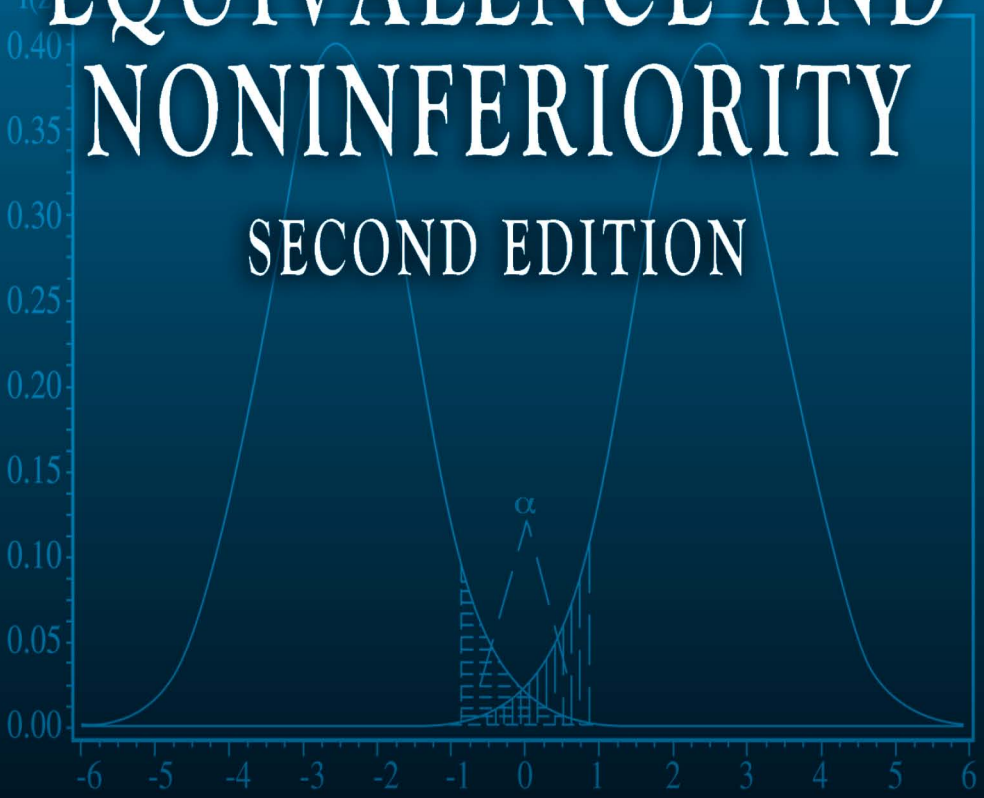
TESTING STATISTICAL HYPOTHESES OF EQUIVALENCE AND NONINFERIORITY

SECOND EDITION



STEFAN WELLEK

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

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

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

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-13: 978-1-4398-0819-1 (Ebook-PDF)

To Brigitte

Contents

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 `WKTSHEQ2 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.

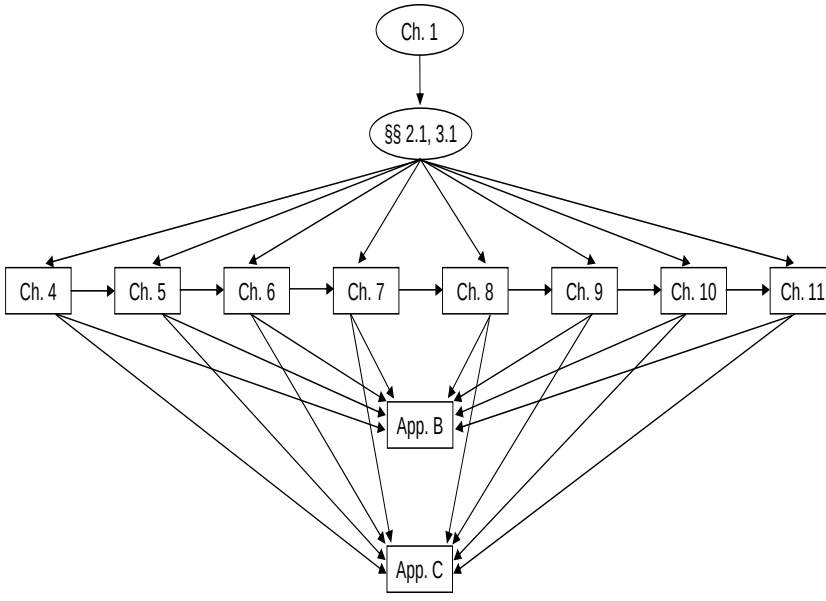
The book can be used in several ways depending on the reader's interests and level of statistical background. Chapter 1 gives a general introduction to the topic and should at least be skimmed by readers of any category. Chapters 2 and 3 deal with general approaches to problems of testing for noninferiority 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).

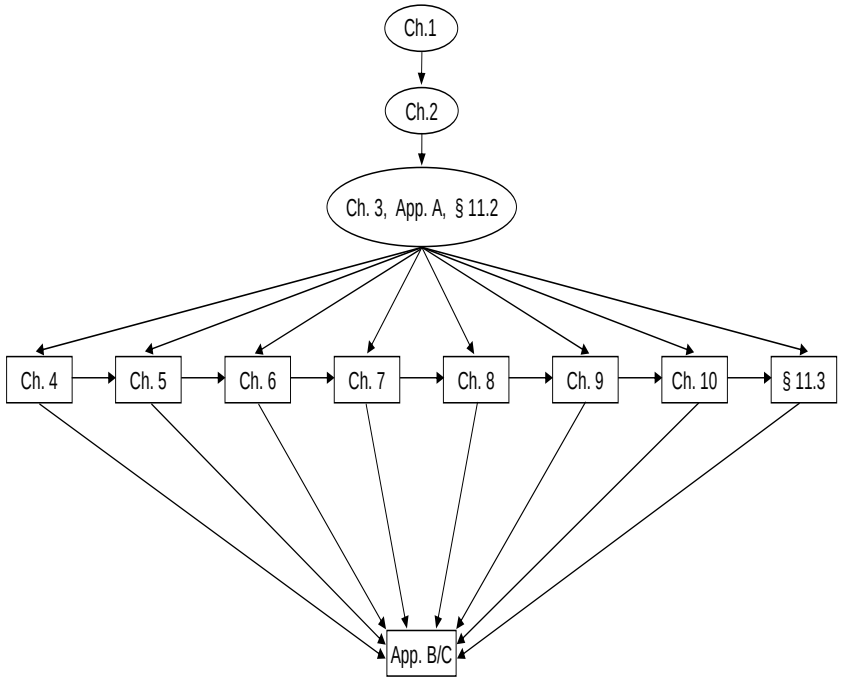
Apart from occasional cross-references, all remaining chapters (and even individual sections of them) can be read independently of each other. The material they contain is to provide the working statistician of any background and level of sophistication with a sufficiently rich repertoire of efficient 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 LaTeX 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 TeX-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/GRAPH. 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
   Web: 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 hypertensiology doubt the clinical relevance of existing data showing the *bio*equivalence 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 hypertensiologists 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 $p_\circ = .60$, versus the two-sided alternative $p \neq p_\circ$. 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 $p_\circ = .60$ 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 $p_\circ$ for considering the deviations practically irrelevant.* For the moment, let us specify this region as the interval $(p_\circ - .10\,,\, p_\circ + .10) = (.50\,,\, .70)$. Hence, by equivalence of p to $p_\circ$ 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 $p_\circ = .60$, the stronger the evidence in favor of equivalence provided by the available data. Thus, a reasonable test for equivalence of p to $p_\circ$ will use a decision rule of the following form: The *null hypothesis of inequivalence is rejected if and only if* the difference $X/n - p_\circ$ *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 $p_\circ$ is as weak as possible without increasing the risk of an erroneous equivalence decision over α, 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 $p_\circ \mp .10$ as the *theoretical* range of equivalence and at level $\alpha = 5\%$, the optimal critical bounds to $X/n - p_\circ$ to be used in a test for equivalence of p to $p_\circ = .60$ 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 $X/n - p_\circ$ 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 $X/n - p_\circ = -4.0\%$. 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 $|p - p_\circ| < .10$.

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 biovailability 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 *bio*equivalence 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 [→ §§ 6.5, 7.4]
- additivity of main effects [→ § 9.3.1]
- identity of carryover effects in crossover trials [→ § 9.3.2].

Establishing goodness of fit by means of equivalence testing procedures is even an important issue in genetic epidemiology. This will will 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/nonferiority 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.

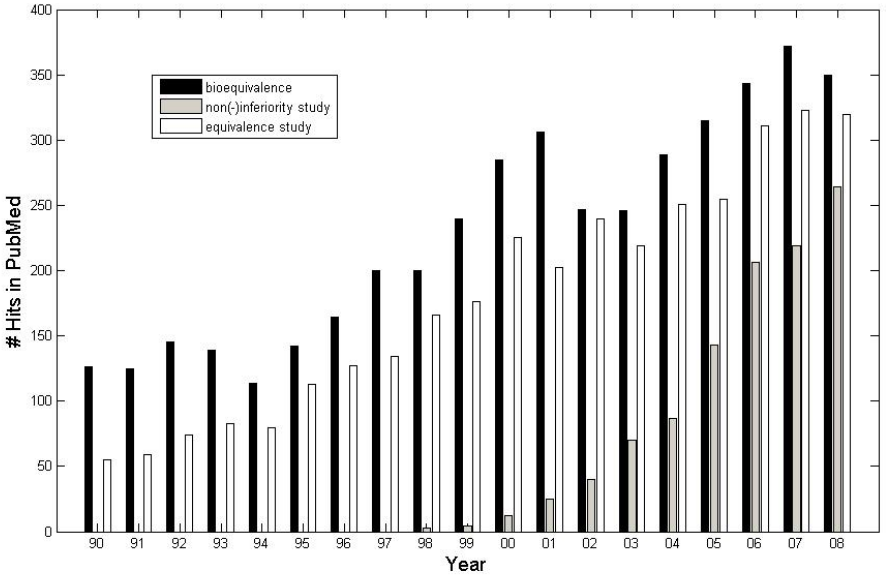


Figure 1.1 *Count of hits obtained by entering the keywords (i) bioequivalence, (ii) non(-)inferiority study, and (iii) equivalence study in PubMed, by year of publication.*

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 millenium, 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 θ 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 μ_1 and μ_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 θ is contained in a suitable neighborhood around some reference value $\theta_\circ$ taken on by θ if and only if the distributions under comparison are exactly equal. This neighborhood comprises those values of θ whose distance from $\theta_\circ$ 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 $\theta_\circ - \varepsilon_1$ and $\theta_\circ + \varepsilon_2$, respectively. Of course, both ε_1 and ε_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 $\theta_\circ$ is $\theta_\circ = 0$, and the equivalence interval is frequently chosen symmetrical about $\theta_\circ$, 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*

$$H : \theta \leq \theta_\circ - \varepsilon_1 \quad \text{or} \quad \theta \geq \theta_\circ + \varepsilon_2 \tag{1.1a}$$

of nonequivalence, versus the *equivalence assumption*

$$K : \theta_\circ - \varepsilon_1 < \theta < \theta_\circ + \varepsilon_2 \tag{1.1b}$$

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 α. 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 $H_0 : \theta = \theta_\circ$ of the corresponding two-sided testing problem, irrespective of what particular positive values are assigned to the constants ε_1 and ε_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 $\{\theta_\circ\}$ 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 θ. But continuity of the power function $\theta \mapsto \beta(\theta)$, say, clearly implies, that the test can maintain level α on $\{\theta \neq \theta_\circ\}$ only if its power against $\theta = \theta_\circ$ exceeds α neither. Consequently, if we tried to test the null hypothesis $\theta \neq \theta_\circ$ against the alternative $\theta = \theta_\circ$, we would not be able to replace the trivial "test" rejecting the null hypothesis independently of the data with probability α, 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 θ within which the difference between the actual value of θ and its reference value $\theta_\circ$ 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 θ and $\theta_\circ$ is now bounded to below only whereas excesses in value of θ over $\theta_\circ$ 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 $\theta_\circ + \varepsilon_2$ of the equivalence interval is replaced with $+\infty$ or, in cases where the parameter space Θ of θ is bounded to the right itself, by $\theta^* = \sup\Theta$. The corresponding hypothesis testing problem reads

$$H_1 : \theta \leq \theta_\circ - \varepsilon \quad \text{versus} \quad K_1 : \theta > \theta_\circ - \varepsilon \tag{1.2}$$

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 θ keeps denoting a meaningful measure for the extent of superiority of a new treatment A over some standard treatment B and $\theta = \theta_\circ$ indicates identity in effectiveness of both treatments, testing $\theta \leq \theta_\circ + \varepsilon$ versus $\theta > \theta_\circ + \varepsilon$ rather than $\theta \leq \theta_\circ - \varepsilon$ versus $\theta > \theta_\circ - \varepsilon$ 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 θ [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 δ and ρ as the difference $p_1 - p_2$ and the odds ratio $p_1(1-p_2)/\big((1-p_1)p_2\big)$, 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 δ, 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, δ), the range of δ-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.

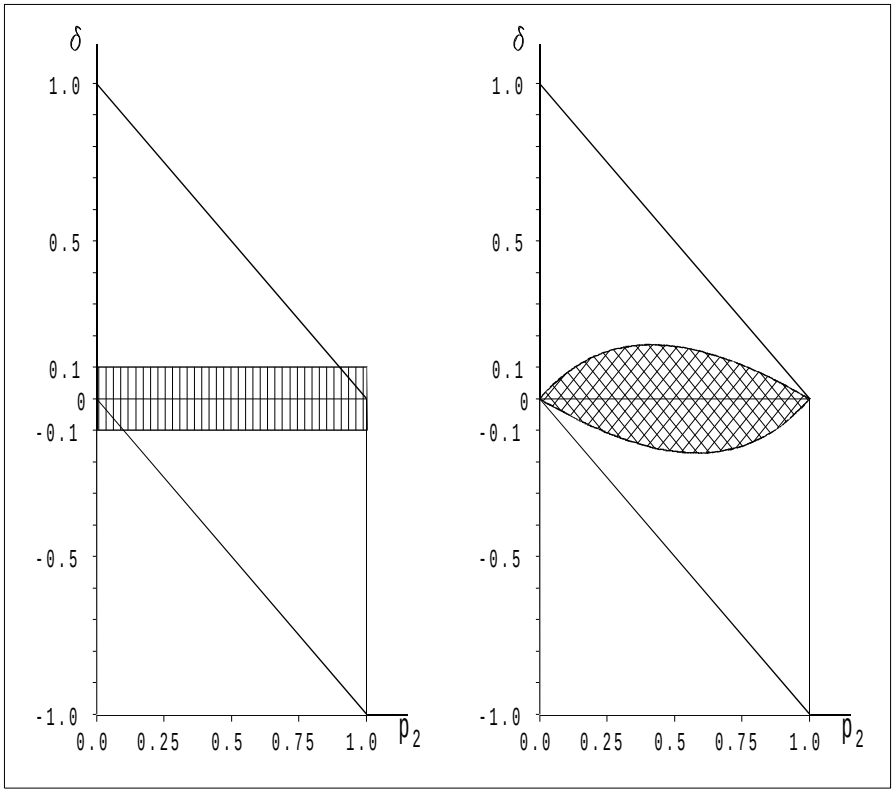


Figure 1.2 *Equivalence hypotheses in terms of the difference between both responder rates [left] and the odds ratio [right] as regions in the $p_2 \times \delta$-plane with $\delta = p_1 - p_2$. [Rhomboid $\widehat{=}$ set of possible values of (p_2, δ).]*

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 μ_1, μ_2 and σ, 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 σ is large enough, whereas the areas under the corresponding densities are next to disjoint if σ 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 $\theta_\circ - \varepsilon_1$, $\theta_\circ + \varepsilon_2$ 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 $\theta_\circ$. 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 $\theta_\circ - \varepsilon_1 = 2\log(.80) \approx -.446$, $\theta_\circ + \varepsilon_2 = 2\log(1.20) \approx .365$ 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 symmetricized 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 ε: Strict Choice	Liberal 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 t-Test	δ/σ	0	.25	.50
(v)	Two-Sample t-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

$^{\dagger)}$ $D \equiv$ intraindividual difference for a randomly chosen observational unit
$^{\ddagger)}$ $X, Y \equiv$ 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:

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

where Φ^{-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:

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

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 ε, 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.

2

General techniques for dealing with noninferiority problems

2.1 Standard solution in the case of location parameter families

Whenever the target parameter θ 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 $\theta \leq \theta_\circ$ versus $\theta > \theta_\circ$ 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 θ in the location of the distribution of D away from $\theta_\circ = 0$ 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) $F_\circ : \mathbb{R} \to [0, 1]$ symmetric about zero. For the time being, we do not specify whether the baseline cdf has some known form (e.g., $F_\circ = \Phi(\cdot/\sigma)$ with Φ 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 $(X_1, Y_1), \ldots, (X_n, Y_n)$ 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 $(D_1, \ldots, D_n)$ is independent and identically distributed (iid) ;

(b) $P[D_i \leq d] = F_\circ(d - \theta)$ for arbitrary $z \in \mathbb{R}$ and all $i = 1, \ldots, n$, with $F_\circ(-d) = 1 - F_\circ(d)\ \forall d$ and θ denoting the parameter of interest. (By convention, positive values of θ 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 α for the traditional one-sided testing problem $H_1^\circ : \theta \leq 0$ versus $K_1^\circ : \theta > 0$ [$\rightarrow$ (1.2), specialized to the case that $\theta_\circ = 0$, $\varepsilon = 0$] is well known to be given by

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

where $T(\cdot)$ denotes a suitable real-valued function of n arguments (usually called the test statistic), and c_α the upper 100α percentage point of the distribution of the random variable $(D_1, \ldots, D_n)$ 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 $(\tilde{D}_1, \ldots, \tilde{D}_n)$ obviously satisfies again (a) and (b), provided the parameter θ 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 $H_1 : \theta \leq -\varepsilon$ [$\rightarrow$ relevant inferiority] if and only if we find that $T(D_1 + \varepsilon, \ldots, D_n + \varepsilon) > c_\alpha$ where $T(\cdot)$ and c_α 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 (300μg) and ethinyl estradiol (30μg) 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* [$\tilde{D}_i = D_i + \varepsilon$; $\tilde{R}_i^+$ = *rank of the ith 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 θ 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 $i = 1, \ldots, 28$, 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 $X_1, \ldots, X_m, Y_1, \ldots, Y_n$. 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 $X_i \stackrel{\mathrm{d}}{=} Y_j + \theta$ where, as usual (cf. Randles and Wolfe, 1979, §1.3), the symbol "$\stackrel{\mathrm{d}}{=}$" 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 $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ 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 θ 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 $T(X_1, \ldots, X_m, Y_1, \ldots, Y_n) > c_\alpha$, where the test statistic $T(\cdot)$ is a real-valued function of $m + n$ arguments and c_α stands for the upper 100α percentage point of the distribution of $T(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ 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 $i = 1, \ldots, m$) and $\tilde{Y}_j = Y_j$ (for $j = 1, \ldots, n$), 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 θ_a of the shift parameter to which the proposed hypothesis is referring. Then, the power of the test for noninferiority with margin ε 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 β under this alternative, is likewise obtained by replacing θ_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

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

and

$$\mathrm{POW}_0 = 1 - G^*_{\sqrt{mn/N}\varepsilon/\sigma}(t_{N-2;1-\alpha}) \tag{2.1b}$$

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, σ_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 $\mathbf{X} = (X_1, \ldots, X_m, Y_1, \ldots, Y_n)$]. 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 θ of interest (also called a strictly totally positive family of order 2 or STP_2 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 ε, there is a test for $H_1 : \theta \leq \theta_\circ - \varepsilon$ versus $K_1 : \theta > \theta_\circ - \varepsilon$ which is uniformly most powerful among all tests at the same significance level α depending on the data only through T. The rejection region of this test is given by

$$\left\{ \mathbf{x} \,|\, T(\mathbf{x}) > k \right\} \ . \tag{2.2}$$

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 $\theta = \theta_\circ - \varepsilon$. In the continuous case, the optimal k is obtained by solving the equation

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

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, γ) of real numbers the second of which belongs to the half-open unit interval $[0\,,\,1)$ such that there holds

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

For $\gamma > 0$, the exact level-α 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 ruled out for continuously distributed $T(\mathbf{X})$, the null hypothesis H_1 of (relevant) inferiority has to be rejected [accepted] with probability γ $[1-\gamma]$. 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_2-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 $P_{\theta_\circ - \varepsilon}[\,\cdot\,]$ 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-α 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