of natural numbers
$\mathbb{N}_0$	" " nonnegative integers
$\mathbb{R}$	" " real numbers
$\mathbb{R}_+$	" " positive real numbers
$\mathbb{Z}$	" " integers
$\# A$	number of elements of an arbitrary finite set $A$
$I_A(x)$	indicator function of a given subset $A \subseteq \mathbb{R}$ at $x \in \mathbb{R}$
$\mathbf{1}_{(r,s)}$	$r \times s$ matrix with all elements equal to 1.
$d^2(\mathbf{u}, \mathbf{v})$	squared Euclidean distance between any two vectors $\mathbf{u}$ , $\mathbf{v}$ of the same dimension
$\equiv$	equal by definition
$\stackrel{d}{=}$	equal in distribution
$\times$	Cartesian product of sets
$\xrightarrow{P}$	convergence in probability
$\xrightarrow{\mathcal{L}}$	convergence in law
$X \sim \mathcal{D}$	random variable $X$ follows distribution $\mathcal{D}$
$\mathcal{U}(a, b)$	uniform distribution over the interval $(a, b) \subseteq \mathbb{R}$
$\mathcal{B}(n, p)$	binomial distribution with parameters $n \in \mathbb{N}_0$ and $p \in [0, 1]$
$b(k; n, p)$	probability mass function of $\mathcal{B}(n, p)$ evaluated at $k \in \{0, 1, \dots, n\}$

$\mathcal{M}(n; \theta_1, \dots, \theta_k)$	multinomial distribution of dimension $k \geq 2$ with parameters $n \in \mathbb{N}_0$ and $(\theta_1, \dots, \theta_k) \in [0, 1]^k$ , $\sum_{j=1}^k \theta_j = 1$ .
$h_s^{m,n}(x; \rho)$	probability mass function at $x$ of the conditional distribution of $X \sim \mathcal{B}(m, p_1)$ given $X + Y = s$ , with $Y \sim \mathcal{B}(n, p_2)$ independent of $X$ and $\rho = p_1(1 - p_2)/(1 - p_1)p_2$
$\mathcal{E}(\sigma)$	exponential distribution with scale parameter $\sigma \in \mathbb{R}_+$
$\Gamma(x)$	gamma function (complete gamma integral) at $x > 0$
$\mathcal{N}(\xi, \sigma^2)$	normal distribution with expectation $\xi \in \mathbb{R}$ and variance $\sigma^2 \in \mathbb{R}_+$
$\varphi(\cdot)$	density function of $\mathcal{N}(0, 1)$
$\Phi(\cdot)$	cdf of $\mathcal{N}(0, 1)$
$\Phi^{-1}(\cdot)$	quantile function of $\mathcal{N}(0, 1)$
$u_{1-\alpha}$	$\Phi^{-1}(1 - \alpha)$ , for arbitrary $\alpha \in (0, 1)$
$\chi_{\nu; \gamma}^2$	$\gamma$ -quantile ( $0 < \gamma < 1$ ) of a central $\chi^2$ -distribution with $\nu$ degrees of freedom
$C_\alpha(\psi)$	square root of the $\alpha$ -quantile ( $0 < \alpha < 1$ ) of a noncentral $\chi^2$ -distribution with a single degree of freedom and $\lambda_{nc}^2 = \psi^2$ ( $\psi \in \mathbb{R}_+$ )
$C_{\alpha; \tilde{\varepsilon}}$	$= C_\alpha(\tilde{\varepsilon})$ , with $(-\tilde{\varepsilon}, \tilde{\varepsilon})$ as the equivalence range for the expectation of a Gaussian distribution with unit variance
$\chi_{\nu; \gamma}^2(\psi^2)$	$\gamma$ -quantile ( $0 < \gamma < 1$ ) of a noncentral $\chi^2$ -distribution with $\nu$ degrees of freedom and $\lambda_{nc}^2 = \psi^2$
$f_\nu^\chi(\cdot)$	density of a $\chi$ -distribution with $\nu$ degrees of freedom
$F_\nu^T(\cdot)$	cdf of a central $t$ -distribution with $\nu$ degrees of freedom
$t_{\nu; \gamma}$	$\gamma$ -quantile ( $0 < \gamma < 1$ ) of a central $t$ -distribution with $\nu$ degrees of freedom
$F_{\nu; \psi}^T(\cdot)$	cdf of a noncentral $t$ -distribution with $\nu$ degrees of freedom and noncentrality parameter $\psi \in \mathbb{R}$
$t_{\nu; \gamma}(\psi)$	$\gamma$ -quantile ( $0 < \gamma < 1$ ) of a noncentral $t$ -distribution with $\nu$ degrees of freedom and noncentrality parameter $\psi \in \mathbb{R}$
$F_{\nu_1, \nu_2}$	cdf of a central $F$ -distribution with $\nu_1, \nu_2$ degrees of freedom
$F_{\nu_1, \nu_2; \gamma}(\psi^2)$	$\gamma$ -quantile of a noncentral $F$ -distribution with $\nu_1, \nu_2$ degrees of freedom and $\lambda_{nc}^2 = \psi^2$
$\mathcal{F}_{\nu_1, \nu_2}(\psi^2)$	random variable following a noncentral $F$ -distribution with $\nu_1, \nu_2$ degrees of freedom and $\lambda_{nc}^2 = \psi^2$
$p_+$	$P[D_i > 0]$ , with $D_i$ as the $i$ th intra-subject difference obtained from a sample of paired observations
$p_0$	$P[D_i = 0]$ , with $D_i$ as above
$p_-$	$P[D_i < 0]$ , " "

$q_+$	$P[D_i + D_j > 0]$ , with $D_i, D_j$ as the intra-subject difference for two different observational units in a paired-data setting
$q_0$	$P[D_i + D_j = 0]$ , with $D_i, D_j$ as above
$q_-$	$P[D_i + D_j < 0]$ , " " " "
$U_+$	$U$ -statistic estimator of $q_+$
$U_0$	$U$ -statistic estimator of $q_0$
$\pi_+$	$P[X_i > Y_j]$ , with $X_i$ and $Y_j$ as independent observations from two different univariate distributions under comparison
$\pi_0$	$P[X_i = Y_j]$ , with $X_i$ and $Y_j$ as above
$W_+$	$U$ -statistic estimator of $\pi_+$
$W_0$	$U$ -statistic estimator of $\pi_0$
$  S_1 - S_2  $	$\sup_{t>0}  S_1(t) - S_2(t) $ , with $S_1(\cdot), S_2(\cdot)$ as two survivor functions of the continuous type
$\lambda_\nu(t)$	hazard function at $t \geq 0$ of the $\nu$ th sample ( $\nu = 1, 2$ )
$I_N(\hat{\beta})$	observed information evaluated at the maximum likelihood estimate of the regression coefficient $\beta \in \mathbb{R}$
$\phi_T, \phi_R$	"direct" effect of drug formulation $T$ ( $\leftrightarrow$ "Test") and $R$ ( $\leftrightarrow$ "Reference"), respectively
$\pi_k$	$k$ th period effect in a $2 \times 2$ crossover trial ( $k = 1, 2$ )
$\sigma_S^2$	between-subject variance in a standard BE trial
$\sigma_{eT}^2, \sigma_{eR}^2$	within-subject variance associated with drug formulation $T$ and $R$ , respectively
$\sigma_{TT}^2, \sigma_{TR}^2$	total variance of the bioavailability of drug formulation $T$ and $R$ , respectively
$X_{ki}$	bioavailability measured in the $i$ th subject ( $i = 1, \dots, m$ ) of sequence group $T/R$ during the $k$ th period ( $k = 1, 2$ ) of a standard bioequivalence trial
$X_i^-$	$X_{1i} - X_{2i}$
$\mu^-$	$E(X_i^-)$
$Y_{kj}$	bioavailability measured in the $j$ th subject ( $j = 1, \dots, n$ ) of sequence group $R/T$ during the $k$ th period ( $k = 1, 2$ ) of a standard bioequivalence trial
$Y_j^-$	$Y_{1j} - Y_{2j}$
$\nu^-$	$E(Y_j^-)$
$\lambda^{(1)}, \lambda^{(2)}$	carryover effect of the treatment administered in the first period of a $2 \times 2$ crossover trial for the 1st ( $\leftrightarrow T/R$ ) and 2nd ( $\leftrightarrow R/T$ ) sequence group

---

## References

- Abramowitz, M. and Stegun, I. (1965) *Handbook of Mathematical Functions*. New York: Dover Publications, Inc.
- Agresti, A. (2002) *Categorical Data Analysis. Second Edition*. Hoboken, NJ: John Wiley & Sons, Inc.
- Altman, D.G. and Bland, J.M. (1995) Absence of evidence is not evidence of absence. *British Medical Journal* 311: 485.
- Andersen, P.K., Borgan, Ø., Gill, R.D. and Keiding, N. (1993) *Statistical Models Based on Counting Processes*. Berlin: Springer-Verlag.
- Andersen, P.K. and Gill, R.D. (1982) Cox's regression model for counting processes: A large sample study. *Annals of Statistics* 10: 1100–20.
- Anderson, S. and Hauck, W.W. (1983) A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistics: Theory and Methods* 12: 2663–92.
- Anderson, S. and Hauck, W.W. (1990) Consideration of individual bioequivalence. *Journal of Pharmacokinetics and Biopharmacology* 18: 259–73.
- Anderson, T.W. (1984) *An Introduction to Multivariate Statistical Analysis. Second Edition*. New York: John Wiley & Sons, Inc.
- Armitage, P. (1955) Tests for linear trends in proportions and frequencies. *Biometrics* 11: 375–86.
- Bailey, C.C., Gnekow, A., Wellek, S., Jones, M., Round, C., Brown, J., Philips, A. and Neidhardt, M.K. (1995) Prospective randomized trial of chemotherapy given before radiotherapy in childhood medulloblastoma. International Society of Paediatric Oncology SIOP and the German Society of Paediatric Oncology GPO: SIOP II. *Medical Pediatric Oncology* 25: 166–78.
- Barnard, G.A. (1947) Significance tests for 2x2 tables. *Biometrika* 34: 123–38.
- Berger, J.O. (1985) *Statistical Decision Theory and Bayesian Analysis. Second Edition*. New York: Springer-Verlag.
- Berger, R.L. (1982) Multiparameter hypothesis testing and acceptance sampling. *Technometrics* 24: 295–300.

- Berger, R.L. and Boos, D.D. (1994) P values maximised over a confidence set for the nuisance parameter. *Journal of the American Statistical Association* 89: 1012–16.
- Berger, R.L. and Hsu, J.C. (1996) Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science* 11: 283–319.
- Berger, R.L. and Sidik, K. (2003) Exact unconditional tests for a  $2 \times 2$  matched pairs design. *Statistical Methods in Medical Research* 12: 91–108.
- Bickel, P.J. and Doksum, K.A. (2001) *Mathematical Statistics: Basic Ideas and Selected Topics. Vol. I. Second Edition.* Upper Saddle River, NJ: Prentice Hall.
- Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975) *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, MA: MIT Press.
- Blackwelder, W.C. (1982) “Proving the null hypothesis” in clinical trials. *Controlled Clinical Trials* 3: 345–53.
- Blackwelder, W.C. (1993) Sample size and power for prospective analysis of relative risk. *Statistics in Medicine* 12: 691–8.
- Bondy, W.H. (1969) A test of an experimental hypothesis of negligible difference between means. *The American Statistician* 23: 28–30.
- Boschloo, R.D. (1970) Raised conditional level of significance for the  $2 \times 2$ -table when testing the equality of two probabilities. *Statistica Neerlandica* 24: 1–35.
- Box, G.P.E. and Tiao, G.C. (1973) *Bayesian Inference in Statistical Analysis.* Reading, MA: Addison-Wesley.
- Brown, B.W. (1980) The crossover experiment for clinical trials. *Biometrics* 36: 69–79.
- Brown, L.D., Hwang, J.T.G. and Munk, A. (1997) An unbiased test for the bioequivalence problem. *Annals of Statistics* 25: 2345–67.
- Casella, G. and Berger, R.L. (1987) Reconciling bayesian and frequentist evidence in one-sided testing problems. *Journal of the American Statistical Association* 82: 106–11.
- Chan, I.S.F. (1998) Exact tests of equivalence and efficacy with a non-zero lower bound for comparative studies. *Statistics in Medicine* 17: 1403–13.
- Chan, I.S.F. (2003) Proving non-inferiority or equivalence of two treatments with dichotomous endpoints using exact methods. *Statistical Methods in Medical Research* 12: 37–58.
- Chan, I.S.F. and Zhang, Z. (1999) Test-based exact confidence intervals for the difference of two binomial proportions. *Biometrics* 55: 1202–9.

- Chen, M.L. (1997) Individual bioequivalence – A regulatory update. *Journal of Biopharmaceutical Statistics* 7: 5–11.
- Chester, S.T. (1986) Review of equivalence and hypothesis testing. In *ASA Proceedings of the Biopharmaceutical Section*. Washington: American Statistical Association, 177–82.
- Chinchilli, V.M. and Elswick Jr., R.K. (1997) The multivariate assessment of bioequivalence. *Journal of Biopharmaceutical Statistics* 7: 113–23.
- Chow, S.C. and Liu, J.P. (2008) *Design and Analysis of Bioavailability and Bioequivalence Studies, Third Edition*. Boca Raton, FL: Chapman & Hall/CRC Press.
- Cox, D.R. (1958) Some problems connected with statistical inference. *Annals of Mathematical Statistics* 29: 357–72.
- Cox, D.R. and Hinkley, D.V. (1974) *Theoretical Statistics*. London: Chapman & Hall.
- D'Angelo, G., Potvin, D. and Turgeon, J. (2001) Carry-over effects in bioequivalence studies. *Journal of Biopharmaceutical Statistics* 11: 35–43.
- Dannenberg, O., Dette, H. and Munk, A. (1994) An extension of Welch's approximate *t*-solution to comparative bioequivalence trials. *Biometrika* 81: 91–101.
- Davies, R. (1971) Rank tests for “Lehmann's alternative.” *Journal of the American Statistical Association* 66: 879–83.
- Davis, P.J. and Rabinowitz, P. (1975) *Methods of Numerical Integration*. New York: Academic Press.
- Devlin, B. and Roeder, K. (1999) Genomic control for association studies. *Biometrics* 55: 997–1004.
- Dunnett, C.W. and Gent, M. (1977) Significance testing to establish equivalence between treatments, with special reference to data in the form of 2 × 2 tables. *Biometrics* 33: 593–602.
- EMEA (1998) *CPMP/ICH/363/96*.  
<http://www.emea.europa.eu/pdfs/human/ich/03636en.pdf>.
- EMEA (2005) *CPMP/EWP/2158/99*.  
<http://www.emea.europa.eu/pdfs/human/ewp/21589en.pdf>.
- Farrington, C.P. and Manning, G. (1990) Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unit relative risk. *Statistics in Medicine* 9: 1447–54.
- Feller, W. (1968) *An Introduction to Probability Theory and Its Applications*.

- Volume I. Third Edition. New York: John Wiley & Sons, Inc.
- Feller, W. (1971) *An Introduction to Probability Theory and Its Applications*. Volume II. Second Edition. New York: John Wiley & Sons, Inc.
- Firle, E.A. (1998) *Asymptotically Distribution-Free Tests for Equivalence for Paired Observations Allowing for Ties*. Diploma thesis, Department of Mathematics, University of Mainz, Mainz, Germany.
- Fisher, R.A. (1934) *Statistical Methods for Research Workers*. 5th Edition. Edinburgh: Oliver & Boyd.
- Flühler, H., Grieve, A.P., Mandallaz, D., Mau, J. and Moser, H. (1983) Bayesian approach to bioequivalence assessment: An example. *Journal of Pharmacological Sciences* 72: 1178–81.
- Food and Drug Administration FDA (1997) *Bioequivalence Studies-II. Gender Studies with non-replicate Designs*. Center for Drug Evaluation and Research at <http://www.fda.gov/bioequivdata>.
- Food and Drug Administration FDA (2001) *Guidance for Industry: Statistical Approaches to Establishing Bioequivalence*. Rockville, MD: Center for Drug Evaluation and Research CDER.
- Food and Drug Administration FDA (2008) *Orange Book, Version 12/2008*. <http://www.fda.gov/cder/orange/obreadme.htm>.
- Freeman, P. (1989) The performance of the two-stage analysis of two treatment, two period crossover trials. *Statistics in Medicine* 8: 1421–32.
- Freitag, G. (2005) Methods for assessing noninferiority with censored data. *Biometrical Journal* 47: 88–98.
- Freitag, G., Lange, S. and Munk, A. (2006) Non-parametric assessment of non-inferiority with censored data. *Statistics in Medicine* 25: 1201–17.
- Frick, H. (1987) On level and power of Anderson and Hauck's procedure for testing equivalence in comparative bioavailability trials. *Communications in Statistics, Theory and Methods* 16: 2771–8.
- Giani, G. and Finner, H. (1991) Some general results on least favorable parameter configurations with special reference to equivalence testing and the range statistic. *Journal of Statistical Planning and Inference* 28: 33–47.
- Giani, G., Straßburger, K. and Seefeldt, F. (2005) *SeParATe, Version 3.0B*. Diabetesforschungsinstitut an der Heinrich-Heine-Universität, Abt. Biometrie & Epidemiologie, Düsseldorf, Germany.
- Goddard, K.A.B., Ziegler, A. and Wellek, S. (2009) Adapting the logical basis of tests for Hardy-Weinberg equilibrium to the real needs of association studies in human and medical genetics. *Genetic Epidemiology* 33: 569–80.

- Grizzle, J.E. (1965) The two-period change-over design and its use in clinical trials. *Biometrics* 21: 467–80.
- Guilbaud, O. (1993) Exact inference about the within-subject variability in  $2 \times 2$  crossover trials. *Journal of the American Statistical Association* 88: 939–46.
- Hájek, J., Šidák, Z. and Sen, P.K. (1999) *Theory of Rank Tests*. San Diego: Academic Press.
- Hájek, P. and Havránek, T. (1978) *Mechanizing Hypothesis Formation*. Berlin: Springer-Verlag.
- Halmos, P.R. (1974) *Measure Theory*. Berlin: Springer-Verlag.
- Harkness, W.L. (1965) Properties of the extended hypergeometric distribution. *Annals of Mathematical Statistics* 36: 938–45.
- Hauck, W.W. and Anderson, S. (1996) Comment on “Bioequivalence trials, intersection-union tests and equivalence confidence sets.” *Statistical Science* 11: 303.
- Hauck, W.W., Hauschke, D., Diletti, E., Bois, F.Y., Steinijans, V.W. and Anderson, S. (1997) Choice of Student’s  $t$ - or Wilcoxon-based confidence intervals for assessment of average bioequivalence. *Journal of Biopharmaceutical Statistics* 7: 179–89.
- Hauschke, D., Steinijans, V.W. and Diletti, E. (1990) A distribution-free procedure for the statistical analysis of bioequivalence studies. *International Journal of Clinical Pharmacology and Therapeutics* 28: 72–8.
- Hauschke, D., Steinijans, V.W. and Hothorn, L.A. (1996) A note on Welch’s approximate  $t$ -solution to bioequivalence assessment. *Biometrika* 83: 236–7.
- Hauschke, D., Steinijans, V.W. and Pigeot, I. (2007) *Bioequivalence Studies in Drug Development: Methods and Applications*. Chichester: John Wiley & Sons, Inc.
- Hodges, J.L. and Lehmann, E.L. (1954) Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society, Series B* 16: 262–8.
- Hoffelder, T. (2006) *Two-Sample Equivalence Tests for Multivariate Normal Data*. Diploma thesis, Department of Mathematics, University of Mainz, Mainz, Germany.
- Holder, D.J. and Hsuan, F. (1993) Moment-based criteria for determining bioequivalence. *Biometrika* 80: 835–46.
- Holzgreve, H., Distler, A., Michaelis, J., Philipp, T. and Wellek, S. (1989)

- Verapamil versus hydrochlorothiazide for the treatment of hypertension. Results of a long-term double-blind comparative trial. *British Medical Journal* 299: 881–6.
- Hotelling, H. (1931) The generalization of Student's ratio. *Annals of Mathematical Statistics* 2: 360–78.
- Hsu, J.C., Hwang, J.T.G., Liu, H.K. and Ruberg, S.J. (1994) Confidence intervals associated with tests for bioequivalence. *Biometrika* 81: 103–14.
- Hsueh, H.M., Liu, J.P. and Chen, J.J. (2001) Unconditional exact tests for equivalence or noninferiority for paired binary endpoints. *Biometrics* 57: 478–83.
- Janssen, A. (1999) Testing nonparametric statistical functionals with application to rank tests. *Journal of Statistical Planning and Inference* 81: 71–93.
- Janssen, A. (2000) Nonparametric bioequivalence tests for statistical functionals and their efficient power functions. *Statistics & Decisions* 18: 49–78.
- Janssen, A. and Wellek, S. (2008) Exact linear rank tests for two-sample equivalence problems with continuous data. In *Technical Report*. University of Duesseldorf, Düsseldorf, Germany.
- Jeffreys, H. (1961) *Theory of Probability. Third Edition*. Oxford: Clarendon Press.
- Johnson, A.D. and O'Donnell, C.J. (2009) An open access database of genome-wide association results. *BMC Medical Genetics* 10: 6.
- Johnson, N.L., Kotz, S. and Balakrishnan, N. (1994) *Continuous Univariate Distributions. Volume 1. Second Edition*. New York: John Wiley & Sons, Inc.
- Johnson, N.L., Kotz, S. and Balakrishnan, N. (1995) *Continuous Univariate Distributions. Volume 2. Second Edition*. New York: John Wiley & Sons, Inc.
- Johnson, N.L., Kotz, S. and Balakrishnan, N. (1997) *Discrete Multivariate Distributions*. New York: John Wiley & Sons, Inc.
- Johnson, N.L., Kotz, S. and Kemp, A.W. (1992) *Univariate Discrete Distributions. Second Edition*. New York: John Wiley & Sons, Inc.
- Jones, B. and Kenward, M.G. (1989) *Design and Analysis of Cross-Over Trials*. London: Chapman & Hall.
- Jones, B. and Kenward, M.G. (2003) *Design and Analysis of Cross-Over Trials*.

- als. Second Edition. Boca Raton, FL: Chapman & Hall/CRC Press.
- Kalbfleisch, J.D. and Prentice, R.L. (2002) *The Statistical Analysis of Failure Time Data, 2nd Ed.* Hoboken, NJ: John Wiley & Sons, Inc.
- Kallenberg, W.C.M. (1984) *Testing Statistical Hypotheses: Worked Solutions.* Amsterdam: Centrum voor Wiskunde en Informatica.
- Karlin, S. (1956) Decision theory for Pólya type distributions. Case of two actions, I. In J. Neyman, ed., *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley, CA: California University Press., 115–28.
- Karlin, S. (1957a) Pólya type distributions, II. *Annals of Mathematical Statistics* 28: 281–308.
- Karlin, S. (1957b) Pólya type distributions, III: Admissibility for multi-action problems. *Annals of Mathematical Statistics* 28: 839–60.
- Karlin, S. (1968) *Total Positivity. Volume 1.* Stanford, CA: Stanford University Press.
- Kingman, J.F.C. and Taylor, S.J. (1973) *Introduction to Measure and Probability.* Cambridge: Cambridge University Press.
- Kotz, S., Balakrishnan, N. and Johnson, N.L. (2000) *Continuous Multivariate Distributions, Vol. 1: Models and Applications.* New York: John Wiley & Sons, Inc.
- Lachenbruch, P.A. and Lynch, C.J. (1998) Assessing screening tests: extensions of McNemars's test. *Statistics in Medicine* 17: 2207–17.
- Lambert, J.C., Araria-Goumidi, L., Myllykangas, L., Ellis, C., Wang, J.C., Bullido, M.J., Harris, J.M., Artiga, M.J., Hernandez, D., Kwon, J.M., Frigard, B., Petersen, R.C., Cumming, A.M., Pasquier, F., Sastre, I., Tienari, P.J., Frank, A., Sulkava, R., Morris, J.C., St Clair, D., Mann, D.M., Wavrant-DeVrièze, F., Ezquerro-Trabalón, M., Amouyel, P., Hardy, J., Haltia, M., Valdivieso, F., Goate, A.M., Pérez-Tur, J., Lendon, C.L. and Chartier-Harlin, M.C. (2002) Contribution of APOE promoter polymorphisms to Alzheimer's disease risk. *Neurology* 59: 59–66.
- Lee, A.J. (1990) *U-Statistics. Theory and Practice.* New York: Marcel Dekker Inc.
- Lehmann, E.L. (1963) Nonparametric confidence intervals for a shift parameter. *Annals of Mathematical Statistics* 34: 1507–12.
- Lehmann, E.L. (1975) *Nonparametrics: Statistical Methods Based on Ranks.* San Francisco: Holden-Day, Inc.

- Lehmann, E.L. (1986) *Testing Statistical Hypotheses. Second Edition.* New York: John Wiley & Sons, Inc.
- Lehmann, E.L. and Romano, J.P. (2005) *Testing Statistical Hypotheses. Third Edition.* New York: Springer.
- Lindley, D.V. (1970) *Introduction to Probability and Statistics from a Bayesian Viewpoint. Part 2: Inference.* Cambridge: Cambridge University Press.
- Little, R.J.A. (1989) Testing the equality of two independent binomial proportions. *The American Statistician* 43: 283–8.
- Liu, J.P. and Chow, S.C. (1992) On the assessment of variability in bioavailability/bioequivalence studies. *Communications in Statistics: Theory and Methods* 21: 2591–607.
- Lloyd, C.J. (2008a) A new exact and more powerful unconditional test of no treatment effect from binary matched pairs. *Biometrics* 64: 716–23.
- Lloyd, C.J. (2008b) Exact P-values for discrete models obtained by estimation and maximisation. *Australian & New Zealand Journal of Statistics* 50: 329–45.
- Lloyd, C.J. and Moldovan, M.V. (2008) A more powerful exact test of non-inferiority from binary matched pairs data. *Statistics in Medicine* 27: 3540–9.
- Lu, Y. and Bean, J.A. (1995) On the sample size for one-sided equivalence of sensitivities based upon McNemar's test. *Statistics in Medicine* 14: 1831–9.
- MacLeod, M.J., Dahiyat, M.T., Cumming, A.M., Meiklejohn, D., Shaw, D. and St Clair, D. (1999) No association between Glu/Asp polymorphism of NOS3 gene and ischemic stroke. *Neurology* 53: 418–20.
- Makuch, R.W. and Simon, R. (1978) Sample size requirements for evaluating a conservative therapy. *Cancer Treatment Reports* 62: 1037–40.
- Mandallaz, D. and Mau, J. (1981) Comparison of different methods for decision-making in bioequivalence assessment. *Biometrics* 37: 213–22.
- Martín, A.A. (1998) Fisher's exact and Barnard's tests. In S. Kotz, C. Reid and D.L. Banks, eds., *Encyclopedia of Statistical Sciences*, vol. 2. New York: John Wiley & Sons, Inc., 250–8.
- Mau, J. (1987) On Cox's confidence distribution. In R. Viertl, ed., *Probability and Bayesian Statistics*. New York: Plenum., 346–56.
- Mau, J. (1988) A statistical assessment of clinical equivalence. *Statistics in Medicine* 7: 1267–77.
- McDonald, L.L., Davis, B.M., Bauer, H.R. and Laby, B. (1981) Algorithm

- AS161: critical regions of an unconditional non-randomized test of homogeneity in  $2 \times 2$  contingency tables. *Applied Statistics* 30: 182–9.
- McDonald, L.L., Davis, B.M. and Milliken, G.A. (1977) A nonrandomized unconditional test for comparing two proportions in a  $2 \times 2$  contingency table. *Technometrics* 19: 145–50.
- McNemar, I. (1947) Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12: 153–7.
- Mehring, G. (1993) On optimal tests for general interval hypotheses. *Communications in Statistics, Theory and Methods* 22: 1257–97.
- Mehta, C.R., Patel, N.R. and Tsiatis, A.A. (1984) Exact significance testing to establish treatment equivalence with ordered categorical data. *Biometrics* 40: 819–25.
- Meredith, M.P. and Heise, M.A. (1996) Comment on “Bioequivalence trials, intersection-union tests and equivalence confidence sets.” *Statistical Science* 11: 304–6.
- Metzler, C.M. (1974) Bioavailability — a problem in equivalence. *Biometrics* 30: 309–17.
- Miettinen, O. and Nurminen, M. (1985) Comparative analysis of two rates. *Statistics in Medicine* 4: 213–26.
- Miller, K.W., Williams, D.S., Carter, S.B., Jones, M.B. and Mishell, D.R. (1990) The effect of olestra on systemic levels of oral contraceptives. *Clinical Pharmacology and Therapeutics* 48: 34–40.
- Morgan, W.A. (1939) A test for the significance of the difference between the two variances in a sample from a normal bivariate population. *Biometrika* 31: 13–9.
- Morikawa, T., Yanagawa, T., Endou, A. and Yoshimura, I. (1996) Equivalence tests for pair-matched binary data. *Bulletin of Informatics and Cybernetics* 28: 31–45.
- Moses, L.E. (1953) Non-parametric methods. In H. Walker and J. Lev, eds., *Statistical Inference*. New York: Holt, Rinehart and Winston, 426–50.
- Müller-Cohrs, J. (1988) Size and power of the Anderson-Hauck test and the double t-test in equivalence problems. In *Proceedings of the SAS*. Copenhagen: European Users Group Internations Conference., 30–8.
- Munk, A. (1993) An improvement on commonly used tests in bioequivalence assessment. *Biometrics* 49: 1225–30.
- Munk, A. (1996) Equivalence and interval testing for Lehmann's alternative. *Journal of the American Statistical Association* 91: 1187–96.

- Nam, J. (1998) Power and sample size for stratified prospective studies using the score method for testing relative risk. *Biometrics* 54: 331–6.
- Nam, J. (2006) Non-inferiority of new procedure to standard procedure in stratified matched-pair design. *Biometrical Journal* 48: 966–77.
- Nam, J. and Blackwelder, W.C. (2002) Analysis of the ratio of marginal probabilities in a matched-pair setting. *Statistics in Medicine* 21: 689–99.
- Neyman, J. and Pearson, E.S. (1936) Contributions to the theory of testing statistical hypotheses I. Unbiased critical regions of type A and type A1. *Statistical Research Memoirs* 1: 1–37.
- Patterson, S. and Jones, B. (2005) *Bioequivalence and Statistics in Clinical Pharmacology*. Boca Raton, FL: Chapman & Hall/CRC Press.
- Pearson, E.S. (1947) The choice of statistical tests illustrated on their interpretation of data classified in a  $2 \times 2$  table. *Biometrika* 34: 139–67.
- Perlman, M.D. and Wu, L. (1999) The emperor's new tests. *Statistical Science* 14: 355–69.
- Pfeiffer, N., Grehn, F., Wellek, S. and Schwenn, O. (1994) Intraocular thymoxamine or acetylcholine for the reversal of mydriasis. *German Journal of Ophthalmology* 3: 422–6.
- Philipp, T., Anlauf, M., Distler, A., Holzgreve, H., Michaelis, J. and Wellek, S. (1997) Randomised, double blind multicentre comparison of hydrochlorothiazide, atenolol, nitrendipine, and enalapril in antihypertensive treatment: Results of the HANE study. *British Medical Journal* 315: 154–9.
- Pitman, E.J.G. (1939) A note on normal correlation. *Biometrika* 31: 9–12.
- Pratt, J.W. (1960) On interchanging limits and integrals. *Annals of Mathematical Statistics* 31: 74–7.
- Racine-Poon, A., Grieve, A.P., Flühler, H. and Smith, A.F.M. (1987) A two-stage procedure for bioequivalence studies. *Biometrics* 43: 847–56.
- Randles, R.H. and Wolfe, D.A. (1979) *Introduction to The Theory of Non-parametric Statistics*. New York: John Wiley & Sons, Inc.
- Rao, C.R. (1973) *Linear Statistical Inference and Its Applications. Second Edition*. New York: John Wiley & Sons, Inc.
- Reich, K., Mössner, R., König, I.R., Westphal, G., Ziegler, A. and Neumann, C. (2002) Promoter polymorphisms of the genes encoding tumor necrosis factor- $\alpha$  and interleukin-1 $\beta$  are associated with different subtypes of psoriasis characterized by early and late disease onset. *The Journal of Investigative Dermatology* 118: 155–63.

- Richard, F., Fromentin-David, I., Ricolfi, F., Ducimetière, P., Di Menza, C., Amouyel, P. and Helbecque, N. (2001) The angiotensin I converting enzyme gene as a susceptibility factor for dementia. *Neurology* 56: 1593–5.
- Rodary, C., Com-Nougue, C. and Tournade, M.F. (1989) How to establish equivalence between treatments: A one-sided clinical trial in paediatric oncology. *Statistics in Medicine* 8: 593–8.
- Rodda, B.E. (1990) Bioavailability: Designs and analysis. In D.A. Berry, ed., *Statistical Methodology in the Pharmaceutical Sciences*. New York: Marcel Dekker Inc, 57–81.
- Rodda, B.E. and Davis, R.L. (1980) Determining the probability of an important difference in bioavailability. *Clinical Pharmacology and Therapeutics* 28: 247–52.
- Roebruck, P. and Kühn, A. (1995) Comparison of tests and sample size formulae for proving therapeutic equivalence based on the difference of binomial probabilities. *Statistics in Medicine* 14: 1583–94.
- Röthmel, J. (2005) Problems with existing procedures to calculate exact unconditional p-values for non-inferiority/superiority and confidence intervals for two binomials and how to resolve them. *Biometrical Journal* 47: 37–47.
- Röthmel, J. and Mannsmann, U. (1999a) Unconditional non-asymptotic one-sided tests for independent binomial proportions when the interest lies in showing non-inferiority and/or superiority. *Biometrical Journal* 41: 149–70.
- Röthmel, J. and Mannsmann, U. (1999b) Letter to the editor on “Exact tests of equivalence and efficacy with a non-zero lower bound for comparative studies” by S.F. Chan. *Statistics in Medicine* 18: 1734–5.
- Romano, J.P. (2005) Optimal testing of equivalence hypotheses. *Annals of Statistics* 33: 1036–47.
- Roy, S.N. (1953) On a heuristic method of test construction and its use in multivariate analysis. *Annals of Mathematical Statistics* 24: 220–38.
- Sasieni, P.D. (1997) From genotype to genes: Doubling the sample size. *Biometrics* 53: 1253–61.
- Savage, I.R. (1956) Contributions to the theory of rank order statistics - the two-sample case. *Annals of Mathematical Statistics* 27: 590–615.
- Scaglione, F. (1990) Comparison of the clinical and bacteriological efficacy of clarithromycin and erythromycin in the treatment of streptococcal pharyngitis. *Current Medical Research and Opinion* 12: 25–33.
- Schall, R. (1995) Assessment of individual and population bioequivalence us-

- ing the probability that bioavailabilities are similar. *Biometrics* 51: 615–26.
- Schall, R. and Luus, H.G. (1993) On population and individual bioequivalence. *Statistics in Medicine* 12: 1109–24.
- Schoenfeld, D. (1981) The asymptotic properties of nonparametric tests comparing survival distributions. *Biometrika* 68: 316–9.
- Schuirmann, D.J. (1987) A comparison of the two one-sided tests procedure and the power approach for assessing equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics* 15: 657–80.
- Schuirmann, D.J. (1996) Comment on “Bioequivalence trials, intersection-union tests and equivalence confidence sets.” *Statistical Science* 11: 312–3.
- Schumann, G., Rujescu, D., Szegedi, A., Singer, P., Wiemann, S., Wellek, S., Giegling, I., Klawe, C., Anghelescu, I., Heinz, A., Spanagel, R., Mann, K., Henn, F.A. and Dahmen, N. (2003) No association of alcohol dependence with a NMDA-receptor 2B gene variant. *Molecular Psychiatry* 8: 11–2.
- Searle, S.R. (1987) *Linear Models for Unbalanced Data*. New York: John Wiley & Sons, Inc.
- Selwyn, M.R., Dempster, A.P. and Hall, N.R. (1981) A Bayesian approach to bioequivalence for the  $2 \times 2$  changeover design. *Biometrics* 37: 11–21.
- Selwyn, M.R. and Hall, N.R. (1984) On Bayesian methods for bioequivalence. *Biometrics* 40: 1103–8.
- Sheiner, L.B. (1992) Bioequivalence revisited. *Statistics in Medicine* 11: 1777–88.
- Shorack, R. (1967) Tables of the distribution of the Mann-Whitney-Wilcoxon U-statistic under Lehmann alternatives. *Technometrics* 9: 666–77.
- Sidik, K. (2003) Exact unconditional tests for testing non-inferiority in matched-pairs design. *Statistics in Medicine* 22: 265–78.
- Starks, T.H. (1979) An improved sign test for experiments in which neutral responses are possible. *Technometrics* 21: 525–30.
- Steinijans, V.W. and Diletti, E. (1983) Statistical analysis of bioavailability studies: Parametric and nonparametric confidence intervals. *European Journal of Clinical Pharmacology* 24: 127–36.
- Steinijans, V.W. and Diletti, E. (1985) Generalisation of distribution-free confidence intervals for bioavailability ratios. *European Journal of Clinical Pharmacology* 28: 85–8.
- Stevens, W.L. (1938) Estimation of blood-group gene frequencies. *Annals of*

- Eugenics 8: 362–75.
- Stuart, A. and Ord, K. (1994) *Kendall's Advanced Theory of Statistics. Volume 1 - Distribution Theory. Sixth Edition*. London: Edward Arnold.
- Suissa, S. and Shuster, J.J. (1985) Exact unconditional sample sizes for the  $2 \times 2$  binomial trial. *Journal of the Royal Statistical Society, Series A* 148: 317–27.
- Tango, T. (1998) Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Statistics in Medicine* 17: 891–908.
- Ungersböck, K. and Kempski, O.S. (1992) *A study of the effect of increased intracranial pressure on the cortical microcirculation of New Zealand white rabbits*. Department of Neurosurgery, University of Mainz. Unpubl. Techn. Report.
- Victor, N. (1987) On clinically relevant differences and shifted nullhypotheses. *Methods of Information in Medicine* 26: 155–62.
- Wang, W. (1997) Optimal unbiased tests for equivalence in intrasubject variability. *Journal of the American Statistical Association* 92: 1163–70.
- Welch, B.L. (1938) The significance of the difference between means when the population variances are unequal. *Biometrika* 29: 350–62.
- Wellek, S. (1990) Vorschläge zur Reformulierung der statistischen Definition von Bioäquivalenz. In G. Giani and R. Repges, eds., *Biometrie und Informatik - Neue Wege zur Erkenntnisgewinnung in der Medizin*. Berlin: Springer-Verlag, 95–9.
- Wellek, S. (1991) Zur Formulierung und optimalen Lösung des Bioäquivalenznachweis-Problems in der klassischen Theorie des Hypothesentestens. In J. Vollmar, ed., *Bioäquivalenz sofort freisetzender Arzneiformen*. Stuttgart: Gustav Fischer Verlag, 91–109.
- Wellek, S. (1993a) Basing the analysis of comparative bioavailability trials on an individualized statistical definition of equivalence. *Biometrical Journal* 35: 47–55.
- Wellek, S. (1993b) A log-rank test for equivalence of two survivor functions. *Biometrics* 49: 877–81.
- Wellek, S. (1994) *Statistische Methoden zum Nachweis von Äquivalenz*. Stuttgart: Gustav Fischer Verlag.
- Wellek, S. (1996) A new approach to equivalence assessment in standard comparative bioavailability trials by means of the Mann-Whitney statistic. *Biometrical Journal* 38: 695–710.

- Wellek, S. (2000a) Bayesian construction of an improved parametric test for probability-based individual bioequivalence. *Biometrical Journal* 42: 1039–52.
- Wellek, S. (2000b) On a reasonable disaggregate criterion on population bioequivalence admitting of resampling-free testing procedures. *Statistics in Medicine* 19: 2755–67.
- Wellek, S. (2004) Tests for establishing compatibility of an observed genotype distribution with Hardy-Weinberg equilibrium in the case of a biallelic locus. *Biometrics* 60: 694–703.
- Wellek, S. (2005) Statistical methods for the analysis of two-arm non-inferiority trials with binary outcomes. *Biometrical Journal* 47: 48–61.
- Wellek, S. and Michaelis, J. (1991) Elements of significance testing with equivalence problems. *Methods of Information in Medicine* 30: 194–8.
- Wellek, S., Ziegler, A. and Goddard, K.A.B. (2009) A confidence-limit-based approach to the assessment of Hardy-Weinberg equilibrium. In *Technical Report, 2008/9 - 2*. Department of Biostatistics, CIMH Mannheim.
- Westlake, W.J. (1972) Use of confidence intervals in analysis of comparative bioavailability trials. *Journal of Pharmacological Sciences* 61: 1340–1.
- Westlake, W.J. (1976) Symmetrical confidence intervals for bioequivalence trials. *Biometrics* 32: 741–4.
- Westlake, W.J. (1979) Statistical aspects of comparative bioavailability trials. *Biometrics* 35: 273–80.
- Westlake, W.J. (1981) Comments on “Bioequivalence testing – A need to rethink.” *Biometrics* 37: 591–3.
- Wiens, B.L. and Iglewicz, B. (1999) On testing equivalence of three populations. *Journal of Biopharmaceutical Statistics* 9: 465–83.
- Wiens, B.L. and Iglewicz, B. (2000) Design and analysis of three treatment equivalence trials. *Controlled Clinical Trials* 21: 127–37.
- Willems, J.L., Abreu-Lima, C., Arnaud, P., Brohet, C.R., Denis, B., Gehring, J., Graham, I., van Herpen, G. and Machado, H. (1990) Evaluation of ECG interpretation results obtained by computer and cardiologists. *Methods of Information in Medicine* 29: 308–16.
- Windeler, J. and Trampisch, H.J. (1996) Recommendations concerning studies on therapeutic equivalence. *Drug Information Journal* 30: 195–200.
- Witting, H. (1985) *Mathematische Statistik I*. Stuttgart: B.G. Teubner.
- Wittke-Thompson, J.K., Pluzhnikov, A. and Cox, N.J. (2005) Rational inferences about departures from Hardy-Weinberg equilibrium. *American*

- Journal of Human Genetics* 76: 967–86.
- Ziegler, A. and König, I.R. (2006) *A Statistical Approach to Genetic Epidemiology*. Weinheim: Wiley-VCH.
- Zinner, D.D., Duany, L.F. and Chilton, N.W. (1970) Controlled study of the clinical effectiveness of a new oxygen gel on plaque, oral debris and gingival inflammation. *Pharmacology and Therapeutics in Dentistry* 1: 7–15.