Springer Series in Statistics

# Sequential Experimentation in Clinical Trials

## Design and Analysis

Springer

# Springer Series in Statistics

For further volumes:
http://www.springer.com/series/692

Jay Bartroff • Tze Leung Lai • Mei-Chiung Shih

# Sequential Experimentation in Clinical Trials

## Design and Analysis

Springer

Jay Bartroff
Department of Mathematics
University of Southern California
Los Angeles, California, USA

Tze Leung Lai
Department of Statistics
and Cancer Institute
Stanford University
Stanford, California, USA

Mei-Chiung Shih
Department of Health Research and Policy
Stanford University
VA Cooperative Studies Program
Stanford, California, USA

ISSN 0172-7397
ISBN 978-1-4614-6113-5 ISBN 978-1-4614-6114-2 (eBook)
DOI 10.1007/978-1-4614-6114-2
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2012951627

Printed on acid-free paper

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

*To our parents*

*Jack and Barbara Bartroff*

*Chi Y. and Wai C. Lai*

*Ming-Tsan Shih and Min-Hui Hsu*

# Preface

The idea of writing a book on sequential experimentation in clinical trials arose 25 years ago when Lai was at Columbia University, collaborating with Dan Anbar of Abbott Laboratories on a university-industry cooperative research project, “Sequential Statistical Methods in Biopharmaceutical Research,” funded by the National Science Foundation. Anbar and Lai, together with Gordon K.K. Lan and Anastasio Tsiatis (at that time, at the NIH and Harvard, respectively), formed a focused research group that held a week-long meeting every 2 months and organized an annual workshop, with invited speakers from academia, industry, and the FDA and NIH, and dedicated to the development and discussion of sequential methods in the design and analysis of clinical trials. Although substantial progress was made by the group to advance this new area that attracted considerable attention from the pharmaceutical industry after the early termination of the Beta-Blocker Heart Attack Trial in 1981, the book project could not materialize when Lai moved to the West Coast in 1987, joining Stanford University, while the other collaborators remained on the East Coast but were busy with their own moves to new positions. On the other hand, the group members continued their separate research efforts in this area. These efforts and those by other researchers led to major advances and eventual widespread use of group sequential designs and interim analysis methods by the pharmaceutical industry and their acceptance by the FDA.

At the turn of the new century, the monograph by Jennison and Turnbull (2000) appeared, giving a comprehensive overview of group sequential methods developed up to that time. Besides continual developments in interim analysis and group sequential methods, the past decade has also witnessed new developments and growing interest in adaptive designs of clinical trials. The books by Proschan et al. (2006), Chow and Chang (2006), Chang (2007), and Berry et al. (2011) describe some of these developments and their applications. However, as pointed out in Chap. 8, there is substantial disagreement in the literature concerning the appropriateness of these adaptive designs, which either use inefficient test statistics that are not supported by mainstream statistical principles to adjust for the adaptation in maintaining the Type I error of the test or use Bayesian posterior probabilities that do not guarantee the prescribed Type I error. Chapter 8 describes our recent work that provides a new class of adaptive designs which are both flexible and efficient, thereby resolving the dilemma between efficiency and flexibility in the adaptive design literature. Prior to this work, we have also developed a comprehensive methodology of flexible and efficient group sequential designs, to which Chap. 4 is devoted. In fact, the new adaptive designs in Chap. 8 are modifications of the corresponding group sequential designs in Chap. 4, and a unified approach is provided for the methodology and implementation of group sequential and adaptive designs.

Besides giving an up-to-date account of these flexible designs, we also present in Chap. 7 a comprehensive overview, including the most recent developments of inference after the termination of these clinical trials. Chapter 6 describes the Beta-Blocker Heart Attack Trial as an example for the design and analysis of clinical trials with failure-time endpoints and interim analyses. The material in Chaps. 4, 6, 7, and 8 can be used for short courses on group sequential and adaptive designs. We have given short courses based on this material in the First Joint Biostatistics Symposium in Beijing, July 2010, the Applied Statistics Symposium of the International Chinese Statistical Association in New York, June 2011, and the Workshop on the Design and Analysis of Clinical Trials at National University of Singapore, October 2011. We were greatly encouraged by the enthusiastic response and stimulating comments of the participants.

This book has also benefited from the Third International Workshop in Sequential Methodologies at Stanford University, June 2011. The workshop was very well attended and was truly international in nature. There it was pointed out that despite a resurgence of interest in sequential analysis, the subject was not in the graduate curriculum of most statistics departments. One reason that was mentioned was the lack of textbooks that could present the material in an appealing way to today's graduate students. In fact, only a handful of such books had been written and they were published more than 20 years ago. Although there are more recent books which we have mentioned in the second paragraph of this preface, they all deal with the specialized topics of group sequential and adaptive designs rather than general methods and principles in sequential analysis. Another reason that came up during workshop discussions was that sequential methods and adaptive designs seemed to involve special techniques and ideas that are detached from mainstream topics taught in the modern graduate statistics curriculum, e.g., likelihood inference, regression analysis, resampling, semiparametric theory, to name a few. Spurred by these comments, we have made particular efforts to change this perception in the selection and presentation of the materials. To make it suitable for an introductory course on sequential analysis, the book covers the much broader subject of sequential experimentation that includes group sequential and adaptive designs of Phase II and III clinical trials, which have attracted much attention in the past three decades. In particular, the broad scope of design and analysis problems in sequential experimentation clearly requires a wide range of statistical methods and models from nonlinear regression analysis, experimental design, dynamic programming, survival analysis, resampling, and likelihood and Bayesian inference. The background material in these building blocks is summarized in Chaps. 2 and 3 and certain sections in Chaps. 6 and 7. Besides group sequential tests and adaptive designs, we also introduce sequential change-point detection methods in Chap. 5 in connection with pharmacovigilance and public health surveillance. Together with dynamic programming and approximate dynamic programming in Chap. 3, the book therefore covers all basic topics for a graduate course in sequential analysis.

Different parts of the book can be used for short courses on clinical trials, translational medical research, and sequential experimentation. Lai has used an early draft of the book to teach a course on innovative clinical trial designs and statistical methods for second-year Ph.D. students in the Department of Statistics at Stanford University. The course has led to supplements and exercises for various chapters and also to the web site for the book, http://meichiun.web.stanford.edu/clinicaltrials/, to which different parts of the book refer for links to software.

We thank the teaching assistant Alex Deng and the students in the aforementioned course for their stimulating interest and helpful comments on a first draft of the book that was used as lecture notes. In particular, we want to mention Olivia Liao who subsequently wrote her Ph.D. thesis on the subject and contributed to some recent publications that we have referenced. We also thank our collaborators Philip Lavori, Zheng Su, Ray Zhu, Joe Heyse, Jie Chen, Sam Wong, and Hock Peng Chan for working with us on related projects and for their valuable suggestions. We are particularly grateful to Balasubramanian Narasimhan who has been helping us develop open-source software for the innovative designs and analyses that will be posted at the book's web site and to Cindy Kirby who has put together the separate chapters in an efficient and timely fashion. We also acknowledge grant support for the research projects related to this book by the National Science Foundation (DMS-0907241 and 0403105 for Bartroff and 1106535 and 0805879 for Lai), the National Institutes of Health (GMS-068968 for Bartroff and 1 P30 CA124435 for Lai), the U.S. Department of Veterans Affairs Cooperative Studies Program (for Shih), and the Stanford Center for Clinical and Translational Education and Research (for Lai and Shih).

Los Angeles, California Jay Bartroff
Stanford, California Tze Leung Lai
Stanford, California Mei-Chiung Shih

# Contents

# Chapter 1
# Introduction

This chapter gives an overview of (a) the prevalence of sequential experimentation in translational medical research and (b) developments of statistical methods to design and analyze these sequential experiments in evidence-based medical research. In this connection it also gives an outline of the topics covered in the subsequent chapters and discusses the complementary roles of Bayesian and frequentist approaches to sequential design and analysis.

## 1.1 Sequential Experimentation in Translational Medical Research

"From bench to bedside," a maxim of translational medical research, reflects the sequential nature of the experiments involved. "Bench" refers to laboratory experiments to study new biochemical principles and discover novel treatments. The experiments with promising results are followed by preclinical animal studies. After understanding the effect of the treatment (say, a new drug) on animals (e.g., rodents), the next stage of drug development consists of clinical trials that involve human subjects, starting with Phase I studies to determine a safe dose or dosage regimen and to collect information on the pharmacokinetics (PK) and pharmacodynamics (PD) of the drug. PK is concerned with the concentration versus time curve that is associated with the kinetics of drug absorption, distribution, and elimination. PD is concerned with the steady-state relationship of drug concentration at an effector site to the effect/response produced. The information collected and the dosage regimen determined from Phase I studies are used to design Phase II clinical trials to evaluate the efficacy of the drug for particular indications (endpoints) in patients with the disease. Phase II trials are precursors of Phase III trials whose goal is to demonstrate effectiveness of the drug for its approval by the regulatory agency (the Food and Drug Administration in the United States) and to provide adequate evidence for its labeling. Besides testing efficacy, Phase III trials also collect safety information

J. Bartroff et al., *Sequential Experimentation in Clinical Trials,*
Springer Series in Statistics 298, DOI 10.1007/978-1-4614-6114-2_1,

from the relatively large samples of patients accrued to the trial. The safety of the drug is evaluated from the data obtained from all three phases of clinical trials prior to marketing approval of the drug and continues to be evaluated through post-marketing Phase IV trials.

Despite the sequential nature of Phase I–III trials, the trials are often planned separately, treating each trial as an independent study whose design depends on results from studies in previous phases. An advantage of this is that the reproducibility of the results of the trial can be evaluated on the basis of the prescribed design, without worrying about the statistical variability of the results of earlier-phase trials that determine the prescribed design. A disadvantage lies in the fact that the sample sizes of the trials are often inadequate because of the separate planning. A different strategy is to expand a trial seamlessly from one phase into the next phase; the Phase II–III cancer trial design in Sect. 6.7 is an example. Although Phase II–III design, which is an active area of current research undergoing new advances, is beyond the scope of this book, we give a brief introduction in Sect. 6.7 to show the power of an overarching sequential experimentation approach to translational medicine.

This book focuses on sequential methods for the design and analysis of Phase I, II, and III clinical trials, thereby providing the background for understanding and developing the new advances. Although these methods are developed in the context of clinical trials, they are also applicable to other fields that involve sequential experimentation. We therefore give an introduction to the statistical methods and the underlying principles and also relate them to basic topics taught in typical graduate statistics programs that assume the data to be generated by nonsequential designs. For example, while Chap. 2 considers Phase I clinical trials, it starts with nonlinear regression and experimental design before relating them to basic pharmacologic principles and models underlying dose determination.

## 1.2 Sequential Analysis: From Weapons Testing to Confirmatory Clinical Trials

The subject named *sequential analysis*, which also includes sequential design of experiments, was born in response to demands for more efficient testing of antiaircraft gunnery during World War II, which led to Wald's development of the sequential probability ratio test (SPRT) in 1943 (Wallis 1980). Let $X_1, X_2, \ldots$ be i.i.d. random variables with common density function $f$. To test $H_0 : f = f_0$ versus $H_1 : f = f_1$, the SPRT stops sampling at stage

$$N = \inf\left\{ n \geq 1 : \prod_{i=1}^{n} \big(f_1(X_i)/f_0(X_i)\big) \notin (A, B) \right\}, \tag{1.1}$$

where $0 < A < 1 < B$ are the stopping boundaries. When stopping occurs, $H_0$ or $H_1$ is rejected according to whether the likelihood ratio $\prod_{i=1}^{N}(f_1(X_i)/f_0(X_i))$ crosses the upper boundary $B$ or the lower boundary $A$. In Chap. 3 we give a summary of the theory of sequential tests of hypotheses, beginning with the SPRT on testing a simple null versus a simple alternative hypothesis and describing important subsequent developments that led to a relatively complete theory for composite hypotheses.

Within a few years after Wald's introduction of the SPRT, it was recognized that sequential hypothesis testing might provide a useful tool in clinical trials to test the efficacy of new medical treatments. A number of papers appeared during the 1950s on modifications of the SPRT for the design of clinical trials, and an overview of these developments was given in Armitage (1960). In 1969, Armitage et al. proposed a new alternative to the SPRT and its variants, called the *repeated significance test* (RST). The underlying motivation for the RST is that, since the strength of evidence in favor of a treatment from a clinical trial is conveniently indicated by the results of a conventional significance test, it is appealing to apply the significance test, with nominal significance level $\alpha$, repeatedly during the trial. Noting that the overall significance level $\alpha^*$, which is the probability that the nominal significance level is attained at some stage, is larger than $\alpha$, they developed a recursive numerical algorithm to compute $\alpha^*$ in the case of testing a normal mean $\theta$ with known variance $\sigma^2$, for which the RST of $H_0 : \theta = 0$ is of the form

$$T = \inf\left\{n \le M : |S_n| \ge a\sigma\sqrt{n}\right\}, \tag{1.2}$$

rejecting $H_0$ if $T < M$ or if $T = M$ and $|S_M| \ge a\sigma\sqrt{M}$, where $S_n = X_1 + \cdots + X_n$. Haybittle (1971) proposed the following modification of the RST to increase its power. The stopping rule has the same form as (1.2) but the rejection region is modified to $T < M$ or $|S_M| \ge c\sigma\sqrt{M}$, where $a(\ge c)$ is so chosen that the overall significance level is equal to some prescribed number. In particular, $a = \infty$ gives the fixed sample size test while $a = c$ gives the RST.

In double-blind multicenter clinical trials, it is not feasible to arrange for continuous examination of the data as they accumulate to perform the RST. This led Pocock (1977) to introduce a "group sequential" version of (1.2), in which the $X_n$ represents an approximately normally distributed statistic of the data in the $n$th group (instead of the $n$th observation) and $M$ represents the maximum number of groups. Instead of the square-root boundary $a\sigma\sqrt{n}$, O'Brien and Fleming (1979) proposed to use a constant stopping boundary in

$$T = \inf\{n \le M : |S_n| \ge b\}, \tag{1.3}$$

which corresponds to the group sequential version of an SPRT.

While sequential analysis had an immediate impact on weapons testing when it was introduced during World War II to reduce the sample sizes of such tests (Wallis 1980), its refinements for testing new drugs and treatments received little attention from the biomedical community until the Beta-Blocker Heart Attack Trial (BHAT) that was terminated in October 1981, prior to its prescheduled end in June 1982.

The main reason for this lack of interest is that the fixed sample size (i.e., the number of patients accrued) for a typical trial is too small to allow further reduction while still maintaining reasonable power at the alternatives of interest. On the other hand, BHAT, which was a multicenter, double-blind, randomized placebo-controlled trial to test the efficacy of long-term therapy with propranolol given to survivors of an acute myocardial infarction, drew immediate attention to the benefits of sequential methods not because it reduced the number of patients but because it shortened a 4-year study by 8 months, with positive results for a long-awaited treatment for MI patients.

The "success story" of BHAT paved the way for major advances in the development of group sequential methods in clinical trials and for the steadily increasing adoption of group sequential design. Chapter 4 gives a review of these advances and describes the current methodology that has moved far beyond the Pocock and O'Brien–Fleming boundaries (1.2) and (1.3). Chapter 6 presents the design details of BHAT and the interim analysis results considered by its Data and Safety Monitoring Board. Inspired by the statistical issues raised by BHAT, a number of important and difficult problems concerning the design and analysis of clinical trials with failure-time endpoints and interim analyses have been resolved in the past 2 decades, and Chap. 6 also describes the "time-sequential" methodology developed in this connection. Chapter 5, however, shows that the fully sequential methodology summarized in Chap. 3 has recently emerged as a standard for prelicensure (Phase III) vaccine safety trials and post-marketing (Phase IV) safety studies.

Analysis of the data at the conclusion of a clinical trial typically involves tests and confidence intervals not only for the primary endpoint but also for different secondary endpoints. The use of a stopping rule whose distribution depends on these parameters introduces substantial difficulties for such inference. Siegmund (1978) developed a method, based on ordering the sample space in a certain way, to construct confidence intervals for its mean of a normal population with known variance following a RST. Alternative orderings of the sample space were subsequently introduced for group sequential tests by Rosner and Tsiatis (1988) and Emerson and Fleming (1990). By making use of resampling methods, Chuang and Lai (1998, 2000) developed a general resampling approach to constructing accurate confidence intervals following sequential tests. Subsequently, Lai and Li (2006) introduced a general ordering scheme that can be used in conjunction with resampling to completely solve the long-standing problem of constructing valid confidence intervals for the primary endpoint of a group sequential trial. Chapter 7 summarizes these developments and describes the methods. Analysis of secondary endpoints following a group sequential trial is also considered in Chap. 7, which reviews the bias-correction approach of Whitehead (1986), Liu et al. (2000), Whitehead et al. (2000), and Hall and Yakir (2003) and describes the hybrid resampling methods of Lai et al. (2009).

## 1.3 Adaptation and Sequential Optimization

After sequential analysis was introduced in response to more efficient testing of weapons during World War II, it was soon realized that sequential methods could be used to address statistical problems for which there are no solutions with fixed sample sizes. While Dantzig (1940) had shown that no fixed sample size test exists for the problem of testing the null hypothesis $H_0 : \mu = \mu_0$, with prescribed error probabilities $\alpha$ and $\beta$ at $\mu_0$ and $\mu_0 + \delta$, for the mean $\mu$ of a normal distribution whose variance $\sigma^2$ is unknown, Stein (1945) showed that a two-stage procedure that uses the first stage to estimate $\sigma^2$ and thereby to determine an appropriate second-stage sample size can have power independent of $\sigma$. Stein's two-stage design is the first example to show that one can use data during the course of an experiment to learn about the unknown parameters and thereby adapt the experimental design (which is the sample size in Stein's example) as the experiment progresses. It also paved the way for the next generation of adaptive designs in clinical trials in the 1990s that are described in Chap. 8. These adaptive designs, however, are inefficient because they do not incorporate the uncertainties of the parameter estimates at the end of the first stage. Chapter 8 also describes a new class of adaptive designs, introduced by Bartroff and Lai (2008a,b), which use an additional stage to accommodate the uncertainties in the first-stage estimates.

Adaptation via sequential learning of unknown parameters is also a central idea in the theory of nonlinear optimal experimental design. As shown in Sect. 2.3, the optimal design measure involves the unknown parameter vector $\theta$ in a nonlinear regression model. To circumvent these difficulties, Fedorov (1972) and others proposed that designs be constructed sequentially, using observations made to date to estimate $\theta$ and choosing the next design point by replacing the unknown $\theta$ in the optimal design by the current estimate. Lai et al. (2012c) have recently shown the advantages of adaptation in an integrated plan for developing a new drug, which is mentioned earlier in the second paragraph of Sect. 1.1. In the development of a new drug, an important component of the effort and costs involves clinical trials to provide clinical data to support a beneficial claim of the drug and, in case such claim is not valid, to support the termination of the development. The clinical trials progress in steps and are labeled Phase I, II, and III trials, as we have already noted in Sect. 1.1. A project team steers their operations in which intensity, cost, and duration increase with the phase; in particular, Phase III often involves over 3000 professionals, several years to reach completion, and over $100 million in cost. In addition, there is a core team that makes decisions guided by a clinical development plan (CDP). The CDP maps out the clinical development pathway, beginning with first-in-man studies and ending with submission to the regulatory agency or termination of development. It defines the number and type of clinical studies and their objectives, determines the time sequence of the studies, some of which may be carried out in parallel, identifies key risk areas, and sets key decision points and go/no-go criteria. Julious and Swank (2005) have noted that statistical methods for clinical trial design have focused primarily on "optimizing individual clinical trials" but are lacking "at a more global level in the optimization of clinical development plans." In practice, however, it is often difficult to specify in advance the cost of each clinical trial in the sequence and the prior probabilities of a go or no-go decision to perform the optimization of CDPs "at a more global level." Lai et al. (2012c) use ideas from adaptive design of clinical trials, in particular, seamless Phase II–III designs, to adapt a CDP to information acquired during the course of its execution.

Optimization is an important technique in formulating and computing statistical procedures, which can be regarded as statistical decision rules. When the decision rule consists of a sequence of actions, determination of the optimal rule involves dynamic programming. In Chap. 3 we give an introduction to dynamic programming and use it to prove the optimality of the SPRT for simple hypotheses and to derive approximately optimal tests based on generalized likelihood ratio statistics for composite hypotheses. We also give an introduction to recent advances in approximate dynamic programming and apply it to address the treatment versus experimentation dilemma in Phase I cancer trial designs.

## 1.4 Two Time Scales and Time-Sequential Survival Analysis

As pointed out in Sect. 1.2, the early termination of BHAT paved the way for major advances in the development of group sequential designs. These advances are summarized in Chap. 4, but BHAT and other trials with failure-time endpoints require more subtle methods than those described in Chap. 4. In Chap. 6 we describe these methods that address two time scales in time-sequential survival analysis. "Time-sequential" means that interim analyses are conducted over calendar times, rather than on the time scale measured by the number of subjects at each interim analysis as in group sequential methods in Chap. 4, for which the number of subjects is proportional to the variance (under the null hypothesis) of the test statistic. We begin Chap. 6 with a review of traditional (nonsequential) survival analysis, focusing on how the variances of the commonly used test statistics can be derived with relative ease, despite the complexities due to right censoring, by making use of martingale theory. In the time-sequential setting, calendar time is one time scale, and the other time scale is "information time," which is measured by the null variance of the test statistic at the time of interim analysis. There is no simple connection between the two time scales and it has been a long-standing problem concerning how to address the difficulties caused by the two time scales in the design and analysis of time-sequential clinical trials with failure-time endpoints.

In Sect. 6.5 we discuss these difficulties and describe the methods that have been developed to address them. These include a comprehensive asymptotic distribution theory for time-sequential censored rank statistics, relatively simple and yet efficient modified Haybittle–Peto tests, and interim Bayesian estimation of the maximum information at the scheduled end of the trial for futility stopping. Section 6.7

describes some recent advances, including the Phase II–III cancer trial designs that we have mentioned in Sect. 1.2 and a method that allows multiple test statistics to increase power if the trial should proceed to its scheduled end. Sections 7.3 and 7.5 describe an innovative hybrid resampling approach to statistical inference from survival data following a time-sequential trial.

## 1.5 Bayesian and Frequentist Approaches and Associated Software

Berry et al. (2011, p. 1) say that a primary purpose of their book is to describe the Bayesian approach as an alternative to the traditional frequentist approach, which is “the standard statistical approach to designing and analyzing clinical trials and other medical experiments.” They find the “flexibility in both design and analysis” and the “decision-oriented” underpinning of the Bayesian approach particularly suited to sequential analysis and adaptive design of clinical trials. On the other hand, they acknowledge that for Phase III confirmatory trials, which are “typically overseen and judged by a regulatory agency,” the statistical hurdle for regulatory approval of the new treatment is “to get a statistically significant result at a specified type I error,” and the type I error of an adaptive Bayesian design is “extremely difficult, if not impossible, to calculate” and has to be computed by Monte Carlo simulations. Their approach is to adjust the rejection threshold of the Bayesian adaptive/sequential test by using the Monte Carlo simulations carried out under some chosen parameter configuration(s) belonging to the null hypothesis. However, for a composite null hypothesis, there is no guarantee that the worst parameter configuration in the null hypothesis has been chosen for these simulations. An example is given by Lai et al. (2012a, Sect. 4.4) in their comparison of the Phase II–III design described in Chap. 6 with the Bayesian counterpart developed by Huang et al. (2009). Their numerical study shows that because the Bayesian design uses simulations under certain assumed survival rates to control the type I error, the type I error can be substantially inflated under other survival rates belonging to the highly composite null hypothesis. In contrast, the frequentist semiparametric approach used in their design and analysis is shown to maintain the prescribed type I error.

The argument of Berry et al. (2011, Chap. 1) that the Bayesian approach can handle adaptation and sequential learning much more efficiently than the frequentist approach is fair for the prevailing frequentist methods cited in their references, but it overlooks the possibility that *suitably chosen* frequentist methods can work as well, if not better. In fact, there is already a versatile arsenal of statistical methods and theories, including likelihood inference, semiparametric models for censored survival data, bootstrap, and other resampling methods, for nonsequential settings. We shall show in the subsequent chapters how these time-tested methods can be extended to sequential experiments and adaptive designs. In fact, in Chap. 3, we show that these extensions can also be derived as approximations to Bayes rules. Our viewpoint, therefore, is that Bayesian and frequentist approaches should complement each other. One may start with a Bayesian formulation and end up with a frequentist implementation that may be more convenient and appropriate for the problem at hand, for example, confirmatory testing for drug approval. This idea is illustrated in Sect. 3.7 that starts with Bayes sequential tests of one-sided hypotheses and ends up with sequential generalized likelihood ratio tests which have approximately optimal Bayesian and frequentist properties and are also convenient for implementation and description. Another example, which is beyond the scope of this book, is the classical multiarmed bandit problem; see the survey in Lai (2001, pp. 337–339) which shows that while the Bayesian formulation of the infinite-horizon version of the problem has a solution in terms of the "Gittins index" for each arm, a closed-form approximation of the Gittins index yields an upper confidence bound for the arm's mean parameter. Not only does this frequentist approximation provide an intuitive interpretation of the Bayes solution but it also leads to approximately optimal solutions of finite-horizon bandit problems with a frequentist formulation. Conversely, one may start with a frequentist problem and ends up with a Bayes solution. A classic example is the optimality theorem of Wald's SPRT in Sect. 3.6. As explained in Sect. 3.3, Wald conjectured this result on the basis of certain lower bounds for the expected sample sizes under the simple null and alternative hypotheses. Section 3.6 shows that the proof of the conjecture requires solving an auxiliary Bayes problem to which dynamic programming can be applied.

Other than Bayesian designs for Phase I trials, which usually have small sample sizes, considered in Chap. 2 and Sect. 3.8, and the interplay between Bayesian and frequentist approaches to sequential hypothesis testing discussed in Chap. 3, we focus in the subsequent chapters on the frequentist approach and refer readers to the comprehensive treatment of Bayesian methods for clinical trials in Berry et al. (2011). On the other hand, we want to discuss here an irreconcilable difference, which is widely recognized and somewhat controversial, between frequentist and Bayesian inference at the conclusion of a clinical trial with a group sequential or adaptive design. Bayesian inference (e.g., credible sets for parameters) is based on the posterior distribution given the randomly stopped sample, and no adjustment is needed for early stopping or adaptive randomization. In contrast, frequentist inference such as confidence sets has to make adjustments to ensure the correctness of the prescribed coverage probability. In nonsequential designs, the difference between credible and confidence intervals is small for large sample sizes because of the central limit theorem and higher-order expansions of the posterior distribution (Johnson 1970) and the sampling distribution (Gross and Lai 1996) of the approximate pivot used to construct credible or confidence intervals. However, for group sequential designs, this large-sample theory no longer holds, and in fact, an approximate pivot in the fixed sample size case is no longer approximately pivotal in the group sequential setting, as will be explained in Chap. 7 which also describes how valid confidence intervals can be constructed by a resampling procedure, similar to Efron's (1987) bootstrap method to construct confidence intervals based on samples of fixed size.

A particularly attractive feature of the Bayesian designs in Berry et al. (2011) is that software programs are available at the book's website for users to implement the design and analysis. This is what we try to emulate for the novel procedures described in our book. Open-source software is being developed and tested and will be posted at the website mentioned in the Preface. The introduction of flexibility in the timing of interim analysis by using the Lan–DeMets spending function and other flexible group sequential methods described in Sect. 4.1 made the application of these methods feasible in practice. A major obstacle for wide use of group sequential methods has been the lack of statistical software for performing the needed calculations. Programming code was developed by individual researchers for their own use but was not widely accessible to the clinical trials community until the mid-1990s when the commercial and academic worlds became connected via the Internet. A brief but fairly exhaustive review of the software packages available as of 2006 can be found in Wassmer and Vandemeulebroecke (2006). In 2000, University of Wisconsin at Madison made available, free of charge, an interactive FORTRAN program developed by Reboussin, DeMets, Kim, and Lan that enabled users to design group sequential trials, including those with failure-time endpoints under the assumption of proportional hazards. Its usage required certain sophistication. A similar package, focusing exclusively on failure-time endpoints but with considerably more options for the user, was developed by Gu and Lai (1999) and is downloadable from Gu's website at the Chinese University of Hong Kong, but the package has not been maintained and will be replaced by the R package currently being developed for Chap. 6. Other FORTRAN code has been made available by various researchers to provide design tools to facilitate the use of specific methodology. An example is the FORTRAN code available from Christopher Jennison of the University of Bath, United Kingdom. SAS® provides a number of design PROCs for calculating a variety of group sequential boundaries. R offers similar functionality. S+ contains a module S+ SeqTrial® that offers a wide variety of group sequential methods, point, and interval estimation options at the end of a group sequential trial under a variety of distributional assumptions.

The first package that was made available commercially on the market was PEST developed at Lancaster University, UK, by John Whitehead (formerly at the University of Reading, UK) and his collaborators. The latest version of PEST, PEST 4.4, is written in C. The software provides tools for calculating group sequential triangular boundaries for binary, normal, and time-to-failure variables. There are a number of stand-alone packages today that are available commercially. The best known are (a) East® developed by Cytel Statistical Software and Services; (b) ADDPLAN Adaptive Designs—Plans and Analyses®, currently available in Release 3.1, developed initially at the University of Köln, Germany, and commercially offered under different licensing agreements; (c) PASS 2005® distributed by NCSS, Inc.; and (d) STOPP® developed by Edward Lakatos and licensed commercially by BiostatHaven, Inc. PASS offers the traditional tools for the design of group sequential trials but does not offer analysis and trial monitoring tools as do the other packages. The most established and widely used commercial package is East. In its most current version, Version 5, East contains a broad menu of stopping boundaries for group sequential testing for normal, binary, and time-to-failure variables. Both ADDPLAN and East offer graphical user interfaces enabling the user to obtain outputs in tabular as well as graphical format. Both offer design as well as analysis and simulation tools. Recently, Cytel Corporation released two additional modules EastSurv® and EastAdapt®. EastSurv offers design and analysis capabilities outside of the proportional hazards framework through simulations. EastAdapt offers a tool for sample size recalculations using three different conditional power approaches. It also offers a simulation tool for verifying the model performance. Both EastSurv and EastAdapt are offered in addition to the core East package under separate licenses for additional fees. ADDPLAN offers a similar scope of procedures as East but its focus is more on adaptive trials rather than the more traditional group sequential designs that are the core of East. STOPP offers the same design tools as East except that its survival methodology is based on methods which do not assume proportional hazards and do not rely on simulations. As discussed above, there are at least two commercially available software packages that offer a wide variety of group sequential and adaptive designs. However, as Wassmer and Vandemeulebroecke (2006) commented in their review of existing packages, “further developments of packages or add-on modules that include, e.g., the planning, simulation and analysis of adaptive seamless designs are mandatory when the rapid development in this area is taken into account.”

# Chapter 2
# Nonlinear Regression, Experimental Design, and Phase I Clinical Trials

In typical Phase I studies in the development of relatively benign drugs, the drug is initiated at low doses and subsequently escalated to show safety at a level where some positive response occurs, and healthy volunteers are often used as study subjects. In Sect. 2.2 we describe some basic pharmacologic principles and models underlying dose determination. These models are typically nonlinear in certain parameters and therefore nonlinear regression models are used. Section 2.1 gives an introduction to nonlinear regression and also describes in this connection nonlinear mixed effects models (NONMEMs), which play a central role in population pharmacokinetics and pharmacodynamics in Sect. 2.2. In connection with Phase I studies, Sect. 2.3 gives an overview of the theory of optimal experimental design. The design and analysis of Phase I studies are described in Sect. 2.4.

This paradigm in Sect. 2.4 does not work for diseases like cancer, for which a non-negligible probability of severe toxic reaction has to be accepted to give the patient some chance of a favorable response to the treatment. Moreover, in many such situations, the benefits of a new therapy may not be known for a long time after enrollment, but toxicities manifest themselves in a relatively short time period. Therefore, patients (rather than healthy volunteers) are used as study subjects, and given the hoped-for (rather than observed) benefit for them, one aims at an acceptable level of toxic response in determining the dose. The objective of Phase I cancer trials is to find a *maximum tolerated dose* (MTD) with the ethical constraint of protecting the study subjects from toxicities in excess of what they can tolerate. To address this constraint, $3+3$ designs are often used and they are described in Sect. 2.5.1. However, simulation studies by O'Quigley et al. (1990) showed the performance of these designs to be "dismal," for which they provided the following explanation: "Not only do (these designs) not make efficient use of accumulated data, they make use of no such data at all, beyond say the previous three, or sometimes six, responses." They proposed an alternative design, called the *continual reassessment method* (CRM), which uses parametric modeling of the dose–response relationship and a Bayesian approach to estimate the MTD, or more generally the dose level $x$ such that the probability $F(x)$ of a toxic event is

J. Bartroff et al., *Sequential Experimentation in Clinical Trials,*
Springer Series in Statistics 298, DOI 10.1007/978-1-4614-6114-2_2,

$p$ (1/3 in the case of MTD). Section 2.5.2 describes the CRM and other model-based designs. However, because of the ethical demands for treating patients in the study at safe doses even though they may not be effective, 3+3 designs and their variants are still widely used despite their inadequacy in generating dose-toxicity information for the posttrial estimate of the MTD, for which the model-based designs are more efficient. Bartroff and Lai (2010) have provided a mathematical representation of this dilemma between safe treatment of current patients in the dose-finding cancer trial and efficient experimentation to gather information about the MTD for future patients. The next chapter will describe their formulation of a stochastic optimization problem that addresses this dilemma and summarize their solution of the problem, leading to a class of hybrid designs.

## 2.1 Nonlinear Regression Models

### *2.1.1 Nonlinear Least Squares*

As in linear regression models, the method of least squares is commonly used to estimate the unknown parameter vector $\boldsymbol{\theta}$ in the nonlinear regression model

$$y_j = f_{\boldsymbol{\theta}}(\boldsymbol{x}_j) + \varepsilon_j, \qquad j = 1, \ldots, n, \tag{2.1}$$

in which $f_{\boldsymbol{\theta}}(\cdot)$ is a given nonlinear function of $\boldsymbol{\theta}$ and $\varepsilon_j$ are unobservable independent random errors with zero means and

(a) $\mathrm{var}(\varepsilon_j) = \sigma^2$ (constant variance error models), or
(b) $\mathrm{var}(\varepsilon_j) = f_{\boldsymbol{\theta}}^2(\boldsymbol{x}_j)\sigma^2$ (constant coefficient of variation error models), or
(c) $\mathrm{var}(\varepsilon_j) = f_{\boldsymbol{\theta}}(\boldsymbol{x}_j)\sigma^2$ (Poisson-type error models).

We can estimate $\boldsymbol{\theta}$ by generalized least squares (GLS), that is, by minimizing

$$S(\boldsymbol{\theta}) = \sum_{j=1}^{n} w_j [y_j - f_{\boldsymbol{\theta}}(\boldsymbol{x}_j)]^2, \tag{2.2}$$

where the weights are inversely proportional to $\mathrm{var}(\varepsilon_j)$.

To compute the minimizer $\hat{\boldsymbol{\theta}}$ of (2.2), we write $f_{\boldsymbol{\theta}}(\boldsymbol{x}_j) = f(\boldsymbol{\theta}, \boldsymbol{x}_j)$, initialize with $\hat{\boldsymbol{\theta}}^{(0)}$ and approximate $f(\boldsymbol{\theta}, \boldsymbol{x}_j)$ after the $k$th iteration, which yields $\hat{\boldsymbol{\theta}}^{(k)}$, by

$$f(\boldsymbol{\theta}, \boldsymbol{x}_j) \approx f\left(\hat{\boldsymbol{\theta}}^{(k)}, \boldsymbol{x}_j\right) + \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{(k)}\right)^T \nabla f\left(\hat{\boldsymbol{\theta}}^{(k)}, \boldsymbol{x}_j\right)$$

so that (2.1) can be approximated by the linear regression model

$$y_j - f\left(\hat{\boldsymbol{\theta}}^{(k)}, x_j\right) = \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{(k)}\right)^T \nabla f\left(\hat{\boldsymbol{\theta}}^{(k)}, x_j\right) + \varepsilon_j. \tag{2.3}$$

The GLS estimate $\hat{\boldsymbol{\theta}}^{(k+1)}$ of $\boldsymbol{\theta}$ in (2.3) is given explicitly, and the iterative scheme is called the *Gauss–Newton algorithm*.

The Gauss increment $\delta_{k+1} := \hat{\boldsymbol{\theta}}^{(k+1)} - \hat{\boldsymbol{\theta}}^{(k)}$ may produce an increase in $S(\boldsymbol{\theta})$ when it is outside the region where the linear approximation holds. To ensure a decrease in $S(\boldsymbol{\theta})$, use a step factor $0 < \lambda \leq 1$ so that $S(\hat{\boldsymbol{\theta}}^{(k)} + \lambda\delta^{(k)}) < S(\boldsymbol{\theta}^{(k)})$. A commonly used method is to start with $\lambda = 1$ and halve it until we have $S(\boldsymbol{\theta}^{(k+1)}) < S(\boldsymbol{\theta}^{(k)})$. A commonly used criterion for numerical convergence is the size of the parameter increment relative to the parameter value. Another criterion is that the relative change in $S(\boldsymbol{\theta})$ be small. A third criterion is that $Y - \eta(\boldsymbol{\theta}^{(k)})$ be nearly orthogonal to the tangent space of $\eta(\boldsymbol{\theta}) := (f(\boldsymbol{\theta}, \boldsymbol{x}_1), \ldots, f(\boldsymbol{\theta}, \boldsymbol{x}_n))^T$ at $\boldsymbol{\theta}^{(k)}$. The Gauss–Newton algorithm is aborted at the $k$th step when one gets a singular (or nearly singular) coefficient matrix in the linear equation defining GLS. It may also stop after reaching a prescribed upper bound on the number of iterations without convergence. When one does not get an answer from the Gauss–Newton algorithm, one should choose another starting value and repeat the algorithm.

### *2.1.2 Nonlinear Mixed Effects Models*

As will be explained in the next section, two important pharmacologic models are the poly-exponential model $f_{\boldsymbol{\theta}}(t) = \sum_{k=1}^{K} \alpha_k e^{-\lambda_k t}$, with $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_k, \lambda_1, \ldots, \lambda_k)^T$ and $t$ denoting time, and the Michaelis–Menten model $f_{\boldsymbol{\theta}}(u) = \nu u/(\alpha + u)$, with $\boldsymbol{\theta} = (\nu, \alpha)^T$ and $u$ denoting drug concentration. A Phase I trial collects data from $I$ subjects, yielding $(y_{ij}, x_{ij})$, $i = 1, \ldots, I$, $j = 1, \ldots, n_i$. In the analysis of these data, it is more flexible to allow subject-specific parameters $\boldsymbol{\theta}_i$ in (2.1). This leads to a NONMEM of the form

$$y_{ij} = f_i(t_{ij}, \boldsymbol{\theta}_i) + \varepsilon_{ij}, \quad \boldsymbol{\theta}_i = \boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\beta}) + \boldsymbol{b}_i \qquad (1 \leq j \leq n_i,\ 1 \leq i \leq I), \tag{2.4}$$

in which $\boldsymbol{\theta}_i$ is a $1 \times r$ vector of the $i$th subject's parameters whose regression function on the subject's observed covariate $\boldsymbol{x}_i$ is given by $\boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\beta})$ with $1 \times s$ parameter vector $\boldsymbol{\beta}$, which is the "fixed effect" to be estimated. The "random effects" $\boldsymbol{b}_i$ in (2.4) are assumed to be independent and identically distributed, having common distribution $G$ with mean 0. The $i$th subject's response $y_{ij}$ at $t_{ij}$ has mean $f_i(t_{ij}, \boldsymbol{\theta}_i)$, in which $f_i$ is a known function and $t_{ij}$ may represent time or some covariate value (such as drug concentration) at that time. Given $\boldsymbol{\theta}_i$, the random errors $\varepsilon_{ij}$ are assumed to be normal with mean 0 and standard deviation $\sigma w(\boldsymbol{\theta}_i)$, in which $w$ is a given function and $\sigma$ is an unknown parameter. The regression function $\boldsymbol{g}$ relates $\boldsymbol{\theta}_i$ to the $i$th subject's physiologic characteristics that constitute the covariate vector $\boldsymbol{x}_i$ in (2.4). The first equation of (2.4) is often called the *individual measurement model* and the second equation the *population structure model*. The population distribution $G$ is usually assumed to be normal with mean 0 and covariance matrix $\boldsymbol{\Sigma}$ so that $\boldsymbol{\beta}$, $\sigma$, $\boldsymbol{\Sigma}$ can be estimated by maximum likelihood. However, unlike linear mixed effects (LME) models in which the normal assumption on $G$ yields closed-form expressions of the likelihood, the normality of $G$ in NONMEM leads to computationally intensive likelihoods that involve $I$ integrals. A commonly used approach, as adopted in the software package NONMEM (Beal and Sheiner 1992) or the `nlme` procedure in R, is to develop iterative schemes based on first-order approximations of $f_i(t_{ij}, \boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\beta}) + \boldsymbol{b}_i)$ in (2.4), so that the normal assumption on $G$ can be used to reduce the problem to that of a linear Gaussian mixed effects model at each iterative step.

Unless otherwise stated, we shall assume throughout the sequel that the random errors $\varepsilon_{ij}$ in model (2.4) have common variance $\sigma^2$ (so $w(\boldsymbol{\theta}) \equiv 1$). The likelihood function $L(\boldsymbol{\beta}, \sigma, \boldsymbol{\Sigma})$ is proportional to

$$|\boldsymbol{\Sigma}|^{-I/2} \prod_{i=1}^{I} \int_{\mathbb{R}^r} \sigma^{-n_i} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{j=1}^{n_i} [y_{ij} - f_i(t_{ij}, \boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\beta}) + \boldsymbol{b}_i)]^2 - \frac{1}{2} \boldsymbol{b}_i \boldsymbol{\Sigma}^{-1} \boldsymbol{b}_i^T \right\} d\boldsymbol{b}_i, \tag{2.5}$$

where $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$. For the case of more general $w(\boldsymbol{\theta}_i)$, simply replace $\sigma$ in (2.5) by $\sigma w(\boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\beta}) + \boldsymbol{b}_i)$. Computing the maximum likelihood estimate of $(\boldsymbol{\beta}, \sigma, \boldsymbol{\Sigma})$ via numerical integration and nonlinear optimization becomes prohibitively difficult for large $I$. Letting $\boldsymbol{\eta} = (\sigma, \boldsymbol{\Sigma})$, Lindstrom and Bates (1990) proposed the following iterative procedure that involves successive linear approximations to $f_i(t_{ij}, \boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\beta}) + \boldsymbol{b}_i)$. At the $m$th iteration, the Lindstrom–Bates procedure consists of a pseudo-data step and a LME step:

(a) *The pseudo-data step*
Given the current estimate $\hat{\boldsymbol{\eta}}^{(m)}$ of $\boldsymbol{\eta}$, compute $\hat{\boldsymbol{\beta}}^{(m)} = \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\eta}}^{(m)})$ and $\hat{\boldsymbol{b}}_i^{(m)} = \hat{\boldsymbol{b}}_i(\hat{\boldsymbol{\eta}}^{(m)})$, $1 \le i \le I$, that jointly minimize

$$\sum_{i=1}^{I} \left\{ (\hat{\sigma}^{(m)})^{-2} S_i(\boldsymbol{\beta}, \boldsymbol{b}) + \boldsymbol{b}_i \left( \hat{\boldsymbol{\Sigma}}^{(m)} \right)^{-1} \boldsymbol{b}_i^T \right\}, \tag{2.6}$$

where

$$S_i(\boldsymbol{\beta}, \boldsymbol{b}) = \sum_{j=1}^{n_i} [y_{ij} - f_i(t_{ij}, \boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\beta}) + \boldsymbol{b})]^2 .$$

This can be carried out by modifying a standard nonlinear least squares routine; see Sect. 6.1 of Lindstrom and Bates (1990). Define the $s \times n_i$, $r \times n_i$, and $1 \times n_i$ matrices

$$\boldsymbol{X}_i^{(m)} = \left( \frac{\partial f_i}{\partial \boldsymbol{\beta}} \left( t_{ij}, \boldsymbol{g}\left(\boldsymbol{x}_i, \boldsymbol{\beta}\right) + \hat{\boldsymbol{b}}_i^{(m)} \right) \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{(m)}} \right)_{1 \le j \le n_i},$$

$$\boldsymbol{Z}_i^{(m)} = \left( \frac{\partial f_i}{\partial \boldsymbol{b}_i} \left( t_{ij}, \boldsymbol{g}\left(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}}^{(m)}\right) + \boldsymbol{b}_i \right) \Big|_{\boldsymbol{b}_i=\hat{\boldsymbol{b}}_i^{(m)}} \right)_{1 \le j \le n_i},$$

$$\boldsymbol{Y}_i^{(m)} = \left( y_{ij} - f_i \left( t_{ij}, \boldsymbol{g}\left(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}}^{(m)}\right) + \hat{\boldsymbol{b}}_i^{(m)} \right) \right)_{1 \le j \le n_i} + \hat{\boldsymbol{\beta}}^{(m)} \boldsymbol{X}_i^{(m)} + \hat{\boldsymbol{b}}_i^{(m)} \boldsymbol{Z}_i^{(m)}.$$

(b) *The LME step*
Linear approximation to $f_i(t_{ij}, \boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\beta}) + \boldsymbol{b}_i)$ around $(\hat{\boldsymbol{\beta}}^{(m)}, \hat{\boldsymbol{b}}_i^{(m)})$ leads to the LME model

$$\boldsymbol{Y}_i^{(m)} = \boldsymbol{\beta} \boldsymbol{X}_i^{(m)} + \boldsymbol{b}_i \boldsymbol{Z}_i^{(m)} + (\varepsilon_{i1}, \ldots, \varepsilon_{in_i}). \tag{2.7}$$

The integrals in (2.5) for the likelihood function of the LME model (2.7) (instead of (2.4)) have closed-form expressions, yielding maximum likelihood estimates of the form

$$\hat{\boldsymbol{\beta}} = \left( \sum_{i=1}^{I} \boldsymbol{Y}_i^{(m)} \boldsymbol{V}_{i,m}^{-1} \boldsymbol{X}_i^{(m)T} \right) \left( \sum_{i=1}^{I} \boldsymbol{X}_i^{(m)} \boldsymbol{V}_{i,m}^{-1} \boldsymbol{X}_i^{(m)T} \right)^{-1}, \tag{2.8}$$

where $\boldsymbol{V}_{i,m} = \boldsymbol{Z}_i^{(m)T} \hat{\boldsymbol{\Sigma}} \boldsymbol{Z}_i^{(m)} + \hat{\sigma}^2 \boldsymbol{I}_{n_i}$ and $\hat{\boldsymbol{\eta}} = (\hat{\sigma}, \hat{\boldsymbol{\Sigma}})$ is computed via the Newton–Raphson algorithm to maximize the likelihood.

Wolfinger (1993) derives the above pseudo-data step by using Laplace's approximation arguments. Vonesh (1996) directly approximates the integrals in (2.5) with $\sigma, \boldsymbol{\beta}, \boldsymbol{\Sigma}$ fixed, by using Laplace's asymptotic formula

$$\int_{\mathbb{R}^r} e^{\ell_i(\boldsymbol{b})} \, d\boldsymbol{b} \sim (2\pi)^{r/2} \left\{ \det\left( -\ddot{\ell}_i(\hat{\boldsymbol{b}}_i) \right) \right\}^{-1/2} e^{\ell_i(\hat{\boldsymbol{b}}_i)}, \tag{2.9}$$

where $\hat{\boldsymbol{b}}_i$ is the maximizer of $\ell_i(\boldsymbol{b})$ and $\ddot{\ell}_i$ is the Hessian matrix of second partial derivatives of $\ell_i$ with respect to the components of $\boldsymbol{b}$. Noting that Laplace's approximation to an integral corresponds to adaptive Gaussian quadrature with one quadrature point, Pinheiro and Bates (1995) use adaptive Gaussian quadrature with $q$ quadrature points to compute the integrals in (2.5). Lai and Shih (2003b) have developed a hybrid method that uses (2.9) if the minimum eigenvalue $\lambda_{\min}(-\ddot{\ell}_i(\hat{\boldsymbol{b}}_i))$ exceeds a prescribed threshold and uses Monte Carlo simulations otherwise. Lai et al. (2006b) introduce importance sampling to refine the Monte Carlo component of the hybrid method. They also point out the importance of approximating the likelihood function adequately with relative ease for selecting good predictive models $\boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\beta})$.

Since the normality assumption on $G$ only provides numerically tractable maximum likelihood estimates after various approximations, a natural alternative is to try estimating $G$ nonparametrically by a distribution with finite support, with the number of support points depending on the sample size. However, even for the simple case $n_i \equiv n$ and $f_i(t_{ij}, \boldsymbol{\theta}_i) = \boldsymbol{\theta}_i$ with known $\boldsymbol{\beta}$ and $\sigma$, it is difficult to estimate $G$ well since the optimal rate of convergence of the estimate to $G$ is very slow when $G$ has a smooth density function, as pointed out by Fan (1991). Lai and Shih (2003a) have developed a nonparametric maximum likelihood estimator (MLE) of $G$ when there are $I' \leq I$ subjects whose $\boldsymbol{\theta}_i$ can be well estimated by the nonlinear least squares estimator $\tilde{\boldsymbol{\theta}}_i$ based on $\{(y_{ij}, t_{ij}) : 1 \leq j \leq n_i\}$. Because of the low resolution in estimating $G$ nonparametrically, however, the nonparametric approach does not yield a better estimate of $f(\cdot, \cdot)$ in the simulation study reported by Lai and Shih (2003a) who consider the case of $f_i$ being all equal (to $f$).

## 2.2 Pharmacokinetics and Pharmacodynamics

The nonlinear regression and NONMEM in the preceding section are basic statistical methods in pharmacology, which is the science dealing with interactions between living systems and molecules, especially chemicals introduced from outside the system. This broad definition includes clinical pharmacology (whose objective is to prevent, diagnose, and treat diseases with drugs) and the pathogenesis of diseases due to chemicals in the environment; see Katzung (1995). A drug is defined as a small molecule that, when introduced into the body, alters the body's function. The component of a cell or organism that interacts with a drug and initiates the chain of biochemical events leading to the drug's therapeutic and toxic effects is called a *receptor*. The receptor concept has become the central focus of investigation of *pharmacodynamics* (PD), which is the study of drug effects and their mechanisms of action. The relation between the dose of a drug and its clinically observed effects can be quite complex. In carefully controlled in vitro systems, however, the relation between the concentration of a drug at the site(s) of action and its effects can often be described by relatively simple mathematical models. How a drug dose produces its effects involves not only pharmacodynamics but also *pharmacokinetics* (PK). The latter is concerned with the concentration–time curve that is associated with the following "history" of a single administration of a drug:

(a) *Absorption phase of the drug into the body:* Transfer of the drug from its site of administration (via oral, or inhalational, or intravenous, or other route) into the bloodstream.
(b) *Distribution phase:* Distribution of the drug to different compartments of the body, including receptor-binding sites in the target tissue, and resulting in rapid decline in plasma concentration.
(c) *Elimination phase:* Excretion of chemically unchanged drug or elimination via metabolism that converts the drug into one or more metabolites (e.g., at the liver).

Drug administration can be divided into two phases, a PK phase in which the kinetics of drug absorption, distribution, and elimination translate into drug concentration–time relationships in the body, and a PD phase in which the drug concentration at the site(s) of action leads to the response/effects produced. Knowledge of both phases is important for the design of a dosage regimen to achieve the therapeutic objective. Since both the desired response and toxicity of the drug are functions of the drug concentration at the site(s) of action, the therapeutic objective can be achieved only when the drug concentration lies within a "therapeutic window," outside which the therapy is either ineffective or has unacceptable toxicity. Drug concentrations, however, can rarely be measured directly at the sites of action and are typically measured at the plasma, which is a more accessible site. An optimal dosage regimen can therefore be defined as one that maintains the plasma concentration of a drug within the therapeutic window. This can be achieved for many drugs by giving an initial dose to yield a plasma concentration within the therapeutic window and then maintaining the concentration within this window by periodic doses to replace the drug lost over time.

A basic goal of PD models is to describe and quantify the steady-state relationship of drug concentration $(C)$ at an effector site to the drug effect $(E)$. The simplest PD model for one drug is the so-called Emax model defined by $E = e_{\max}C/(C+c_{50})$, where $e_{\max}$ is the maximum effect that the drug can produce and $c_{50}$ is the concentration that yields 50 % of $e_{\max}$. This equation is the same as the Michaelis–Menten model in enzyme kinetics. A generalization to incorporate the baseline effect $e_0$ leads to

$$E = e_0 + e_{\max}C/(C+c_{50}). \tag{2.10}$$

A convenient surrogate for the drug concentration at an effector site, which is difficult to measure directly, is dose $(D)$. In empirical studies, $C$ and $c_{50}$ in (2.10) are replaced by $D$ and $ED_{50}$.

There is a large literature on PK models, which can roughly be classified as "mechanistic" and "empirical"; see Rowland and Tozer (1989). In mechanistic models, the body is viewed in terms of kinetic compartments between which the drug distributes and from which elimination occurs. The kinetics is often described by a linear system of ordinary differential equations, which have explicit solutions involving exponential functions. On the other hand, the rate constants of a compartmental model may be functions of the concentration of the drug itself or another metabolite/interacting drug, leading to a system of nonlinear differential equations that have to be solved numerically. Empirical PK models are typically poly-exponential models of the form $\sum \alpha_i e^{-\lambda_i t}$. One such model that is commonly used is the one-compartment model

$$y_j = \frac{Dk_a}{V(k_a-k_e)}(e^{-k_e t_j} - e^{-k_a t_j}) + \varepsilon_j, \quad 1 \le j \le n, \tag{2.11}$$

in which $y_j$ is the concentration at time $t_j$ after the administration of a single oral dose $D$. Here $V$, $k_a$, $k_e$ are the volume of distribution, absorption rate constant, and elimination rate constant, respectively. Note that (2.11) has the form of a bi-exponential model $\alpha_1 e^{-\lambda_1 t} + \alpha_2 e^{-\lambda_2 t}$ with $\alpha_1 = -\alpha_2$.

So far we have considered estimation of the PK/PD parameters of a subject from the data in a study on the subject. In many PK/PD studies, however, data are collected from a number of subjects, some of whom may have intensive blood sampling while others only have sparse data. A primary objective of these studies is to study the PK/PD characteristics of the entire population, such as how they vary with certain covariates. This requires embedding the individual parametric PK/PD models in a population model. For example, the $y_j$ in (2.11) are now replaced by $y_{ij}$, where $i$ denotes the subject number. Since the dose, volume of distribution, absorption, and elimination rate constants may vary from subject to subject, we also have to replace $D, V, k_a, k_e, n$ by $D_i, V_i, k_{ai}, k_{ei}$, and $n_i$ in (2.11). Let $\boldsymbol{\theta}_i$ be the vector consisting of the logarithms of the PK parameters $V_i, k_{ai}, k_{ei}$. The unknown $\boldsymbol{\theta}_i$ may vary with certain covariates, such as the subject's age and body weight. How can the individual subjects' data be used to analyze such relationships for the target population, of which the subjects can be regarded as a sample? The NONMEM provides a valuable tool to address this problem. The subject's data are often too sparse to provide an adequate estimate $\hat{\boldsymbol{\theta}}_i$ of $\boldsymbol{\theta}_i$ so that $h(\hat{\boldsymbol{\theta}}_i)$ can be used to estimate $h(\boldsymbol{\theta}_i)$. If $\boldsymbol{\beta}$, $\sigma$, and $G$ are known, then a natural estimate of $h(\boldsymbol{\theta}_i)$ in the mixed effects model is the posterior mean $E_{\boldsymbol{\beta},\sigma^2,G}[h(\boldsymbol{\theta}_i) \,|\, \text{subject } i\text{'s data}]$. Without assuming $\boldsymbol{\beta}$, $\sigma^2$, and $G$ to be known, the empirical Bayes approach replaces them by their estimates $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}^2$, and $\hat{G}$ from the $I$ studies so that $h(\boldsymbol{\theta}_i)$ is estimated by $\widehat{h(\boldsymbol{\theta}_i)} = E_{\hat{\boldsymbol{\beta}},\hat{\sigma}^2,\hat{G}}[h(\boldsymbol{\theta}_i) \,|\, \text{subject } i\text{'s data}]$.

Returning to the PD model (2.10), the variable $C$ refers to concentration at an effector site. It is usually impossible to measure $C$ directly, so some surrogate for $C$ has to be used. On the other hand, if one has a kinetic model for $C$, then it can be used to impute the value of $C$ from the blood/urine measurements. Chapter 9 of Davidian and Giltinan (1995) illustrates how population PK/PD models can be synthesized for such tasks.

## 2.3 Theory of Optimal Design

The conditions under which an experiment is performed affect the quality of information arising from the experiment. *Optimal design of experiments* (or simply *optimal design*) concerns how to choose these conditions, or "settings," in order to maximize the amount of information coming from an experiment and thus optimize the quality of statistical inference that is possible. In the context of clinical trials, these settings may be the treatment dose or dosing regimen, the treatment type, or the characteristics of the patient who may be randomized into one of multiple treatment groups. In what follows we give a brief introduction to the theory emphasizing general concepts over technicalities; for a more complete mathematical treatment, see Fedorov (1972) or Silvey (1980).

### 2.3.1 Optimal Design Theory in Linear Regression Models

Consider a random variable $Y \sim p(y|\boldsymbol{x}, \boldsymbol{\theta}, \sigma)$ such that $\mathrm{Var}(Y|\boldsymbol{x}, \boldsymbol{\theta}, \sigma) = \sigma^2$ and

$$E(Y|\boldsymbol{x}, \boldsymbol{\theta}, \sigma) = \boldsymbol{\theta}^T \boldsymbol{x}, \tag{2.12}$$

where $\boldsymbol{x} = (x_1, \ldots, x_k)^T \in \mathscr{X}$, the *design space*, is a vector of control variables which may be chosen by the experimenter and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T$ and $\sigma$ are unknown parameters. The linear regression model $Y \sim N(\boldsymbol{\theta}^T \boldsymbol{x}, \sigma^2)$ will be referred to as the normal case for which the linearity of (2.12) in $\boldsymbol{\theta}$ greatly simplifies the problem of choosing $\boldsymbol{x}$ in order to get the maximal information about $\boldsymbol{\theta}$ out of $Y$. Before proceeding to the problem, we make two remarks about the assumptions. First, (2.12) can be extended to $E(Y|\boldsymbol{x}, \boldsymbol{\theta}, \sigma) = \theta_1 f_1(\boldsymbol{x}) + \cdots + \theta_k f_k(\boldsymbol{x})$, where $\boldsymbol{f} = (f_1, \ldots, f_k)$ and the $f_i$ are known functions. Replacing $\boldsymbol{x}$ by $\boldsymbol{f}$ and the design space $\mathscr{X}$ by $\boldsymbol{f}(\mathscr{X})$ reduces to the original problem. Second, the variance $\sigma^2$ could be replaced by $\sigma^2 v(\boldsymbol{x})$ for any known function $v$ because this case can again be reduced to the original one with $\tilde{Y} = Y/\sqrt{v(\boldsymbol{x})}$ replacing $Y$.

Suppose we are planning to perform $n$ independent experiments with input variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ which will result in the independent observations $Y_1, \ldots, Y_n$. The least squares estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, or equivalently the MLE in the normal case, has covariance matrix

$$\sigma^2 \left( \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} \tag{2.13}$$

when the $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are such that $\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^T$ is invertible. Two key properties of (2.13) are that it does not depend on $\boldsymbol{\theta}$, which is a direct result of the linear structure of (2.12), and that it depends on $\sigma$ but in a special way such that the minimizer of any function of (2.13) does not depend on $\sigma$. If the desire is to make (2.13) "small" in some sense, then this is equivalent to making the *information matrix*

$$\boldsymbol{M} = \boldsymbol{M}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^T \tag{2.14}$$

"large." Since $\boldsymbol{M}$ is a matrix, there are various criteria for judging $\boldsymbol{M}$ to be "large" so that the optimal design problem is to find the $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ that maximize $\Psi(\boldsymbol{M})$, for some real-valued function $\Psi$. Some popular choices for $\Psi$ include the following:

*D-optimality:* Under the normality assumption, the volume of the confidence ellipsoid for $\boldsymbol{\theta}$ is proportional to $(\det \boldsymbol{M})^{-1/2}$, and minimizing this is equivalent to maximizing $\Psi(\boldsymbol{M}) = \log\det(\boldsymbol{M})$.

***c**-optimality:* For a given $k$-vector $\boldsymbol{c}$, the least squares estimate (or MLE in the normal case) of the linear combination $\boldsymbol{c}^T \boldsymbol{\theta}$ is $\boldsymbol{c}^T \hat{\boldsymbol{\theta}}$, which has variance proportional to

$$\boldsymbol{c}^T \boldsymbol{M}^{-1} \boldsymbol{c}, \tag{2.15}$$

hence, $\Psi(\boldsymbol{M}) = -\boldsymbol{c}^T \boldsymbol{M}^{-1} \boldsymbol{c}$ is the function to be maximized for this criterion.

*E-optimality:* Closely related to $\boldsymbol{c}$-optimality is the criterion which seeks to minimize the maximum of (2.15) over all $\boldsymbol{c}$ in the $k$-dimensional unit sphere, that is, to minimize

$$\max_{\boldsymbol{c}:||\boldsymbol{c}||=1} \boldsymbol{c}^T \boldsymbol{M}^{-1} \boldsymbol{c}.$$

Kiefer (1974) showed that this is equivalent to maximizing the minimum eigenvalue of $\boldsymbol{M}$, which $\Psi$ is taken to be for this criterion.

Because of the discreteness of the problem of maximizing $\Psi(\boldsymbol{M})$ over all choices for $\boldsymbol{x}_1,\dots,\boldsymbol{x}_n$, standard numerical optimization techniques often have difficulty, especially when $n$ is large. Moreover, the value of $n$ itself may not be well motivated in the experimenter's mind prior to the experiment. An elegant solution comes with the identification of $\boldsymbol{x}_1,\dots,\boldsymbol{x}_n$ to a certain probability measure over the design space $\mathscr{X}$, that is, the discrete measure placing mass $1/n$ on each point $\boldsymbol{x}_1,\dots,\boldsymbol{x}_n$, and enlarging the search to include all such probability measures has led to the *approximate theory* of linear optimal design (Kiefer 1974). Letting $\mu$ denote a probability measure on $\mathscr{X}$ and

$$\boldsymbol{M}(\mu) = E_\mu\left(\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T\right), \tag{2.16}$$

where $\tilde{\boldsymbol{x}}$ denotes the random variable with distribution $\mu$, the optimization problem is equivalent to finding the measure $\mu$ that maximizes $\Psi(\boldsymbol{M}(\mu))$. Closed-form analytic solutions are available in some cases, but in general, iterative algorithms are necessary to find optimal designs; see Fedorov (1972, Sect. 2.10).

### 2.3.2 Elfving's Method for c-Optimal Design

In order to give concrete examples we next focus on $\boldsymbol{c}$-optimal designs because, in low dimensions, optimal designs can often be found exactly by using an elegant geometric method of Elfving (1952). Assume that a linear model (2.12) is given and that the design space $\mathscr{X} \subseteq \mathbb{R}^k$ is compact, that is, closed and bounded. For a given vector $\boldsymbol{c} \in \mathbb{R}^k$, the problem is to find the measure $\mu$ on $\mathscr{X}$ maximizing

$$\Psi(\boldsymbol{M}(\mu)) = -\boldsymbol{c}^T \boldsymbol{M}(\mu)^{-1} \boldsymbol{c}$$

(or equivalently, minimizing $\boldsymbol{c}^T \boldsymbol{M}(\mu)^{-1} \boldsymbol{c}$), where $\boldsymbol{M}(\mu)$ is given by (2.16). It follows from the facts that $\Psi$ is a concave function (of matrices), the space of all matrices $\boldsymbol{M}(\mu)$ is convex, and Carathéodory's theorem (see Silvey 1980, p. 72) that a maximizer of $\Psi(\boldsymbol{M}(\mu))$ can be found among the measures $\mu$ with at most $k$ support points, that is, $\mu$ of the form

$$\mu = \sum_{i=1}^k p_i \delta_{\boldsymbol{x}_i}, \quad \text{where } \sum_{i=1} p_i = 1,\ \boldsymbol{x}_i \in \mathscr{X} \text{ and } p_i \geq 0 \text{ for all } i = 1,\dots,k, \tag{2.17}$$

in which $\delta_{\boldsymbol{x}}$ is the degenerate measure putting mass 1 at $\boldsymbol{x}$. Therefore, we can restrict our search for a maximizer of $\Psi(\boldsymbol{M}(\mu))$ to probability measures of the form (2.17).

Elfving's (1952) method for finding this discrete measure is the following. Let $\mathscr{X}^- = \{-\boldsymbol{x} : \boldsymbol{x} \in \mathscr{X}\}$ denote the reflection of $\mathscr{X}$ through the origin, and let $\mathscr{S}$ denote the *convex hull* of $X \cup X^-$, that is, $\mathscr{S}$ is the collection of all points of the form $\sum_{i=1}^k p_i \boldsymbol{z}_i$, where $\sum_{i=1}^k p_i = 1$, $p_i \geq 0$ and $\boldsymbol{z}_i \in X \cup X^-$ for all $i = 1, \ldots, k$. Extend a ray from the origin through the point $\boldsymbol{c}$ and let $\boldsymbol{s}^* \in \mathscr{S}$ be the point where this ray pierces the boundary of $\mathscr{S}$. By the definition of $\mathscr{S}$, $\boldsymbol{s}^*$ can be written as

$$\boldsymbol{s}^* = \sum_{i=1}^k \pm p_i \boldsymbol{x}_i$$

for some choice of signs, where the $p_i$ and $\boldsymbol{x}_i$ satisfy the conditions in (2.17). Then the design measure $\sum_{i=1}^k p_i \delta_{\boldsymbol{x}_i}$ is $\boldsymbol{c}$-optimal, that is, the design that places weight $p_i$ at point $\boldsymbol{x}_i$, $i = 1, \ldots, k$; see Chernoff (1972) for a sketch of the proof.

*Example 2.1.* Suppose that independent responses $Y_i$ to a drug with dose $x_i \in [0, a]$ ($a > 0$ the known "maximum dose") are given by

$$Y_i = \alpha x_i + \beta x_i^2 + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where the $\varepsilon_i$ are i.i.d. $N(0, \sigma^2)$ random variables. This model fits into the form (2.12) by taking $\boldsymbol{x} = (x, x^2)^T$, $\boldsymbol{\theta} = (\alpha, \beta)^T$, and $k = 2$. Not worrying for the moment about what value of $n$ to use, suppose that the ultimate objective of the $n$ measurements to be taken is to estimate optimally the mean response at some critical dose $x_0$, $0 < x_0 \leq a$. Thus, it is appropriate to consider the $\boldsymbol{c}$-optimal design with $\boldsymbol{c} = (x_0, x_0^2)^T$ for optimal estimation of the mean response $\alpha x_0 + \beta x_0^2 = \boldsymbol{c}^T \boldsymbol{\theta}$ at dose $x_0$. The design space

$$\mathscr{X} = \{(x, x^2) \in \mathbb{R}^2 : \ 0 \leq x \leq a\},$$

as well as $\mathscr{X}^-$, are truncated parabolas. Let $\gamma = \sqrt{2} - 1 = .4142136\ldots$.

*Case 1.* If $0 < x_0 \leq \gamma a$, then the ray in direction $(x_0, x_0^2)$ pierces $\mathscr{S}$ at the point

$$\left( \frac{a^2 \gamma (1-\gamma)}{a(\gamma^2+1) - x_0(\gamma+1)}, x_0 \cdot \frac{a^2 \gamma (1-\gamma)}{a(\gamma^2+1) - x_0(\gamma+1)} \right). \tag{2.18}$$

Setting (2.18) equal to

$$p(\gamma a, \gamma^2 a^2) + (1-p)(-a, -a^2)$$

and solving for $p$ gives

$$p = \frac{1 - x_0(\gamma+1) + a\gamma}{(\gamma+1)[a(\gamma^2+1) - x_0(\gamma+1)]}. \tag{2.19}$$

Thus, the $\boldsymbol{c}$-optimal design is $\mu = p\delta_{(\gamma a, \gamma^2 a^2)} + (1-p)\delta_{-(a,a^2)}$ which, in other words, puts the fraction $p$ of observations at dose $x = \gamma a$ and the remaining fraction $1-p$ at dose $x = a$. Note that the design may not be implementable in practice, since $np$, with $p$ given by (2.19), may not be an integer. This is a consequence of using the optimal design formulation that uses a probability measure rather than a discrete collection of $n$ design points to represent a design. In practice, if $np$ is not an integer, then choose the closest integer.

*Case 2.* If $\gamma a \le x_0 \le a$, then the ray in direction $(x_0, x_0^2)$ pierces $\mathscr{S}$ precisely at $(x_0, x_0^2)$; hence, the $\boldsymbol{c}$-optimal design is simply $\delta_{(x_0, x_0^2)}$, that is, the design that puts all measurements at dose $x = x_0$.

### *2.3.3 Extension to Nonlinear Models*

A key feature of the linear design theory in the previous section is that the information matrix (2.14) does not depend on $\boldsymbol{\theta}$. In this section we consider the more general case where

$$Y \sim p(y|\boldsymbol{x}, \boldsymbol{\theta}) \quad \text{and} \quad E(Y|\boldsymbol{x}, \boldsymbol{\theta}) = \eta(\boldsymbol{\theta}, \boldsymbol{x}) \tag{2.20}$$

for some function $\eta$, where we have absorbed the parameter $\sigma$ of the previous section into $\boldsymbol{\theta}$ for notational simplicity since the distinction between parameters of interest and nuisance parameters does not matter in the nonlinear case. To generalize the notion of information matrix used above, we note that (2.13) is the inverse of the Fisher information matrix of independent observations $Y_1, \dots, Y_n$, and therefore it is natural to define $\boldsymbol{M}(\mu) = \boldsymbol{M}(\mu, \boldsymbol{\theta})$ in the nonlinear case as the Fisher information of the design $\mu$

$$\boldsymbol{M}(\mu, \boldsymbol{\theta}) = \int_{\mathscr{X}} E\left[ -\frac{\partial^2 p(Y|\boldsymbol{x}, \boldsymbol{\theta})}{\partial\theta_i \partial\theta_j} \right] d\mu(\boldsymbol{x}), \tag{2.21}$$

where the expectation in (2.21) is taken over $Y$. As the notation suggests, the information matrix $\boldsymbol{M}(\mu, \boldsymbol{\theta})$ depends on $\boldsymbol{\theta}$ in general. The problem of optimal design now becomes more difficult as the optimal design for inference about $\boldsymbol{\theta}$ now depends on $\boldsymbol{\theta}$ itself. Application of linear optimal design theory leads to *locally optimal designs*, that is, designs that are optimal for a given value of $\boldsymbol{\theta}$. A globally optimal design for nonlinear models has to proceed in a sequential fashion, computing a locally optimal design at the current estimate of $\boldsymbol{\theta}$ to obtain a new measurement or measurements, and then updating the estimate and repeating the process until the criterion function or sequence of estimates is judged to converge; see Fedorov (1972, Sect. 4.4). Here, the role of prior information about $\boldsymbol{\theta}$ is important, particularly for beginning the sequential process. If there is prior information about the true value of $\boldsymbol{\theta}$, then the sequential process can begin at that value. Such information may be present if the current experiment is a continuation of a previous experiment or if theoretical knowledge about the current or similar settings is available, and in either of these situations, the prior information may be encoded in a prior distribution on $\boldsymbol{\theta}$ in the Bayesian sense, from which an estimate of the true value of $\boldsymbol{\theta}$ can be obtained. In the absence of such prior information, "preliminary" observations should be performed using some nondegenerate design so that an estimate of $\boldsymbol{\theta}$ can be obtained from them, and then the sequential procedure described above can begin.

Closely related to this sequential approach is the Bayesian approach which puts a prior distribution $\Pi$ on $\boldsymbol{\theta}$ and maximizes

$$\int \Psi(\boldsymbol{M}(\mu,\boldsymbol{\theta}))\,d\Pi(\boldsymbol{\theta}) \tag{2.22}$$

rather than simply $\Psi(\boldsymbol{M}(\mu,\hat{\boldsymbol{\theta}}))$, where $\hat{\boldsymbol{\theta}}$ is the current estimate of $\boldsymbol{\theta}$. In order to produce Bayesian designs for clinical trials that control the chance of overdosing, Haines et al. (2003) propose to modify the Bayesian criterion (2.22) by including a penalty for high doses. That is, for scalar doses $x$ and an unknown target dose $x^*$ with prior distribution $\rho$ induced by $\Pi$, the problem becomes to find the design measure $\mu$ maximizing (2.22) subject to the constraint

$$P_{\mu,\rho}(x \geq x^*) = \int_{\mathscr{X}} \rho(\{x^* : x \geq x^*\})\,d\mu(x) \leq \varepsilon,$$

for some small chosen value of $\varepsilon > 0$. For clinical trials in which patients are assigned doses sequentially, Haines et al. (2003) further extend their method by adding a sequential aspect by replacing the Bayesian information (2.22) by the sequential analog at the $(k+1)$st stage, given by finding the $(k+1)$st dose $x_{k+1}$ maximizing

$$\int \Psi(\{kM(\mu_k,\boldsymbol{\theta}) + M(\delta_{x_{k+1}},\boldsymbol{\theta})\}/(k+1))d\Pi_k(\boldsymbol{\theta}) \quad \text{subject to}$$
$$P_{\rho_k}(x_{k+1} \geq x^*) = \rho_k(\{x^* : x_{k+1} \geq x^*\}) \leq \varepsilon,$$

where $\mu_k$ is the empirical measure of the first $k$ doses, $\delta_x$ is the degenerate measure at $x$, and $\Pi_k$ and $\rho_k$ are the posterior distributions based on the first $k$ doses and responses.

## 2.4 Phase I Clinical Trials for Relatively Benign Drugs

The primary objective of a Phase I clinical trial is to determine the dose and dosing regimen of a new drug and to collect information about drug-related side effects.

The secondary objective is to use the data collected to evaluate the effectiveness of the treatment. Before the Phase I trial, preclinical in vitro and animal studies are conducted to evaluate toxicity and the pharmacologic actions of the drug, thereby coming up with estimates of a good starting dose for Phase I trials with human subjects. Because of safety considerations for subjects in the trial, the drug is usually initiated at a low, safe dose and sequentially escalated to show safety at a level where some therapeutic response occurs. As noted in Sect. 2.2, the PK/PD models are nonlinear, and nonlinear design theory described in Sect. 2.3.3 is particularly suited for efficient estimation of the model parameters. On the other hand, the ultimate goal is not just estimation of these parameters per se, but to find a dose within the therapeutic window. For relatively benign drugs, Phase I trials involve healthy volunteers from whom intensive blood sampling is conducted over time. The next section describes a different paradigm for Phase I trials of cytotoxic treatments in cancer.

Although intersubject variability is seldom considered at the design stage of Phase I trials, such variability should be examined in the analysis of the data. Thus, while a nonlinear regression model of the type (2.1) with the same $\boldsymbol{\theta}$ for all subjects is assumed at the design stage, nonlinear mixed models of the type (2.4) with subject-specific $\boldsymbol{\theta}_i$ can be used to analyze the data. An example is given by Lai et al. (2006b), in which an orally administered cancer drug, temozolomide, was given to 65 adult patients with advanced cancer in four Phase I trials sponsored by the Schering–Plough Research Institute. Once such trial for treating patients who had advanced cancer that was refractory to standard forms of therapy was reported by Newlands et al. (1992). Each of these 65 patients had 10–15 drug concentration measurements from 10 min to 16 h after a single dose, and a total of 756 concentration measurements were collected. These concentrations were modeled by the one-compartment open model (2.11) to identify the influence of patient characteristics on the PK; the patient covariates forming the vector $\boldsymbol{x}_i$ in the analysis were body surface area, gender, age, and creatinine clearance.

## 2.5 Early Phase Clinical Trials for Cytotoxic Cancer Treatments

### *2.5.1 Up-and-Down and Related Designs*

Up-and-down designs are sequential (or cohort-by-cohort) designs for a discrete dose set in which the “next” dose is always equal or adjacent (the next higher or lower) to the current dose, hence the name “up-and-down.” The original idea is often credited to Dixon and Mood (1948), but an earlier paper by Wilson and Worcester (1943) proposed the idea for clinical uses. These designs have a wide range of applications such as for bioassays, explosives testing, metallurgy, and educational testing. In the dose-finding setting, they have the intuitive appeal of not making large jumps within the dose space. Most up-and-down designs are *random walk rules*, sometimes called *first-order Markov procedures*, which choose the next dose based only on the most recent dose and observation. Because of this simplicity, the properties of random walk rules such as the limiting stationary distribution of the dose allocation and its speed of convergence can be obtained exactly using random walk theory.

*Example 2.2.* The biased coin design of Durham and Flournoy (1994) for estimating the $p$th quantile, $0 < p \le 1/2$, of a response curve using available dose set

$$d_1 < d_2 < \cdots < d_L \tag{2.23}$$

utilizes a biased coin that lands heads with probability $p/(1-p)$ and chooses the $(k+1)$st dose $x_{k+1}$ as follows: If the $k$th dose and observed toxicity are $x_k = d_\ell$ and $y_k \in \{0,1\}$, respectively, then

$$x_{k+1} = \begin{cases} d_{(\ell-1)\vee 1} & \text{if } y_k = 1, \\ d_{(\ell+1)\wedge L} & \text{if } y_k = 0 \text{ and the coin lands heads,} \\ d_\ell & \text{if } y_k = 0 \text{ and the coin lands tails.} \end{cases}$$

Durham and Flournoy (1994) show that, if $F(d) := P(y_k = 1|x_k = d)$ is non-increasing in $d$, the limiting distribution of the dose allocation of this up-and-down rule is unimodal with mode essentially equal to the $p$th quantile of $F(x)$.

To derive the limiting distribution and to understand up-and-down designs more generally, describe an up-and-down design by its *transition probabilities*

$$p_{\ell,m} = P(x_{k+1} = d_m | x_k = d_\ell), \qquad \ell, m \in \{1, \ldots, L\},$$

which is the probability of stepping to the $m$th dose $d_m$, given that the current dose is $d_\ell$. For random walk rules in which dose levels are never skipped, we will have $p_{\ell,m} = 0$ whenever $|\ell - m| > 1$ and hence

$$p_{\ell,\ell-1}\mathbf{1}\{\ell > 1\} + p_{\ell,\ell} + p_{\ell,\ell+1}\mathbf{1}\{\ell < L\} = 1.$$

As a Markov chain, the random walk $\{x_k\}$ has a tri-diagonal transition probability matrix $\boldsymbol{P} = \{p_{\ell,m}\}_{\ell,m=1}^{L}$. Given any initial treatment distribution

$$(P(x_1 = d_1), P(x_1 = d_2), \ldots, P(x_1 = d_L))$$

and a $\boldsymbol{P}$ such that any dose level in (2.23) can be eventually reached from any other, the limiting treatment distribution $\pi_\ell = \lim_{k\to\infty} P(x_k = d_\ell)$, $\ell = 1, \ldots, L$, can be found by solving $L$ linear *balance equations*

$$\pi_\ell = \pi_{\ell-1}p_{\ell-1,\ell}\mathbf{1}\{\ell > 1\} + \pi_\ell p_{\ell,\ell} + \pi_{\ell+1}p_{\ell+1,\ell}\mathbf{1}\{\ell < L\}, \qquad \ell = 1, \ldots, L, \tag{2.24}$$

or equivalently, $\boldsymbol{P}^T \boldsymbol{\pi} = \boldsymbol{\pi}$, where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_L)^T$. The unique solution is

$$\pi_\ell \propto \prod_{j=\ell}^{L-1} \frac{p_{j+1,j}}{p_{j,j+1}}, \qquad \ell = 1, \ldots, L, \tag{2.25}$$

with the convention $\prod_{j=L}^{L-1} = 1$, and the proportionality constant in (2.25) is

$$\pi_L = \left(1 + \sum_{\ell=1}^{L-1} \prod_{j=\ell}^{L-1} \frac{p_{j+1,j}}{p_{j,j+1}}\right)^{-1}. \tag{2.26}$$

The form (2.25) of the solution can be used to find the mode of the limiting distribution $\boldsymbol{\pi}$ since it implies that $\pi_\ell \geq \pi_{\ell-1}$ if and only if $p_{\ell-1,\ell} \geq p_{\ell,\ell-1}$. In particular, for Durham and Flournoy's (1994) biased coin design,

$$p_{\ell-1,\ell} = [1 - F(d_{\ell-1})]p/(1-p) \quad \text{and} \quad p_{\ell,\ell-1} = F(d_\ell),$$

hence

$$\pi_\ell \geq \pi_{\ell-1} \Longleftrightarrow \frac{F(d_\ell)}{1 - F(d_{\ell-1})} \leq \frac{p}{1-p},$$

which shows that this design's limiting distribution has its mode at the discrete $p$th quantile of $F(x)$.

### 3+3 Designs

The widely used 3+3 design (see Korn et al. 1994) can be viewed as a truncated mixture of two up-and-down designs. There are many variations on the 3+3 design, but in its simplest form, the design begins at the lowest dose $d_1$ and, treating patients in cohorts of 3, escalates to the next highest dose level if 0 of 3 experiences toxicity, stays at the same level if 1 of 3 experiences toxicity, and de-escalates or stops the trial if at least 2 of 3 experience toxicity. As pointed out earlier by Storer (1989), these designs are difficult to analyze since even a strict quantitative definition of MTD is lacking, "although it should be taken to mean some percentile of a tolerance distribution with respect to some objective definition of clinical toxicity," and the "implicitly intended" percentile seems to be the 33rd percentile (related to 2/6). In particular, the 3+3 design tends to not have the reliable convergence properties of random walk designs and has been widely criticized in dose-finding clinical trials, such as Reiner et al. (1999) who conclude that its "risk of choosing the incorrect level is large."

### Stochastic Approximation

Another class of designs related to up-and-down designs consists of stochastic approximation procedures (Lai and Robbins 1979; Robbins and Monro 1951), one distinguishing feature being that dose selection under a stochastic approximation procedure will typically converge to a point, whereas random walk up-and-down design points converge to a distribution, as mentioned above. If $F(x) = E(y|x)$ is the mean of the outcome $y = y(x)$ at level (e.g., dose) $x$, then the goal of stochastic approximation is to produce a sequence $\{x_n\}$ of estimates converging to the unique root $x^*$ of the equation $F(x) = y^*$, for given $y^*$. Robbins and Monro (1951) introduced stochastic approximation procedures of the form

$$x_{n+1} = x_n - \frac{(y_n - y^*)}{nb}$$

for some constant $b > 0$ and established that $x_n \to x^*$ in probability under the assumption $\sup_x E[y(x)^2] < \infty$. Moreover, if $b < 2F'(x^*)$, then $\sqrt{n}(x_n - x^*)$ converges to the $N(0, \sigma^2/[b(2F'(x^*) - b)])$ distribution, where $\sigma^2 = \lim_{x\to x^*} \mathrm{Var}[y(x)]$, and the choice of $b$ is thus crucial to the performance of this stochastic approximation procedure (Sacks 1958). Since the optimal choice of $b$ depends on the unknown slope $F'(x^*)$, Lai and Robbins (1979) proposed an adaptive stochastic approximation scheme in which $b$ is replaced by an adaptively chosen sequence $b_n$ that is strongly consistent for $F'(x^*)$. They also study the global cost

$$\sum_{n=1}^{N} (x_n - x^*)^2 \tag{2.27}$$

of the stochastic approximation sequence $\{x_n\}_1^N$ and show that it is of order $\sigma^2 \log N$ as long as $b < 2F'(x^*)$. Although this suggests that adaptive stochastic approximation may be a good choice to use in Phase I dose finding, its “out of the box” application to finite dose spaces and logistic regression models has been less than successful than model-based methods, since it is essentially nonparametric and the sample sizes of Phase I studies are typically small. For example, Bartroff and Lai (2010, 2011) have shown that myopic model-based methods perform considerably better than stochastic approximation in terms of “global” cost functions like (2.27) for $N$ patients and that the performance can be further improved by utilizing approximate dynamic programming techniques, as will be discussed further in Sect. 3.8.

### 2.5.2 Model-Based Designs

Even though 3+3 designs and their variants are widely used in Phase I cancer trials, it has also been widely recognized as unsatisfactory on both ethical and efficiency grounds because it results in mostly subtherapeutic doses and inadequate information to estimate the MTD for a subsequent Phase II trial. To address this difficulty, Eisenhauer et al. (2000) suggest to use (a) methods to determine more informative starting doses, (b) pharmacokinetics-guided dose-escalation methods, and (c) model-based methods for dose determination, which are discussed next.

In model-based methods, a patient's response $y$ to treatment at dose level $x$ is usually modeled by a binary random variable taking values 0 or 1, such that $y = 1$ indicates a DLT and whose distribution depends on $x$ and an unknown vector $\boldsymbol{\theta}$ of parameters through the function

$$F_{\boldsymbol{\theta}}(x) = P(y = 1 | \text{dose} = x).$$

We assume that $F_{\boldsymbol{\theta}}(x)$ is an increasing function of $x$, approaching 0 as $x \to -\infty$ and 1 as $x \to \infty$. In a sequential trial with $n$ patients, we assume that $y_1, \dots, y_n$ are independent, except possibly through the choice of the dose levels $x_1, \dots, x_n$, since $x_{k+1}$ will typically be chosen as a function of the previous doses and responses $(x_1, y_1), \dots, (x_k, y_k)$. As defined above, the MTD is then the $p$th quantile of $F_{\boldsymbol{\theta}}$, that is, MTD $= F_{\boldsymbol{\theta}}^{-1}(p)$. Because of its prevalence in the literature and for simplicity, here we take as our working model the two-parameter logistic regression model

$$F_{\boldsymbol{\theta}}(x) = 1 \Big/ \left(1 + e^{-(\alpha + \beta x)}\right) \tag{2.28}$$

where $\boldsymbol{\theta} = (\alpha, \beta)$. For the two-parameter logistic model, MTD $= [\log(p/(1-p)) - \alpha]/\beta$. The methods that follow are not restricted to the model (2.28) and can be applied to other models such as the probit, gamma, and hyperbolic tangent models (see e.g., O'Quigley et al. 1990).

Noting that the nonparametric approach in stochastic approximation seems too ambitious for moderate sample sizes, Wu (1985) proposed to use a parametric modification of the stochastic approximation scheme in Sect. 2.5.1, taking $x_{k+1}$ to be the $p$th quantile of $F_{\hat{\boldsymbol{\theta}}_k}$, where $\hat{\boldsymbol{\theta}}_k$ is the MLE of $\theta$ based on the doses and responses of the first $k$ patients. O'Quigley et al. (1990) proposed a similar design but from a Bayesian point of view, called the CRM, that estimates the MTD at each stage by the posterior mean of $\boldsymbol{\theta}$ with respect to a chosen prior distribution. O'Quigley (2002) extends CRM to allow early stopping through the use of a sequential stopping rule.

Babb et al. (1998) pointed out that the CRM dose, being the mean of the MTD's posterior distribution, can be viewed as the Bayesian design with respect to squared error loss. That is, letting $\mathscr{F}_k$ denote the information set generated by the first $k$ doses and responses, that is, by $(x_1, y_1), \dots, (x_k, y_k)$, CRM chooses the $(k+1)$st dose $x_{k+1}$ to be that minimizing $E[h(x_{k+1}) | \mathscr{F}_k]$, for

$$h(x) = (x - \text{MTD})^2. \tag{2.29}$$

Babb et al. (1998) suggested that the symmetric nature of the squared error loss or its close relative, the absolute error loss, may not be appropriate for modeling the toxic response to a cancer treatment and proposed the "escalation with overdose control" (EWOC) method, which is a Bayesian design with respect to the asymmetric loss function

$$h(x) = \begin{cases} \omega(\text{MTD} - x) & \text{if } x \leq \text{MTD} \\ (1-\omega)(x - \text{MTD}) & \text{if } x \geq \text{MTD} \end{cases} \tag{2.30}$$

where the chosen constant $0 < \omega < 1/2$ is the so-called *feasibility bound*. Note that this loss function penalizes an overdose $x = \text{MTD} + \delta$ more than an underdose $x = \text{MTD} - \delta$ of the same magnitude $\delta > 0$. EWOC can be shown to be equivalent to estimating the MTD at each stage by the $\omega$th quantile of the posterior distribution of the MTD. In the examples in Babb et al. (1998), $\omega$ is chosen to be slightly less than $p$.

Whereas the step-up/down design in traditional Phase I cancer trials focuses on the safety of patients in the study at the expense of being inefficient for the posttrial estimate of the MTD, there has also been much work on $c$- and $D$-optimal experimental designs for such estimation from binary responses. Haines et al. (2003) proposed sequential Bayesian $c$- and $D$-optimal designs, subject to a prescribed upper-bound $\varepsilon$ on the probability of doses exceeding the MTD, as described in the last paragraph of Sect. 2.3.3.

Despite their shortcomings and the development of alternative Bayesian approaches since 1990, conventional dose-escalation designs are still widely used in Phase I cancer trials because of the ethical issue of safe treatment of patients currently in the trial. However, a Phase I design also has the goal of determining the MTD for a future Phase I cancer trial, and needs an informative experimental design to meet this goal. Von Hoff and Turner (1991) have documented that the overall response rates in Phase I trials are low and that substantial numbers of patients are treated at doses that are retrospectively found to be nontherapeutic. Eisenhauer et al. (2000, p. 685) have pointed out that "with a plethora of molecularly defined antitumor targets and an increasingly clear description of tumor biology, there are now more antitumor candidate therapies requiring Phase I study than ever" and that "unless more efficient approaches are undertaken, Phase I trials may be a rate-limiting step in the process of evaluation of novel anticancer agents." The hybrid designs of Bartroff and Lai (2010) that will be described in Sect. 3.8 were motivated by developing one such "more efficient" approach.

## 2.6 Supplements and Problems

1. *Asymptotic theory of nonlinear least squares and Levenberg–Marquardt shrinkage.*
   Let $\hat{\boldsymbol{\theta}}$ be the least squares estimate of $\boldsymbol{\theta}$ in the nonlinear regression model (2.1). Let $\boldsymbol{\theta}_0$ denote the true value of $\boldsymbol{\theta}$. Assuming $w_j \equiv 1$ in (2.2), we have

$$E[S(\boldsymbol{\theta})] = \sum_{t=1}^{n} [f(\boldsymbol{\theta}, \boldsymbol{x}_t) - f(\boldsymbol{\theta}_0, \boldsymbol{x}_t)]^2 + n\sigma^2, \tag{2.31}$$

recalling that $E(\varepsilon_t) = 0$ and $\mathrm{Var}(\varepsilon_t) = \sigma^2$. Therefore,

$$E[S(\boldsymbol{\theta})] - n\sigma^2 \begin{cases} = 0 & \text{if } \boldsymbol{\theta} = \boldsymbol{\theta}_0 \\ \to \infty & \text{if } \boldsymbol{\theta} \neq \boldsymbol{\theta}_0 \end{cases} \tag{2.32}$$

under the assumption

$$\sum_{t=1}^{\infty} [f_t(\boldsymbol{\theta}) - f_t(\boldsymbol{\theta}_0)]^2 = \infty \text{ for } \boldsymbol{\theta} \neq \boldsymbol{\theta}_0, \tag{2.33}$$

where $f_t(\boldsymbol{\theta}) = f(\boldsymbol{\theta}, \boldsymbol{x}_t)$. In the linear case $f_t(\boldsymbol{\theta}) = \boldsymbol{\theta}^T \boldsymbol{x}_t$, (2.33) is equivalent to the convergence of $(\sum_{t=1}^{n} \boldsymbol{x}_t \boldsymbol{x}_t^T)^{-1}$ to $\mathbf{0}$. Since $\hat{\boldsymbol{\theta}}$ is the minimizer of $S(\boldsymbol{\theta})$, (2.32) suggests that $\hat{\boldsymbol{\theta}}$ is consistent. A rigorous proof involves considering $S(\boldsymbol{\theta})$ as a random function of $\boldsymbol{\theta}$ and requires additional assumptions.

Consistency of $\hat{\boldsymbol{\theta}}$ leads easily to its asymptotic normality since we can approximate $f_t(\hat{\boldsymbol{\theta}})$ by $f_t(\boldsymbol{\theta}_0) + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \nabla f_t(\boldsymbol{\theta}_0)$ when $\hat{\boldsymbol{\theta}}$ is near $\boldsymbol{\theta}_0$, assuming that $\nabla_t f(\boldsymbol{\theta})$ is uniformly continuous in $t$ and $\boldsymbol{\theta}$ belonging to some neighborhood of $\boldsymbol{\theta}_0$. The asymptotic properties of $\hat{\boldsymbol{\theta}}$ are therefore the same as those of ordinary least squares (OLS):

$$\hat{\boldsymbol{\theta}} \approx N\left(\boldsymbol{\theta}_0, \sigma^2 \left(\sum_{t=1}^{n} \hat{\boldsymbol{x}}_t \hat{\boldsymbol{x}}_t^T\right)^{-1}\right), \tag{2.34}$$

where $\hat{\boldsymbol{x}}_t = \nabla f_t(\hat{\boldsymbol{\theta}})$. Moreover, $\sigma^2$ can be consistently estimated by

$$\hat{\sigma}^2 = \sum_{t=1}^{n} \left(y_t - f_t(\hat{\boldsymbol{\theta}})\right)^2 \Big/ n. \tag{2.35}$$

For smooth real-valued functions $g(\boldsymbol{\theta}_0)$, we apply the Taylor expansion $g(\hat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta}_0) \doteq (\nabla g(\boldsymbol{\theta}_0))^T (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ to approximate $g(\hat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta}_0)$ by a linear function, providing the asymptotic normality of $g(\hat{\boldsymbol{\theta}})$ with mean $g(\boldsymbol{\theta}_0)$ and covariance matrix

$$\sigma^2 (\nabla g(\boldsymbol{\theta}_0))^T \left(\sum_{t=1}^{n} \hat{\boldsymbol{x}}_t \hat{\boldsymbol{x}}_t^T\right)^{-1} (\nabla g(\boldsymbol{\theta}_0)). \tag{2.36}$$

The square root of (2.36) also gives the estimated standard error for $g(\hat{\boldsymbol{\theta}})$ if we replace the unknown $\sigma$ and $\boldsymbol{\theta}_0$ in (2.36) by $\hat{\sigma}$ and $\hat{\boldsymbol{\theta}}$. The adequacy of this normal approximation to construct confidence intervals for $g(\boldsymbol{\theta}_0)$ is questionable for highly nonlinear $g$, as the one-term Taylor expansion can be quite poor. An alternative to the asymptotic approximations is the *bootstrap method*, which uses Monte Carlo simulations to obtain standard errors and confidence intervals.

The nonlinear least squares procedure is implemented by many numerical software packages. The following are functions in R: `nls.lm`, `nls`. Since the Gauss–Newton scheme is aborted whenever $\sum_{t=1}^n \hat{\boldsymbol{x}}_t^{(k)} \hat{\boldsymbol{x}}_t^{(k)T}$ is singular or nearly singular, where $\hat{\boldsymbol{x}}_t^{(k)} = \nabla f_t(\hat{\boldsymbol{\theta}}^{(k)})$ and $\hat{\boldsymbol{\theta}}^{(k)}$ is defined in Sect. 2.1.1. It is desirable to avoid such difficulties in matrix inversion. This has led to the modification that replaces $\sum_{t=1}^n \hat{\boldsymbol{x}}_t^{(k)} \hat{\boldsymbol{x}}_t^{(k)T}$ by $\sum_{t=1}^n \hat{\boldsymbol{x}}_t^{(k)} \hat{\boldsymbol{x}}_t^{(k)T} + \kappa \boldsymbol{D}$ for the OLS estimate in the $k$th iteration of the Gauss–Newton algorithm, corresponding to using shrinkage as in ridge regression. Here $\boldsymbol{D}$ is a diagonal matrix whose diagonal elements are the same as those of $\sum_{t=1}^n \hat{\boldsymbol{x}}_t^{(k)} \hat{\boldsymbol{x}}_t^{(k)T}$, proposed by Marquardt as a refinement of an earlier proposal $\boldsymbol{D} = \boldsymbol{I}$ by Levenberg.

2. *Generalized linear mixed models.*
The NONMEM in Sect. 2.1.2 have their counterparts for generalized linear models. These are called *generalized linear mixed models* (GLMM) and were introduced by Breslow and Clayton (1993) for longitudinal data $Y_{it}$ to enhance generalized linear models by allowing subject-specific regression parameters $\boldsymbol{b}_i$, called "random effects," thereby extending mixed effects models in linear regression to GLMM. The GLMM assumes the $y_{it}$ to be conditionally independent given the observed covariates $\boldsymbol{x}_{it}$ and $\boldsymbol{z}_{it}$ and such that $y_{it}$ has a conditional density of the form

$$f(y|\boldsymbol{b}_i, \boldsymbol{z}_{it}, \boldsymbol{x}_{it}) = \exp\{[y\theta_{it} - \psi(\theta_{it})]/\sigma + c(y,\sigma)\}, \tag{2.37}$$

in which $\sigma$ is a dispersion parameter and $\mu_{it} = d\psi/d\theta|_{\theta=\theta_{it}}$ satisfies

$$\mu_{it} = g^{-1}\left(\boldsymbol{\beta}^T \boldsymbol{x}_{it} + \boldsymbol{b}_i^T \boldsymbol{z}_{it}\right), \tag{2.38}$$

where $g^{-1}$ is the inverse of a monotone link function $g$, as in the standard generalized linear models for which $\mu_{it} = g^{-1}(\boldsymbol{\beta}^T \boldsymbol{x}_{it})$. The case $g = d\psi/d\theta$ is called the "canonical link," The random effects $\boldsymbol{b}_i$ can contain an intercept term $a_i$ by augmenting the covariate vector to $(1, \boldsymbol{z}_{it})$ in case $a_i$ is not included in $\boldsymbol{b}_i$; $\boldsymbol{\beta}$ is a vector of fixed effects and can likewise contain an intercept term. The density function (2.37) with $\sigma = 1$ is that of an exponential family, which includes the Bernoulli and normal distributions as special cases. Brelow and Clayton assume the $\boldsymbol{b}_i$ in (2.38) to have a common normal distribution with mean 0 and covariance matrix $\boldsymbol{\Sigma}$ that depends on an unknown parameter vector $\boldsymbol{\alpha}$.

The likelihood function of the GLMM defined by (2.37) and (2.38) is of the form $\prod_{i=1}^{n} L_i(\sigma, \boldsymbol{\alpha}, \boldsymbol{\beta})$, where

$$L_i(\sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \int \left\{ \prod_{t=1}^{T} f(y_{it}; \theta_{it}, \sigma) \right\} \phi_{\boldsymbol{\alpha}}(\boldsymbol{b})\, d\boldsymbol{b}, \tag{2.39}$$

in which $\phi_{\boldsymbol{\alpha}}$ denotes the normal density function with mean 0 and covariance matrix depending on an unknown parameter $\boldsymbol{\alpha}$. Analogous to NONMEM described in Sect. 2.1.2, there are three methods to compute the likelihood function, the maximizer of which gives the MLE of $\sigma$, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$:

(a) *Laplace's approximation.* Letting $e^{l_i(\boldsymbol{b}|\sigma, \boldsymbol{\alpha}, \boldsymbol{\beta})}$ be the integrand in the right-hand side of (2.39), Laplace's asymptotic formula for integrals yields the approximation

$$\int e^{l_i(\boldsymbol{b}|\sigma, \boldsymbol{\alpha}, \boldsymbol{\beta})} d\boldsymbol{b} \approx$$
$$(2\pi)^{q/2} \left\{ \det \left[ -\ddot{l}_i \left( \hat{\boldsymbol{b}}_i | \sigma, \boldsymbol{\alpha}, \boldsymbol{\beta} \right) \right] \right\}^{-1/2} \exp \left\{ l_i \left( \hat{\boldsymbol{b}}_i | \sigma, \boldsymbol{\alpha}, \boldsymbol{\beta} \right) \right\}, \tag{2.40}$$

where $q$ is the dimension of $\boldsymbol{b}_i$, $\hat{\boldsymbol{b}}_i = \hat{\boldsymbol{b}}_i(\sigma, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is the maximizer of $l_i(\boldsymbol{b}|\sigma, \boldsymbol{\alpha}, \boldsymbol{\beta})$ and $\ddot{l}_i$ denotes the Hessian matrix consisting of second partial derivatives of $l_i$ with respect to the components of $\boldsymbol{b}$. The R package `lme4` computes the MLE by using the Laplace approximation (2.40) or the restricted pseudo-likelihood approach proposed by Wolfinger and O'Connell (1993), as the user-specified option.

(b) *Gauss–Hermite quadrature.* Laplace's asymptotic formula (2.40) is derived from the asymptotic approximation of $l_i$ by a quadratic function of $\boldsymbol{b}$ in a small neighborhood of $\hat{\boldsymbol{b}}_i$ as $\lambda_{\min}(-\ddot{l}_i(\hat{\boldsymbol{b}}_i|\sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}) \to \infty$, where $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalue of a symmetric matrix. Therefore, such formula may produce significant approximation error for $L_i$ if the corresponding $\lambda_{\min}(-\ddot{l}_i(\hat{\boldsymbol{b}}_i|\sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}))$ is not sufficiently large. One way to reduce the possible approximation error is to compute $L_i$ by using an adaptive Gauss–Hermite quadrature rule, as in Liu and Pierce (1994). The software package SAS uses adaptive Gauss–Hermite quadrature in the *NLMIXED* procedure to compute (2.39); the R package `lmer()` also uses Gaussian quadrature to compute (2.39) but only for certain special cases of the exponential family (2.37). The numerical integration procedures demand a much higher computational effort and become computationally infeasible when $n$ or $q$ is large. To circumvent the issue of high-dimensional numerical integration, some authors propose putting prior distributions on the unknown parameters and estimate them by the Markov chain Monte Carlo (MCMC) method in a Bayesian way; see Berry et al. (2011) for logistic mixed models. The performance of the MCMC method, however, depends on how the prior parameters are set as well as the convergence rate of the Markov chain to its stationary distribution, which may not even exist. Yafune et al. (1998) use direct Monte Carlo integration but point out that the computational time may be too long to be of practical interest.

(c) *Hybrid method.* This is basically the same as the hybrid method for NONMEM, as pointed out by Lai and Shih (2003b, Sect. 5).

3. *Dose individualization and population PK/PD.*
Several physiologic (e.g., maturation of organs in infants) and pathologic (e.g., kidney failure, heart failure) processes require dosage adjustments in individual patients to modify specific PK parameters. Two basic parameters in this connection are *clearance* (a measure of the ability of the body to eliminate the drug) and *volume of distribution* (a measure of the apparent space in the body available to contain the drug). Drug clearance principles are similar to clearance concepts in renal physiology, in which creatinine or urea clearance is defined as the rate of elimination of the compound in the urine relative to the plasma concentration. Thus, clearance CL of a drug is the rate of elimination by all routes relative to the concentration $C$ of the drug in a biologic fluid; it is perhaps the most important PK parameter to be considered in defining a rational drug dosage regimen. In most cases, the clinician would like to maintain steady-state drug concentrations $C_{ss}$ within a known therapeutic window. Steady state will be achieved when the dosing rate (rate of active drug entering the systemic circulation) equals the rate of drug elimination. Therefore,

$$\text{Dosing rate} = \text{CL} \times C_{ss}.$$

The two major sites of drug elimination are the kidneys and the liver. Clearance of unchanged drug in the urine represents renal clearance. Within the liver, drug elimination occurs via biotransformation of the drug to one or more metabolites, or excretion of unchanged drug into the bile, or both. When no other organs are involved in elimination of the drug, $\text{CL} = \text{CL}_{\text{renal}} + \text{CL}_{\text{liver}}$ since the liver and kidneys work in parallel. The volume of distribution $(V)$ is defined as

$$V = \text{Amount of drug in body}/C,$$

where $C$ is the concentration of the drug in blood or plasma, depending on the fluid measured. It reflects the apparent space available in both the general circulation and the tissue of distribution. It does not represent a real volume but should be regarded as the size of the pool of blood fluids that would be required if the drug were distributed equally throughout all parts of the body. From mass balance and steady-state considerations, $V$ is related to clearance via $\text{CL} = k_e V$, where $k_e$ is the elimination rate constant. Note that both $V$ and the elimination rate $k_e$ appear in the one-compartment open model (2.11). Allowing these parameters and the absorption rate $k_a$ in (2.11) to be subject-specific leads to a NONMEM that is used in the second paragraph of Sect. 2.4.

Dose individualization is a major practical goal of population PK. Since the efficacy and toxicity of a drug are directly related to drug concentrations at a target site, which are generally not available but for which blood concentrations are often good surrogates, criteria for determining the dose and dosing regimen for a specific subject often involve functions of the subject's concentrations or functions of the subject's parameter vector $\boldsymbol{\theta} = g(\boldsymbol{x}, \boldsymbol{\beta}) + \boldsymbol{b}$ in (2.4). The subject's blood samples are often too sparse to provide an adequate estimate of $\boldsymbol{\theta}$. The empirical Bayes approach considered by Lai et al. (2006b) borrows information from healthy volunteers in Phase I studies who have undergone intensive blood sampling and also from clinical patients for whom intensive blood sampling is not feasible. Combining an individual patient's characteristics (as measured by $\boldsymbol{x}$) and sparse concentration data with a large database from other subjects is one of the main motivations for building population PK models. Making use of the hybrid method, the last two paragraphs of Sect. 2.2 discuss how empirical Bayes estimates of $h(\theta)$ can be computed from (a) the patient's data and (b) the population model fitted from other subjects' data.

An emerging trend in cancer therapeutics is to use biomarkers to personalize the treatment and treatment strategy for cancer patients; see Lai et al. (2012b). Personalization (or individualization) of treatments again falls in the domain of nonlinear/generalized LME models. Biomarker-guided personalized therapies for cancer require innovations in design and analysis of early phase and Phase III confirmatory clinical trials and may eventually lead to major breakthroughs in the methodology.

4. In the model $Y = \alpha + \beta x + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$ and $-1 \le x \le 1$, sketch the convex hull $\mathscr{S}$ and use Elfving's method to find the optimal design for estimating (a) the slope $\beta$ and (b) the mean response $\alpha + \beta x_0$ at $x = x_0$, for arbitrary $-1 \le x_0 \le 1$.
5. In the setting of the example in Sect. 2.3.2, fix a value of $a > 0$ and compute the value of
$$\frac{\boldsymbol{c}^T \boldsymbol{M}(\tilde{\mu})^{-1} \boldsymbol{c}}{\boldsymbol{c}^T \boldsymbol{M}(\mu^*)^{-1} \boldsymbol{c}} \qquad \text{for } x_0 = i \cdot a/5,\ i = 1, \ldots, 5,$$
where $\tilde{\mu}$ is the design putting weight $1/3$ at each of the points $x = 0, a/2$, and $a$, and $\mu^*$ is the $\boldsymbol{c}$-optimal design found in the example.
6. Verify (2.18) and (2.19), and show that $p$ given by (2.19) is in $[0,1]$ for arbitrary $a > 0$ and all $0 < x_0 \le \gamma a$.
7. In the example in Sect. 2.3.2, find the $\boldsymbol{c}$-optimal design if $x_0$ is allowed to exceed $a$.
8. Find the Fisher information matrix
$$E\left[-\frac{\partial^2 p(Y|\boldsymbol{x}, \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right]$$
for the logistic regression model

$$P(Y=1|\boldsymbol{x},\boldsymbol{\theta}) = 1/(1+e^{-(\alpha+\beta x)}), \quad P(Y=0|\boldsymbol{x},\boldsymbol{\theta}) = 1 - P(Y=1|\boldsymbol{x},\boldsymbol{\theta}),$$

where $\boldsymbol{x} = (1,x)^T$ and $\boldsymbol{\theta} = (\alpha,\beta)^T$.

9. Making use of the asymptotic theory of nonlinear least squares described in (2.34) and (2.36), explain how optimal linear design theory can be used to construct locally optimal designs in nonlinear regression models.
10. *Discrete dose levels in Phase I cancer trials.*
As pointed out in Sect. 2.5, because of the traditional practice of using up-and-down designs in Phase I cancer trials, the dose levels in dose-finding studies of cancer drugs are usually chosen before the trial as a finite set $\Lambda = \{\lambda_1, \dots, \lambda_d\}$ of possible doses, where $\lambda_1 < \lambda_2 < \cdots < \lambda_d$, unlike the continuous doses we have assumed in Sect. 2.5.2. In this case the MTD has to be redefined as

$$\eta = \begin{cases} \max\{\lambda \in \Lambda : F_{\boldsymbol{\theta}}(\lambda) \le p\} & \text{if } F_{\boldsymbol{\theta}}(\lambda_i) \le p \text{ for some } i, \\ \lambda_1 & \text{otherwise.} \end{cases} \tag{2.41}$$

In many dose-finding trials, the number of discrete dose levels is relatively small, so one can use more robust order-restricted models of toxicity versus dose than the logistic regression model (2.28). Yin and Yuan (2009) have proposed a Bayesian model averaging design based on the monotone dose–toxicity relationship.

# Chapter 3
# Sequential Testing Theory and Stochastic Optimization Over Time

The first seven sections of this chapter give an overview of the theory of fully sequential tests, starting with simple hypotheses involving likelihood ratio statistics and then extending the theory to composite hypotheses via generalized likelihood ratio (GLR) statistics. This theory is of particular relevance to Sect. 5.2 on vaccine clinical trials. The theory is modified in Sect. 3.5 for group sequential designs and later for adaptive designs in Sect. 8.2.3. The classic result of Wald and Wolfowitz on the optimality of the sequential probability ratio test is derived in Sect. 3.6 by using dynamic programming that is also introduced in that section for general stochastic optimization over time. Dynamic programming is often difficult to implement directly for nonlinear models and approximate dynamic programming (ADP) methods have been developed to overcome the computational and analytical difficulties. Section 3.8 introduces approximate programming and applies it to resolve the treatment versus experimentation dilemma in Phase I cancer trials.

## 3.1 Likelihood Ratio Statistics and Likelihood Ratio Identity

Let $X_1, X_2, \dots$ be observations drawn from a probability measure under which $g_1$ is the marginal density of $X_1$ and for $i \geq 2$, the conditional distribution of $X_i$ given $X_1, \dots, X_{i-1}$ has density function $g_i(\cdot|X_1, \dots, X_{i-1})$ with respect to some measure $\nu_i$. To test a simple null hypothesis $H_0 : g_i = p_i$ versus a simple alternative hypothesis $H_1 : g_i = q_i$, the likelihood ratio test based on a sample $X_1, \dots, X_n$ of fixed size $n$ rejects $H_0$ if

$$L_n = \prod_{i=1}^{n} \{q_i(X_i|X_1, \dots, X_{i-1}) / p_i(X_i|X_1, \dots, X_{i-1})\} \tag{3.1}$$

J. Bartroff et al., *Sequential Experimentation in Clinical Trials,*
Springer Series in Statistics 298, DOI 10.1007/978-1-4614-6114-2_3,

exceeds the threshold $c$ for which the type I error probability $P_0\{L_n \geq c\}$ is equal to some prescribed level $\alpha$. The Neyman–Pearson lemma says that among all tests whose type I error probabilities do not exceed $\alpha$, the *likelihood ratio test* is most powerful in the sense that it maximizes the probability of rejecting the null hypothesis under the alternative hypothesis. One can also control the type II error probability (or $1-$power) of the likelihood ratio test by choosing the sample size $n$ appropriately. Instead of using a fixed sample size $n$, an alternative approach is to continue sampling until $L_n$ shows enough evidence against $H_0$ or $H_1$. In the case of i.i.d. $X_t$, this is the idea behind Wald's SPRT in Sect. 1.2.

The likelihood ratio statistics in (3.1) are closely related to change of measures; in fact, (3.1) is the likelihood ratio (Radon–Nikodym derivative) of the measure $Q$ (with conditional densities $q_i$) relative to the measure $P$ (with conditional densities $p_i$). The optimality of the likelihood ratio test (Neyman–Pearson lemma) is a consequence of this change of measures. Regarding a test of $H_0$ versus $H_1$ as a function $\varphi$ from the sample space $\mathscr{X}$ into $[0,1]$ (i.e., $\varphi(X_1,\ldots,X_n)$ is the probability of rejecting $H_0$), the likelihood ratio test $\varphi^*$ can be characterized by $\varphi^* = 1$ if $L_n > c$ and $\varphi^* = 0$ if $L_n < c$. Since $(\varphi^* - \varphi)(L_n - c) \geq 0$, $E_0\{(\varphi^* - \varphi)L_n\} \geq cE_0(\varphi^* - \varphi)$. Changing the measures from $P_1$ to $P_0$ then yields

$$E_1(\varphi^* - \varphi) = E_0\{(\varphi^* - \varphi)L_n\} \geq cE_0(\varphi^* - \varphi), \tag{3.2}$$

in which the equality is a special case of Wald's likelihood ratio identity described below. From (3.2), it follows that if the type I error of $\varphi$ does not exceed that of $\varphi^*$ (i.e., $E_0\varphi \leq E_0\varphi^*$), then $E_1\varphi^* \geq E_1\varphi$, proving the Neyman–Pearson lemma.

Wald (1945) extended the preceding argument involving change of measures to a *likelihood ratio identity*, which can be stated generally as follows. A stopping time $T$ is a positive integer-valued random variable such that the event $\{T = n\}$ depends on the observations $X_1,\ldots,X_n$ up to time $n$. Let $F$ be an event that depends on the observations $X_1,\ldots,X_T$ up to a stopping time $T$. The likelihood ratio identity states that

$$\begin{aligned} Q\left(F \cap \{T < \infty\}\right) &= E_P\left\{L_T I_{F\cap\{T<\infty\}}\right\}, \\ P\left(F \cap \{T < \infty\}\right) &= E_Q\left\{L_T^{-1} I_{F\cap\{T<\infty\}}\right\}. \end{aligned} \tag{3.3}$$

Not only does (3.3) provide a powerful tool to analyze error probabilities in sequential analysis but it also plays a basic role for *importance sampling* in Monte Carlo computation of the probabilities of rare events under the measure $P$. Direct Monte Carlo may fail to generate the event after many simulation runs, but changing the measure to $Q$ can generate the event in a manageable number of simulations.

## 3.2 Wald's SPRT and Its Error Probabilities

In the first paragraph of Sect. 1.2, we have described Wald's SPRT which was introduced by Wald to test a simple null hypothesis versus a simple alternative hypothesis based on i.i.d. observations. Like the Neyman–Pearson test, the SPRT also uses likelihood ratio statistics. But instead of using a fixed sample size as in the Neyman–Pearson test, the SPRT uses a stopping rule $N$ given by (1.1), in which the stopping boundaries $A$ and $B$ for the likelihood ratio statistics depend on the type I and type II error probabilities $\alpha$ and $\beta$. Wald (1945) developed simple formulas for $\alpha$ and $\beta$ in terms of $A$ and $B$, from which he could solve for $A$ and $B$ when $\alpha$ and $\beta$ are given. These formulas are corollaries of the likelihood ratio identity and therefore apply to more general settings than the i.i.d. case considered in (1.1). In fact, we can generalize (1.1) to $N = \inf\{n \geq 1 : L_n \notin (A,B)\}$, in which $L_n$ is defined by (3.1) for general, not necessarily independent $X_i$. In this general framework, if $P_i(N < \infty) = 1$ for $i = 0,1$, then (3.3) yields

$$P_0\{L_N \geq B\} \leq B^{-1}P_1\{L_N \geq B\}, \quad P_1\{L_N \leq A\} \leq AP_0\{L_N \leq A\}, \tag{3.4}$$

and $\leq$ can be replaced by $=$ in (3.4) if $L_N$ has to fall on either boundary exactly (i.e., there is no "overshoot"). Ignoring overshoots, (3.4) treated as the approximate equalities can be used to solve for the error probabilities $\alpha = P_0\{L_N \geq B\}$ and $\beta = P_1\{L_N \leq A\}$:

$$\alpha \approx \frac{1-A}{B-A}, \qquad \beta \approx A\left(\frac{B-1}{B-A}\right). \tag{3.5}$$

Writing (3.5) as equations of $(A,B)$ in terms of $\alpha$ and $\beta$ yields the explicit solutions for the stopping boundaries:

$$A \approx \frac{\beta}{1-\alpha}, \qquad B \approx \frac{1-\beta}{\alpha}. \tag{3.6}$$

## 3.3 Wald's Equation and Lower Bounds of Wald and Hoeffding

Besides the likelihood ratio identity, another tool developed by Wald to analyze the SPRT is *Wald's equation*

$$E\left(\sum_{i=1}^{T} X_i\right) = \mu ET, \tag{3.7}$$

in which $X_1, X_2, \ldots,$ are i.i.d. random variables with mean $\mu$ and $T$ is a stopping time such that $ET < \infty$ in the case $\mu = 0$. To prove (3.7), note that

$$E\left(\sum_{i=1}^{T} X_i\right) = E\left(\sum_{i=1}^{\infty} X_i I_{\{T \geq i\}}\right) = \sum_{i=1}^{\infty}(EX_i)P(T \geq i),$$

since $X_i$ is independent of $\{T < i\}$ that only involves $X_1, \ldots, X_{i-1}$. From this and $\mu = EX_i$, (3.7) follows.

Let $T$ be the stopping rule of a test of $H_0$ versus $H_1$ with error probabilities $\alpha, \beta$, and let $\delta$ denote its terminal decision rule ($\delta = j$ if $H_j$ is accepted, $j = 0, 1$). Wald's likelihood ratio identity yields

$$\begin{aligned}\alpha = P_0(\delta = 1) &= E_1\left\{L_T^{-1} I(\delta = 1)\right\}\\ &= E_1\{e^{-\log L_T} | \delta = 1\} P_1(\delta = 1)\\ &\geq \exp\{-E_1(\log L_T | \delta = 1)\} P_1(\delta = 1)\\ &= \exp\left\{-E_1\left[(\log L_T) I(\delta = 1)\right] \big/ (1-\beta)\right\}(1-\beta),\end{aligned}$$

in which $\geq$ follows from Jensen's inequality. Therefore,

$$-E_1\left[(\log L_T) I(\delta = 1)\right] \leq (1-\beta)\log\left(\alpha/(1-\beta)\right).$$

A similar argument also gives $-E_1[(\log L_T) I(\delta = 0)] \leq \beta \log((1-\alpha)/\beta)$. Adding the two inequalities then yields

$$(1-\beta)\log\frac{\alpha}{1-\beta} + \beta\log\frac{1-\alpha}{\beta} \geq -E_1(\log L_T) = -E_1\left(\sum_{t=1}^{T}\log\frac{f_1(X_t)}{f_0(X_t)}\right) = -\mu_1 E_1 T,$$

by Wald's equation (assuming that $E_1 T < \infty$), where $\mu_i = E_i[\log(f_1(X_1)/f_0(X_1))]$. This yields lower bound for $E_1(T)$, and a similar argument can be used to prove that for $E_0(T)$, that is,

$$\begin{aligned}E_1(T) &\geq \mu_1^{-1}\left\{(1-\beta)\log\left(\frac{1-\beta}{\alpha}\right) + \beta\log\left(\frac{\beta}{1-\alpha}\right)\right\},\\ E_0(T) &\geq (-\mu_0)^{-1}\left\{(1-\alpha)\log\left(\frac{1-\alpha}{\beta}\right) + \alpha\log\left(\frac{\alpha}{1-\beta}\right)\right\},\end{aligned} \tag{3.8}$$

noting that $\mu_1 > 0 > \mu_0$ under the assumption $P_i\{f_1(X_1) \neq f_0(X_1)\} > 0$ for $i = 0, 1$. Since the right-hand sides of (3.8) are Wald's approximations, ignoring overshoots, to $E_1(N)$ and $E_0(N)$, Wald (1945) conjectured the following *optimality theorem*: The SPRT minimizes both $E_0(T)$ and $E_1(T)$ among all tests that have type I and type II errors $\alpha$ and $\beta$, respectively. This theorem was later proved by Wald and Wolfowitz (1948) by dynamic programming arguments that will be described in Sect. 3.6.

Hoeffding (1960) extended Wald's arguments to derive lower bounds for $E(T)$ when the sequential test of $H_0$ versus $H_1$ has error probabilities $\alpha$ and $\beta$, under another measure that has density function $f$ with respect to $\nu$. One such lower bound involves the Kullback–Leibler information number

$$I(f, f_i) = E[\log(f(X_1)/f_i(X_1))]. \tag{3.9}$$

Let $\tau^2 = E\{\log[f_1(X_1)/f_0(X_1)] - I(f,f_0) + I(f,f_1)\}^2$, $\zeta = \max\{I(f,f_0), I(f,f_1)\}$. Hoeffding's lower bound is

$$E(T) \geq \left\{\left[-\zeta \log(\alpha+\beta) + (\tau/4)^2\right]^{1/2} - \tau/4\right\}^2 \Big/ \zeta^2. \tag{3.10}$$

## 3.4 Lorden's 2-SPRT and Sequential GLR Tests

For fixed sample size tests, a first step to extend the Neyman–Pearson theory from simple to composite hypotheses is to consider one-sided composite hypotheses of the form $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ in the case of parametric families with monotone likelihood ratio in a real parameter $\theta$. In this case, the level-$\alpha$ Neyman–Pearson test of $H : \theta = \theta_0$ versus $K : \theta = \theta_1 (> \theta_0)$ does not depend on $\theta_1$ and therefore can be used to test $H_0$ versus $H_1$. In the sequential setting, however, we cannot reduce the optimality considerations for one-sided composite hypotheses to those for simple hypotheses even in the presence of the monotone likelihood ratio. This led Kiefer and Weiss (1957) to consider the problem of minimizing the expected sample size at a given parameter $\theta^*$ subject to error probability constraints at $\theta_0$ and $\theta_1$. Using dynamic programming arguments that will be described in Sect. 3.7, Lorden (1976) showed that a nearly optimal solution to the Kiefer–Weiss problem is a 2-SPRT with stopping rule of the form

$$T^* = \inf\left\{n : \prod_{i=1}^{n} \left(f_{\theta^*}(X_i)/f_{\theta_0}(X_i)\right) \geq A_0 \quad \text{or} \right.$$
$$\left. \prod_{i=1}^{n} \left(f_{\theta^*}(X_i)/f_{\theta_1}(X_i)\right) \geq A_1 \right\}, \tag{3.11}$$

rejecting $H : \theta = \theta_0$ if $\prod_{i=1}^{T^*}(f_{\theta^*}(X_i)/f_{\theta_0}(X_i)) \geq A_0$ and rejecting $K : \theta = \theta_1$ if the other boundary is crossed upon stopping. He also showed that the 2-SPRT asymptotically attains Hoeffding's lower bound for $E_{\theta^*}(T)$ and provided numerical results showing that $E_{\theta^*}(T^*)$ exceeds the lower bound by at most 10%.

Ideally, $\theta^*$ in (3.11) should be chosen to be the true parameter value $\theta$ which is, however, unknown. Consider the exponential family of density functions $f_\theta(x) = \exp\{\theta x - \psi(\theta)\}$ with respect to some measure on the real line. Replacing $\theta$ by its maximum likelihood ratio estimate $\hat{\theta}_n$ at stage $n$ leads to the sequential GLR test of $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta \geq \theta_1$ that stops sampling at stage

$$\tau = \inf\left\{n : \hat{\theta}_n > \theta_0 \text{ and } \prod_{i=1}^{n} \left(f_{\hat{\theta}_n}(X_i)/f_{\theta_0}(X_i)\right) \geq A_0^{(n)} \quad \text{or} \right.$$
$$\left. \hat{\theta}_n < \theta_1 \text{ and } \prod_{i=1}^{n} \left(f_{\hat{\theta}_n}(X_i)/f_{\theta_1}(X_i)\right) \geq A_1^{(n)} \right\}. \tag{3.12}$$

Note that the first likelihood ratio in (3.12) is the GLR statistic for testing $\theta_0$ and the second one is that for testing $\theta_1$. The test rejects $H_i$ upon stopping if the GLR statistic for testing $\theta_i$ exceeds $A_i^{(n)}$ $(i = 0, 1)$. The test, with $A_0^{(n)} = A_1^{(n)} = 1/c$, has been derived by Schwarz (1962) as an asymptotic solution to the Bayes problem of testing $H_0$ versus $H_1$ with 0–1 loss and cost $c$ per observation, as $c \to 0$ while $\theta_0$ and $\theta_1$ are fixed. For the case of a normal mean $\theta$, Chernoff (1961, 1965) derived a different and considerably more complicated approximation to the Bayes test of $H_0'$ : $\theta < \theta_0$ versus $H_1' : \theta > \theta_0$. In fact, setting $\theta_1 = \theta_0$ in Schwarz's test does not yield Chernoff's test. This disturbing discrepancy between the asymptotic approximations of Schwarz (assuming an indifference zone) and Chernoff (without an indifference zone separating the one-sided hypotheses) was resolved by Lai (1988), who gave a unified solution (to both problems) that uses a stopping rule of the form

$$N(g,c) = \inf\left\{ n : \max\left[ \sum_{i=1}^{n} \log \frac{f_{\hat\theta_n}(X_i)}{f_{\theta_0}(X_i)}, \sum_{i=1}^{n} \log \frac{f_{\hat\theta_n}(X_i)}{f_{\theta_1}(X_i)} \right] \ge g(cn) \right\} \tag{3.13}$$

for testing $H_0$ versus $H_1$ and setting $\theta_1 = \theta_0$ in (3.13) for the test of $H_0'$ versus $H_1'$. The function $g$ in (3.13) satisfies $g(t) \sim \log t^{-1}$ as $t \to 0$ and is the boundary of an associated optimal stopping problem for the Wiener process. By solving the latter problem numerically, Lai (1988) also gave a closed-form approximation to the function $g$. Further details are given in Sect. 3.7, where applications of Wald's likelihood identity and Hoeffding's lower bound, and connections between GLR statistics and posterior probabilities of the null hypothesis, are also given.

This unified theory for composite hypotheses provides a bridge between asymptotically optimal sequential and fixed sample size tests. In the fixed sample size case, the Neyman–Pearson approach replaces the likelihood ratio by the GLR, which is also used in (3.13) for the sequential test. Since the accuracy of $\hat\theta_n$ as an estimate of $\theta$ varies with $n$, (3.13) uses a time-varying boundary $g(\varepsilon n)$ instead of the constant boundary in (3.11) (with $A_0 = A_1$) where $\theta$ is completely specified. Simulation studies and asymptotic analysis have shown that $\hat N$ is nearly optimal over a broad range of parameter values $\theta$, performing almost as well as (3.11) that assumes $\theta$ to be known; see Lai (1988). This broad range covers both fixed alternatives, at which the expected sample size is of the order $O(|\log \varepsilon|)$, and local alternatives $\theta$ approaching $\theta_0$ as $\varepsilon \to 0$, at which the expected sample size divided by $|\log \varepsilon|$ tends to $\infty$. In other words, $N(g,c)$ can adapt to the unknown $\theta$ by learning it during the course of the experiment and incorporating the diminishing uncertainties in its value into the stopping boundary $g(\varepsilon n)$. Lai and Zhang (1994) have extended these ideas to construct nearly optimal sequential GLR tests of one-sided hypotheses concerning some smooth scalar function of the parameter vector in multiparameter exponential families, with an indifference zone separating the null and alternative hypotheses and also without an indifference zone. Lai (1997) has provided further extension to a general class of loss functions and prior distributions.

In practice, one often imposes an upper bound $M$ and also a lower bound $m$ on the total number of observations. With $M/m \to b > 1$ and $\log\alpha \sim \log\beta$, we can replace the time-varying boundary $g(cn)$ in (3.13) by a constant threshold $c$ since $g(t) \sim \log t^{-1}$ and $\log n = \log m + O(1)$ for $m \le n \le M$. The test of $H: \theta = \theta_0$ with stopping rule

$$\tilde{N} = \inf\left\{ n \ge m : \left[\prod_{i=1}^{n} f_{\hat{\theta}_n}(X_i)\right] \Big/ \left[\prod_{i=1}^{n} f_{\theta_0}(X_i)\right] \ge e^c \right\} \wedge M, \tag{3.14}$$

which corresponds to (3.13) with $\theta_1 = \theta_0$, $g(cn)$ replaced by $c$, and $n$ restricted between $m$ and $M$, is called a *repeated GLR test*. The test rejects $H$ if the GLR statistic exceeds $e^c$ upon stopping. The repeated significance test of Armitage et al. (1969) described in Sect. 1.2 is a repeated GLR test. Whereas (3.14) considers the simple null hypothesis $\theta = \theta_0$ in the univariate case, it is straightforward to extend the repeated GLR test to multivariate $\theta$ and composite null hypothesis $H_0 : \theta \in \Theta_0$, by simply replacing $\prod_{i=1}^{n} f_{\theta_0}(X_i)$ in (3.14) by $\sup_{\theta\in\Theta_0}\prod_{i=1}^{n} f_{\theta}(X_i)$.

## 3.5 Modifications for Group Sequential Testing

Lai and Shih (2004) have modified the preceding theory for group sequential tests in a one-parameter exponential family $f_\theta(x) = e^{\theta x - \psi(\theta)}$ of density functions, for which Hoeffding's lower bound (3.10) can be expressed as

$$\begin{aligned} E_\theta(T) \ge -\zeta^{-1}\log(\alpha+\beta) - \left(\zeta^{-2}\sigma/2\right)\{(\sigma/4)^2 - \zeta\log(\alpha \quad +\beta)\}^{1/2} \\ +\zeta^{-2}\sigma^2/8 \end{aligned} \tag{3.15}$$

where $\sigma^2 = (\theta_1-\theta_0)^2\psi''(\theta) = \mathrm{Var}_\theta\{(\theta_1-\theta_0)X_i\}$, $\zeta = \max\{I(\theta,\theta_0), I(\theta,\theta_1)\}$, and

$$I(\theta,\lambda) = E_\theta\left[\log\{f_\theta(X_i)/f_\lambda(X_i)\}\right] = (\theta-\lambda)\psi'(\theta) - (\psi(\theta)-\psi(\lambda)) \tag{3.16}$$

is the Kullback–Leibler information number. The lower bound (3.15) does not take into consideration the fact that $T$ can assume only several possible values in the case of group sequential designs. The first step of Lai and Shih (2004, p. 509) is to take this into consideration by providing an asymptotic lower bound for $T$ in the following theorem. Let $n_0 = 0$.

**Theorem 3.1.** *Suppose the possible values of $T$ are $n_1 < \cdots < n_k$, such that*

$$\liminf (n_i - n_{i-1})/|\log(\alpha+\beta)| > 0 \tag{3.17}$$

*as* $\alpha+\beta \to 0$*, where* $\alpha$ *and* $\beta$ *are the type I and type II error probabilities of the test at* $\theta_0$ *and* $\theta_1$*, respectively. Let* $m_{\alpha,\beta}(\theta) = \min\{|\log\alpha|/I(\theta,\theta_0), |\log\beta|/I(\theta,\theta_1)\}$*. Let* $\varepsilon_{\alpha,\beta}$ *be positive numbers such that* $\varepsilon_{\alpha,\beta} \to 0$ *as* $\alpha+\beta \to 0$*, and let* $\nu$ *be the smallest* $j(\le k)$ *such that* $n_j \ge (1-\varepsilon_{\alpha,\beta})m_{\alpha,\beta}(\theta)$*, defining* $\nu$ *to be* $k$ *if no such* $j$ *exists. Then for fixed* $\theta, \theta_0$ *and* $\theta_1 > \theta_0$*, as* $\alpha+\beta \to 0$*,*

$$P_\theta(T \ge n_\nu) \to 1.$$

*If furthermore* $\nu < k$*,* $|m_{\alpha,\beta}(\theta) - n_\nu|/m_{\alpha,\beta}^{1/2}(\theta) \to 0$ *and*

$$\limsup \frac{m_{\alpha,\beta}(\theta)}{\max\{|\log\alpha|/I(\theta,\theta_0),\ |\log\beta|/I(\theta,\theta_1)\}} < 1,$$

*then* $P_\theta(T \ge n_{\nu+1}) \ge \frac{1}{2} + o(1)$*.*

The $n_j$ in Theorem 3.1 can in fact be random variables independent of $X_1, X_2, \ldots$. In this case the preceding argument can still be applied after conditioning on $(n_1, \ldots, n_k)$. The next step of Lai and Shih (2004, p. 510) is to extend Lorden's result on the asymptotic optimality of the 2-SPRT to the group sequential setting in the following.

**Theorem 3.2.** *Let* $\theta_0 < \theta^* < \theta_1$ *be such that* $I(\theta^*, \theta_0) = I(\theta^*, \theta_1)$*. Let* $\alpha+\beta \to 0$ *such that* $\log\alpha \sim \log\beta$.

*(i) The sample size* $n^*$ *of the Neyman–Pearson test of* $\theta_0$ *versus* $\theta_1$ *with error probabilities* $\alpha$ *and* $\beta$ *satisfies* $n^* \sim |\log\alpha|/I(\theta^*,\theta_0)$*.*

*(ii) For* $L \ge 1$*, let* $\mathscr{T}_{\alpha,\beta,L}$ *be the class of stopping times associated with group sequential tests with error probabilities not exceeding* $\alpha$ *and* $\beta$ *at* $\theta_0$ *and* $\theta_1$ *and with* $k$ *groups and prespecified group sizes such that* (3.17) *holds and* $n_k = n^* + L$*. Then, for given* $\theta$ *and* $L$*, there exists* $\tau \in \mathscr{T}_{\alpha,\beta,L}$ *that stops sampling when*

$$\begin{aligned} &(\theta-\theta_0)S_{n_i} - n_i\{\psi(\theta)-\psi(\theta_0)\} \ge b \\ \textit{or}\quad &(\theta-\theta_1)S_{n_i} - n_i\{\psi(\theta)-\psi(\theta_1)\} \ge \tilde{b} \end{aligned} \tag{3.18}$$

*for* $1 \le i \le k-1$*, with* $b \sim |\log\alpha| \sim \tilde{b}$*, and such that*

$$E_\theta(\tau) \sim \inf_{T \in \mathscr{T}_{\alpha,\beta,L}} E_\theta(T) \sim n_\nu + \rho(\theta)(n_{\nu+1} - n_\nu), \tag{3.19}$$

*where* $\nu$ *and* $m_{\alpha,\beta}(\theta)$ *are defined in Theorem 3.1 and* $0 \le \rho(\theta) \le 1$.

Whereas the group sequential 2-SPRT in Theorem 3.2 requires specification of $\theta$, the group sequential GLR in Sect. 4.2 replaces $\theta$ at the $i$th interim analysis by the maximum likelihood estimate $\hat{\theta}_{n_i}$, similar to (3.12) for the fully sequential case. In Sect. 4.2.2 we describe a group sequential GLR test introduced by Lai and Shih (2004, pp. 511–512) and show that it attains the asymptotic lower bound (3.19) at every fixed $\theta$ and that its power is comparable to the upper bound $1-\beta$ at $\theta_1$, under the assumption that the group sizes satisfy (3.17) with $n_k \sim |\log \alpha|/I(\theta^*, \theta_0)$, as $\alpha+\beta \to 0$ such that $\log \alpha \sim \log \beta$.

## 3.6 Optimality of SPRT and Dynamic Programming

In this section we prove the Wald–Wolfowitz theorem on the optimality of the SPRT and introduce the principle of dynamic programming that is used not only to prove it but also to solve much more general sequential optimization problems. Optimization provides an important tool in formulating and computing statistical procedures. Nonlinear least squares and optimal experimental designs covered in Chap. 2 and maximum likelihood estimators and Bayes rules are all optimization problems, and so are regularization methods for sparse high-dimensional regression such as Lasso and elastic net which amount to convex optimization (Boyd and Vandenberghe 2004). Whereas linear programming is concerned with optimal choice of the variables of a linear objective function subject to linear constraints, dynamic programming (DP) is concerned with optimal sequential choice of variables of the summands of an objective function that is the sum of cost functions of the variables over successive periods. In the case where these costs are random variables specified by some stochastic dynamic system (usually Markovian), DP amounts to stochastic optimization over time. In describing the stochastic costs, it is useful to distinguish the variables, called "controls," that can be chosen to minimize the total cost from the other variables, called "states," that undergo stochastic dynamics which depend on the controls chosen.

### *3.6.1 Dynamic Programming: Finite-Horizon Case*

Consider the problem of choosing the controls $u_t \in U$ sequentially so as to minimize $E\{\sum_{t=1}^N c_t(x_t, u_t) + c_N(x_N)\}$, with state variables $x_t$. Here $c_t$ are given functions, and $N$ is called the "horizon" of the problem. The control $u_t$ affects the dynamics of the future states $x_{t+1}, \ldots, x_N$ and depends on the information set $\mathscr{F}_t$ consisting of $x_t, u_{t-1}, x_{t-1}, \ldots, u_1, x_1$. The summand $c_t(x_t, u_t)$ represents the immediate cost and $c_N(x_N)$ the terminal cost; no control is exercised at time $N$ because it only affects the states after $N$. Noting that the choice of $u_t$ can only depend on $\mathscr{F}_t$, we use the "tower property" of conditional expectations to write

$$E\{c_t(x_t, u_t) + V_{t+1}\} = E\{E[c_t(x_t, u_t) + V_{t+1} | \mathscr{F}_t]\}, \tag{3.20}$$

where $V_{t+1}$ involves the information set $\mathscr{F}_{t+1}$ and will be defined below. The tower property suggests the following *backward induction* algorithm to choose $u_t$ for $t = N-1, N-2, \ldots, 1$, and to define $V_t$, initializing with $V_N = c_N(x_N)$:

$$u_t = \arg\min_{u \in U} \{c_t(x_t, u) + E(V_{t+1}|\mathscr{F}_t)\}, \quad V_t = c_t(x_t, u_t) + E(V_{t+1}|\mathscr{F}_t). \tag{3.21}$$

The functions $V_t$ are called *value functions*. We have assumed that the minimizer exists in (3.21) to define $u_{t+1}$. Although this is indeed the case in our applications, we actually do not need this assumption in the definition of the value functions, which we can define more generally by

$$V_t = \inf_{u \in U} \{c_t(x_t, u) + E(V_{t+1}|\mathscr{F}_t)\}. \tag{3.22}$$

The optimal control $u_t$ in (3.21) is a function of $x_t, u_{t-1}, x_{t-1}, \dots, u_1, x_1$. When $x_t$ is a controlled Markov chain so that the conditional probability of $x_{t+1}$ given $x_s, u_s$ $(s \leq t)$ depends only on $x_t, u_t$, the optimal control $u_t$ can be chosen to depend only on $x_t$.

### 3.6.2 *Infinite-Horizon Dynamic Programming*

Letting $N \to \infty$ in the finite-horizon case leads to the infinite-horizon problem of minimizing $E\{\sum_{t=1}^{\infty} c_t(x_t, u_t)\}$ when the infinite series is summable. The case $c_t(x, u) = \beta^t c(x, u)$ is called the discounted problem, with discount factor $0 < \beta < 1$. When $x_t$ is a controlled Markov chain with state space $S$ and stationary transition probabilities

$$P\{x_{t+1} \in A | x_t = x,\ u_t = u\} = P_{x,u}(A) \tag{3.23}$$

for all $t \geq 0$, $u \in U$, $x \in S$, and $A \subset S$, the value function

$$V(x) := \sup_{u_1, u_2, \dots} E\left\{\sum_{t=1}^{\infty} \beta^t c(x_t, u_t) | x_0 = x\right\}$$

is the solution of the dynamic programming equation

$$V(x) = \inf_{u \in U}\left\{c(x, u) + \beta \int_S V(y)\, dP_{x,u}(y)\right\}, \tag{3.24}$$

which can be derived by letting the horizon $N$ in the finite-horizon equation (3.22) approach $\infty$. The sequence $\boldsymbol{u} = (u_1, u_2, \dots)$ is called a control policy. From (3.24), it follows that the optimal control policy is a *stationary* policy in the sense that $u_t = g(x_t)$ for some time-invariant function $g$.

There are two commonly used methods to solve (3.24) for $V$ and for the stationary policy. One is *value iteration*. Starting with an initial guess $v_0$ of $V$, it uses successive approximations

$$v_{k+1}(x) = \min_{u \in U}\left\{c(x, u) + \beta \int_S v_k(y)\, dP_{x,u}(y)\right\}.$$

The minimizer $u = u_k(x)$ of the right-hand side yields an approximation to the stationary policy at the $k$th iteration. The other is *policy iteration*, which can be applied when the state space $S$ is finite as follows. Denote the states of $S$ by $1,\dots,m$, and let

$$\boldsymbol{c}_g = (c(1,g(1)),\dots,c(m,g(m)))^T, \quad \boldsymbol{P}_g = \left(P_{x,g(x)}(y)\right)_{1\le x\le m,\,1\le y\le m}.$$

Note that $\boldsymbol{P}_g$ is the transition matrix of the controlled Markov chain with stationary control policy $g$ that is specified by the vector $(g(1),\dots,g(m))^T$. Minimization over $g$, therefore, is the same as minimization of $\mathbb{R}^m$. In view of (3.24), we define the affine transformation $T_g(\boldsymbol{x}) = \boldsymbol{c}_g + \beta \boldsymbol{P}_g(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathbb{R}^m$. Letting $T(\boldsymbol{x}) = \min_g T_g(\boldsymbol{x})$ and $\boldsymbol{v} = (V(1),\dots,V(m))^T$, note that (3.24) can be written as $\boldsymbol{v} = T(\boldsymbol{v})$. This suggests the following policy iteration scheme that initializes with a preliminary guess $g_0$ of the optimal $g^*$. At the $k$th iteration, after determining $g_k$, solve the linear system $\boldsymbol{v} = T_{g_k}(\boldsymbol{v})$ and denote the solution by $\boldsymbol{v}_k$. If $T(\boldsymbol{v}_k) = \boldsymbol{v}_k$, stop and set $g^* = g_k$. Otherwise solve the linear system $T_g(\boldsymbol{v}_k) = T(\boldsymbol{v}_k)$ for $(g(1),\dots,g(m))^T$ and set $g_{k+1} = g$.

In the case where the control $u_t$ consists of whether to stop at time $t$, the stochastic optimization problem is called an *optimal stopping* problem. Note that unlike more general stochastic control problems, the stopping rule does not change the dynamics of $x_t$. The control policy reduces to a stopping rule $\tau$ in this case. Suppose the costs are $c(x_t)$ prior to stopping and $h(x_\tau)$ upon stopping, and $c_t(x_t) = 0$ for $t > \tau$. The transition matrix (3.23) now has the form

$$P\{x_{t+1} \in A | x_t = x\} = P_x(A) \qquad \text{for } 0 \le t < \tau,\ u \in U,\ x \in S, \text{ and } A \subset S.$$

The value function $V(x) := \sup_\tau E\{\sum_{t<\tau} c(x_t) + h(x_\tau) | x_0 = x\}$ satisfies the dynamic programming equation

$$V(x) = \min\left\{h(x), c(x) + \int_S V(y)\, dP_x(y)\right\}. \tag{3.25}$$

The notation $\sup_\tau$ above refers to supremum over all stopping rules $\tau$, and the right-hand side of (3.25) refers to either stopping when state $x$ is observed and incurring cost $h(x)$ or continuing after paying cost $c(x)$ and then proceeding optimally thereafter. The minimum in (3.25) refers to choosing continuation or stopping according to which has a smaller expected cost. The optimal stopping rule is a stationary policy of the form $\tau = \inf\{n \ge 0 : h(x_t) \le V(x_t)\}$.

### *3.6.3 Bayes Sequential Tests of Simple Null Versus Simple Alternative*

In this section we apply optimal stopping theory to sequential hypothesis testing in which one has to decide whether to accept the null hypothesis upon stopping.

Consider the problem of testing the simple null hypothesis $H_0 : f = f_0$ versus $H_1 : f = f_1$ based on i.i.d. observations $X_1, X_2, \ldots$ as in Sect. 3.2. There are two probability measures in this problem, one involving $f_0$ and the other $f_1$. Dynamic programming, however, involves a single measure and expectation with respect to that measure. A Bayesian formulation that puts prior probabilities of $H_0$ and $H_1$ would yield a single measure. Let $0 < \pi < 1$ be the prior probability in favor of $H_0$ so that $1 - \pi$ is that in favor of $H_1$. Let $0 < w < 1$ be the loss of wrongly rejecting $H_0$, and let $1 - w$ be that of wrongly accepting it (when $H_1$ is true). In addition, there is a cost $c$ for each observation taken so that the Bayes risk of a sequential test with stopping rule $\tau$ and terminal decision rule $\delta$ (taking the value 0 or 1 according to whether $H_0$ is accepted or not) is

$$r(\tau, \delta) = E\left(wI_{\{\delta=1,\ H_0 \text{ is true}\}} + (1-w)I_{\{\delta=1,\ H_1 \text{ is true}\}} + c\tau\right), \tag{3.26}$$

where $E$ denotes expectation with respect to the measure under which the 0-1 variable $\theta$ has probability $\pi$ of being equal to 0 and $X_1, X_2, \ldots$ are i.i.d. with common density $f_\theta$ given $\theta$. As noted in Sect. 1.5, the optimal Bayes decision rule $\delta^*$ does not depend on the stopping rule and accepts $H_0$ or $H_1$ according to which has the smaller posterior risk. Thus, for any stopping rule $\tau$, $\delta^*$ accepts $H_1$ if $w\pi_\tau \le (1-w)(1-\pi_\tau)$ and accepts $H_0$ otherwise, where

$$\pi_n = \pi f_0(X_1)\ldots f_0(X_n) / \{\pi f_0(X_1)\ldots f_0(X_n) + (1-\pi) f_1(X_1)\ldots f_1(X_n)\} \tag{3.27}$$

is the posterior probability in favor of $H_0$. Putting $\delta = \delta^*$ in (3.26) and applying the tower property of conditional expectations, we obtain

$$r(\tau, \delta^*) = E\{h(\pi_\tau) + c\tau\}, \qquad \text{where } h(p) = \min\{wp, (1-w)(1-p)\}. \tag{3.28}$$

From (3.27), it follows that $\pi_n \propto \pi f_0(X_1)\ldots f_0(X_n)$ and therefore

$$\pi_n = \pi_{n-1} f_0(X_n) / \{\pi_{n-1} f_0(X_n) + (1-\pi_{n-1}) f_0(X_n)\}.$$

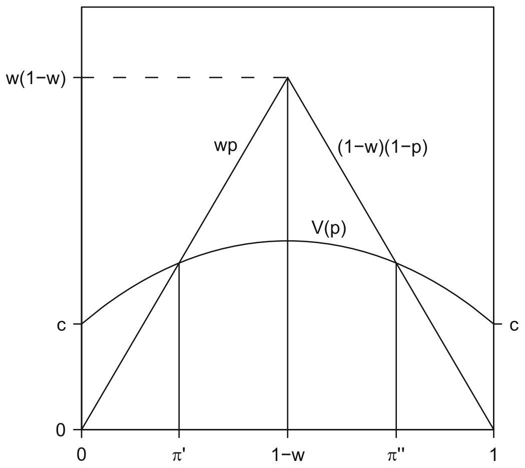
Note that conditional on $\pi_{n-1}$, $\theta = 1$ (or 0) with probability $1 - \pi_n$ (or $\pi_n$) and $X_n$ has density $f_\theta$. Hence $\pi_n$ is a stationary Markov chain on state space $S = (0,1)$. Therefore, as shown in the preceding section, the optimal stopping rule is a stationary policy of the form

$$\tau = \inf\{n \ge 1 : h(\pi_n) \le V(\pi_n)\}. \tag{3.29}$$

For $0 \le p \le 1$, $h(p)$ is the minimum of the linear functions $wp$ and $(1-w)(1-p)$. Moreover, $V(p)$ is the infimum over $\tau$ of

$$\begin{aligned} r(\tau, \delta^*; p) = p\{&wP_0(\delta^* \text{ rejects } H_0) + cE_0(T)\} \\ &+ (1-p)\{(1-w)P_1(\delta^* \text{ accepts } H_0) + cE_1(T)\}, \end{aligned}$$

**Fig. 3.1** The functions $V(p)$ and $h(p)$



which is a linear function of $p$ for given $\tau$. Therefore, $V(p)$ is a concave function of $p$ with $V(0) = c = V(1)$, and there exist $0 < \pi' < \pi'' < 1$ such that $h(p) \le V(p)$ if and only if $p \le \pi'$ or $p \ge \pi''$; see Fig. 3.1. Therefore, the stopping rule (3.29) can be written as

$$\begin{aligned}\tau &= \inf\{n \ge 1 : \pi_n \le \pi' \text{ or } \pi_n \ge \pi''\}\\ &= \inf\left\{n \ge 1 : L_n \le \frac{\pi}{1-\pi}\frac{1-\pi''}{\pi''} \text{ or } L_n \ge \frac{\pi}{1-\pi}\frac{1-\pi'}{\pi'}\right\},\end{aligned}$$

since $\pi_n = 1/\{1 + \frac{1-\pi}{\pi} L_n\}$ by (3.27) and (3.1). Hence, the optimal Bayes solution is the SPRT with stopping boundaries

$$A = \frac{\pi}{1-\pi}\frac{1-\pi''}{\pi''}, \quad B = \frac{\pi}{1-\pi}\frac{1-\pi'}{\pi'}, \tag{3.30}$$

in which $\pi'$ and $\pi''$ depend on $w$ and $c$.

Using $\pi'(w,c)$, $\pi''(w,c)$, and $V(p;w,c)$ to represent $\pi'$, $\pi''$, and $V(p)$ as functions of $w$ and $c$, it can be shown that for fixed $w$, $\pi'$ is nondecreasing and continuous in $c$. Moreover, as the sampling cost $c$ per observation approaches 0, the error probabilities can be made arbitrarily small by using the Neyman–Pearson test with sample size $n$ that is large enough. Therefore, $V(p;w,c) \to 0$ as $c \to 0$, for fixed $p$ and $w$. Since $h(p)$ does not depend on $c$, it then follows that for fixed $w$, $\pi'(w,c) \to 0$ as $c \to 0$. Similarly, for fixed $w$, $\pi''(w,c)$ is nonincreasing and continuous in $c$ and $\pi''(w,c) \to 1$ as $c \to 0$. We next use these properties of $\pi'$ and $\pi''$ to prove the Wald–Wolfowitz theorem on the optimality of SPRT.

### 3.6.4 Auxiliary Bayes Problem and Optimality of the SPRT

As pointed out in Sect. 3.3, based on the fact that, assuming Wald's approximations which ignore overshoots to be exact, the SPRT attains the lower bound (3.8) for the expected sample size of sequential tests of a simple null versus a simple alternative subject to type I and type II error probability constraints, Wald (1945) conjectured the following optimality theorem of the SPRT.

**Theorem 3.3.** *For the problem of testing $H_0 : f = f_0$ versus $H_1 : f = f_1$ based on i.i.d. observations $X_1, X_2, \ldots$ with common density function $f$, the SPRT with type I error probability $\alpha$ and type II error probability $\beta$ minimizes $E_0(T)$ and $E_1(T)$ among all tests (sequential or fixed sample size) with stopping time $T$, which has finite expectation under $H_0$ and $H_1$, and error probabilities $P_0(\text{Reject } H_0) \le \alpha$ and $P_1(\text{Accept } H_0) \le \beta$.*

The theorem, which considers the frequentist optimality of the SPRT, was proved by Wald and Wolfowitz (1948) by showing that the SPRT is the Bayes rule in an auxiliary Bayes problem. Given the stopping boundaries $A$ and $B$ of the SPRT, the auxiliary Bayes problem is concerned with choosing $\pi$, $c$, and $w$ so that $\pi$, $\pi'(w,c)$ and $\pi''(w,c)$, satisfy (3.30). This is the content of the following.

**Lemma 3.1.** *Given $\varepsilon > 0$ and $0 < A \le 1 \le B$, there exists $(\pi, c, w)$ with $0 < \pi < \varepsilon$ such that (3.30) holds for $\pi$, $\pi'(w,c)$, and $\pi''(w,c)$. Moreover, there also exists another triple $(\pi, c, w)$ with $1 - \varepsilon < \pi < 1$ satisfying (3.30).*

*Proof.* Fix $w$ so that $1 - w < A\varepsilon$ and note that

$$\pi'(w,c)\left(1 - \pi''(w,c)\right) / \left\{\pi''(w,c)\left(1 - \pi'(w,c)\right)\right\} \tag{3.31}$$

is a continuous function of $c$ and approaches 0 as $c \to 0$; see the last paragraph of Sect. 3.6.3. Moreover, (3.31) is equal to 1 for sufficiently large $c$ for which $\pi' = \pi''$ (see Fig. 3.1). Hence, there exists $c$ such that (3.31) is equal to $A/B$. With this choice of $c$, let $\pi = \pi'(w,c)/\{A + (1 - A)\pi'(w,c)\}$. Then

$$\frac{1-\pi}{\pi} = A\,\frac{1-\pi'(w,c)}{\pi'(w,c)} = B\,\frac{1-\pi''(w,c)}{\pi''(w,c)}, \tag{3.32}$$

in which the second equality follows from that (3.31) is equal to $A/B$ by the choice of $c$. Note that (3.32) is the same as (3.30). Since $\pi' \le 1 - w \le \pi''$ (see Fig. 3.1), it follows that

$$\pi = \frac{\pi'(w,c)}{A + (1-A)\pi'(w,c)} \le \frac{\pi'(w,c)}{A} \le \frac{1-w}{A} < \varepsilon,$$

recalling that $w$ has been chosen such that $1 - w < A\varepsilon$. An obvious modification of this idea can also yield another triplet $(w, c, \pi)$ with $\pi > 1 - \varepsilon$.

*Proof (Proof of Theorem 3.3).* By Lemma 3.1, there exists $(\pi, c, w)$ such that the SPRT with stopping rule (1.1) is Bayes with prior probability $\pi$ in favor of $H_0$. Since it minimizes the Bayes risk,

$$\begin{aligned}&\pi\{w\alpha + cE_0(N)\} + (1-\pi)\{(1-w)\beta + cE_1(N)\}\\ &\quad \le \pi\left\{w\tilde{\alpha} + cE_0(T)\right\} + (1-\pi)\left\{(1-w)\tilde{\beta} + cE_1(T)\right\}\end{aligned}$$

for any test $(T,\delta)$ that has type I error $\tilde{\alpha} \le \alpha$ and type II error $\tilde{\beta} \le \beta$. This implies

$$\pi\left\{E_0(T) - E_0(N)\right\} + (1-\pi)\left\{E_1(T) - E_1(N)\right\} \ge 0.$$

Since $\pi$ can be chosen less than any $\varepsilon > 0$ by Lemma 3.1, it follows that $E_1(T) \ge E_1(N)$. Similarly, since $\pi$ can also be chosen arbitrarily close to 1, $E_0(T) \ge E_0(N)$.

## 3.7 Approximations to Bayes Sequential Tests and Operating Characteristics

### 3.7.1 *Derivation of Lorden's 2-SPRT as an Approximate Bayes Rule*

The Kiefer–Weiss problem in Sect. 3.4 involves three probability measures, with density functions $f_{\theta_0}$, $f_{\theta_1}$, and $f_{\theta^*}$. To solve the problem by dynamic programming, Lorden (1976) uses first a Bayesian argument to combine the three measures into a single measure after putting prior probabilities on $\theta_0$, $\theta_1$, and $\theta^*$ and then a change of measures to $P_{\theta^*}$, which will be denoted simply by $P_*$ (as $P_i$ has been used to denote $P_{\theta_i}$). This enables him to express the Bayes risk in the form

$$\inf_{\tau} E_*\left\{\tau + \min(uL_\tau^0, u'L_\tau^1)\right\}, \quad \text{where } L_n^j = \prod_{i=1}^{n}\left[f_{\theta_j}(X_i)/f_{\theta^*}(X_i)\right], \tag{3.33}$$

$j = 0, 1$. In (3.33), it is assumed that $c$ (the cost per observation) times the prior probability of $\theta^*$ is equal to 1 as the Kiefer–Weiss problem is concerned with the expected sample size only at $\theta^*$, that the loss for wrongly rejecting $H_0$ times the prior probability of $H_0$ is $u$, and that the loss for wrongly rejecting $H_1$ times the prior probability of $H_1$ is $u'$. Note that $L_n^j$ is the likelihood ratio associated with the change of measures from $P_j$ to $P_*$ so that the expectation can be taken under $P_*$, in view of Wald's likelihood ratio identity (3.3). The Bayes terminal decision rule rejects $H_0$ or $H_1$ according to which gives a smaller posterior risk, and leads to the term $\min(uL_\tau^0, u'L_\tau^1)$ in (3.33).

The dynamic programming problem associated with (3.33) involves a two-dimensional Markov chain $\boldsymbol{x}_n = (uL_n^0, u'L_n^1)$ with stationary transition probabilities, For $\boldsymbol{x} = (x_1, x_2)$, define $h(\boldsymbol{x}) = \min(x_1, x_2)$. As explained in Sect. 3.6.2, the optimal stopping rule has the form

$$\tau = \inf\{n \geq 1 : h(\boldsymbol{x}_n) \leq V(\boldsymbol{x}_n)\}, \tag{3.34}$$

in which $V$ is the value function. However, unlike (3.29) that can be reduced to the stopping time of the SPRT with constant stopping boundaries $A$ and $B$ for the likelihood ratio statistics $L_n = \prod_{i=1}^n (f_1(X_i)/f_0(X_i))$, the stopping boundaries for (3.34) which involve the likelihood ratio pairs $(L_n^0, L_n^1)$ are nonlinear curves. When $f_\theta$ is the $N(\theta, 1)$ density, $\log L_n^\theta$ is a normal random walk, and Lai (1973) studied the optimal stopping problem and found that the nonlinear boundaries can be well approximated by Anderson's test with triangular boundaries for normal random walks.

Lorden noticed that Anderson's test is simply a 2-SPRT, which suggested the possibility of approximating (3.34) by a 2-SPRT. This led to his approach that uses the stopping rule $T^*$ defined by (3.11) in lieu of the stopping rule (3.34) and evaluates asymptotically the Bayes risk $E_*\{T^* + \min(uL_{T^*}^0, u'L_{T^*}^1)\}$ as $\min(A_0^{-1}, A_1^{-1}) \to 0$, with $u$ and $u'$ determined by the boundaries $A_0$ and $A_1$ of (3.11). This asymptotic analysis yields his theorem that the 2-SPRT solves the Kiefer–Weiss problem up to $O(1)$ error.

**Theorem 3.4.** *Let $\alpha$ and $\beta$ denote the error probabilities of the 2-SPRT (3.11). Let $n(A_0, A_1)$ denote the infimum of $E_*(T)$ over all tests satisfying $P_0$(Reject $H_0$) $\leq \alpha$ and $P_1$(Reject $H_1$) $\leq \beta$. If*

$$E_*\left\{\left[\log(f_{\theta^*}(X_1)/f_{\theta_0}(X_1))\right]^2 + \left[\log(f_{\theta^*}(X_1)/f_{\theta_1}(X_1))\right]^2\right\} < \infty,$$

*then $E_*T^* = n(A_0, A_1) + o(1)$ as $A_0 \to \infty$ and $A_1 \to \infty$.*

### *3.7.2 Approximation of Bayes Sequential Tests of One-Sided Hypotheses in an Exponential Family by Sequential GLR Tests*

In Sect. 3.4 we have referred to the Bayes problem of testing one-sided composite hypotheses of the form $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta \geq \theta_1$ for the parameter of an exponential family $f_\theta(x) = \exp(\theta x - \psi(\theta))$ with respect to some measure $\nu$ on the real line. Given a prior distribution $G$ on $\theta$, a loss $l(\theta)$ of accepting the incorrect hypothesis, and a cost $c$ per observation, the Bayes risk of a sequential test $(T, \delta)$ with stopping rule $T$ and terminal decision rule $\delta$ is

$$r(T,\delta) = c\int E_\theta(T)\,dG + \int_{\theta\le\theta_0} l(\theta)P_\theta\{\delta \text{ accepts } H_1\}\,dG$$
$$+\int_{\theta>\theta_1} l(\theta)P_\theta\{\delta \text{ accepts } H_0\}\,dG. \tag{3.35}$$

The support of $G$ is assumed to be contained in the natural parameter space $\Theta = \{\theta : \int e^{\theta x}d\nu(x) < \infty\}$ and to satisfy $\int l(\theta)dG(\theta) < \infty$. The optimal $\delta^*$ does not depend on the stopping rule and accepts the $H_i$ that has smaller posterior risk. However, it is difficult to solve this infinite-horizon DP problem for the optimal stopping rule in this general case because the optimal policy is no longer stationary.

A well-known asymptotic solution to the Bayes problem of minimizing $r(T,\delta)$ is due to Schwarz (1962). Let $\mathscr{B}(c)$ denote the continuation region of the Bayes rule (which continues sampling at stage $n+1$ if and only if $(n,S_n)\in\mathscr{B}(c)$). Assuming that $l(\theta) > 0$ for $\theta \notin (\theta_0,\theta_1)$ and that $G(I) > 0$ for every open interval $I \subset \Theta$, Schwarz's asymptotic theory leads to the following limiting continuation region of the Bayes rule: As $c \to 0$,

$$\frac{\mathscr{B}(c)}{|\log c|} \longrightarrow \left\{(t,w) : 1+\min_{i=0,1}(\theta_i w - t\psi(\theta_i)) > \sup_\theta(\theta w - t\psi(\theta))\right\}. \tag{3.36}$$

Thus, writing $n = t|\log c|$ and $S_n = w|\log c|$, an asymptotic approximation to the Bayes rule is to continue sampling at stage $n+1$ if only if

$$\log c^{-1} + \min_{i=0,1}\{\theta_i S_n - n\psi(\theta_i)\} > \sup_n\{\theta S_n - n\psi(\theta)\}, \tag{3.37}$$

or, equivalently, to stop sampling at stage

$$N_c = \inf\left\{n\ge 1 : \max\left[\prod_{i=1}^n \frac{f_{\hat\theta_n}(X_i)}{f_{\theta_0}(X_i)}, \prod_{i=1}^n \frac{f_{\hat\theta_n}(X_i)}{f_{\theta_1}(X_i)}\right] \ge c^{-1}\right\}, \tag{3.38}$$

where $\hat\theta_n$ is the maximum likelihood estimate of $\theta$. The terminal decision rule $\delta^*$ is to accept $H_1$ (or $H_0$) if $\prod_{i=1}^{N_c} f_{\theta_1}(X_i) >$ (or $\le$) $\prod_{i=1}^{N_c} f_{\theta_0}(X_i)$.

There are two main steps in Schwarz's derivation of the above asymptotic approximation to the Bayes rule. The first step involves upper and lower bounds for the Bayes continuation region $\mathscr{B}(c)$. Let

$$L(n,x) = \frac{\min_{i=0,1}\int_{\Theta_i} l(\theta)\exp\{\theta x - n\psi(\theta)\}\,dG(\theta)}{\int_\Theta \exp\{\theta x - n\psi(\theta)\}\,dG(\theta)} \tag{3.39}$$

be the *stopping risk*, which is the posterior loss due to the wrong decision of the Bayes test if stopping occurs at stage $n$ and $S_n = x$. Let $\mathscr{R}(c) = \{(n,x) : L(n,x) \ge c\}$. Schwarz showed that for sufficiently small $c > 0$,

$$\mathscr{R}(c) \supset \mathscr{B}(c) \supset \mathscr{R}(3\triangle^{-1}c|\log c|), \tag{3.40}$$

where $\triangle = \psi(\theta_0) + \psi(\theta_1) - 2\psi((\theta_0 + \theta_1)/2)$. The next step is to apply Laplace's asymptotic formula (see Sects. 2.1.2 and 3.7.3) to evaluate the integrals in (3.39), leading to the asymptotic approximation

$$\log L(n, S_n) \sim \min_{i=0,1} \{\theta_i S_n - n\psi(\theta_i)\} - \{\hat{\theta}_n S_n - n\psi(\hat{\theta}_n)\}. \tag{3.41}$$

Combining (3.41) and (3.40) gives the "asymptotic shape" (3.37) for the Bayes continuation region $\mathscr{B}(c)$.

A different asymptotic theory in Bayes sequential tests was developed by Chernoff (1961, 1965) in the context of testing $H_0 : \theta < 0$ versus $H_1 : \theta > 0$ for the mean $\theta$ of a normal distribution with unit variance. Instead of assuming an indifference zone $(\theta_0, \theta_1)$ and a general loss function $l(\theta) > 0$ for $\theta \notin (\theta_0, \theta_1)$ as in Schwarz's theory, Chernoff's theory considers the special loss function $l(\theta) = |\theta|$ for $\theta \neq 0$ and assumes a normal prior distribution $G$ with mean 0 and variance $\sigma^2$. The Bayes terminal decision rule accepts $H_0$ (or $H_1$) according as $S_n \leq 0$ (or $S_n > 0$) when stopping occurs at stage $n$. Thus, the Bayes problem reduces to the optimal stopping problem of finding a stopping rule to minimize

$$r(T) = c\int_{-\infty}^{\infty} E_\theta(T)\,dG + \int_{-\infty}^{0} |\theta| P_\theta\{S_T > 0\}\,dG + \int_{0}^{\infty} \theta P_\theta\{S_T \leq 0\}\,dG. \tag{3.42}$$

To study the optimal stopping rule, Chernoff (1961) introduced the normalization

$$t = c^{2/3}(n + \sigma^{-2}), \quad w = c^{1/3} S_n, \tag{3.43}$$

which is different from Schwarz's normalization $t = n/|\log c|$ and $w = S_n/|\log c|$. With the normalization (3.43) for the problem, Chernoff obtained a limiting continuation region of the form $\{(t, w) : |w| < f(t)\}$ as $c \to 0$. The stopping boundary $f(t)$ arises as the solution of the corresponding continuous-time stopping problem involving the Wiener process.

It seems somewhat artificial that there should be two different kinds of asymptotic approximations to Bayes tests of one-sided hypotheses, depending on whether there is an indifference zone. A more natural asymptotic theory should have the property that the approximation in the absence of an indifference zone can be obtained as the limit of the approximations with shrinking indifference zones. To see the feasibility of this unified approach, Lai (1988) studied the problem of testing sequentially $H_0 : \mu \leq -\gamma$ versus $H_1 : \mu > \gamma$ (with an indifference zone $(-\gamma, \gamma)$) and $H_0' : \mu < 0$ versus $H_1' : \mu > 0$ (without an indifference zone) for the drift coefficient $\mu$ of a Wiener process $\{w(t),\ t \geq 0\}$, assuming the $0-1$ loss, a flat prior on $\mu$, and a cost of $t$ for observing the process for a period of length $t$. Note that the Bayes terminal decision rule for either problem accepts the null, or alternative, hypothesis according as $w(t) < 0$ or $w(t) > 0$ when stopping occurs at time $t$. For the problem of testing $H_0' : \mu < 0$ versus $H_1' : \mu > 0$, the posterior loss $L_0(t, w)$ of stopping at time $t$ if $w(t) = w$ is observed and $H_0'$ or $H_1'$ is accepted according as $w < 0$ or $w > 0$ is given by

$$L_0(t,w) = t + \Phi(-|w|t^{-1/2}), \tag{3.44}$$

where $\Phi$ is the standard normal distribution function. For the problem of testing $H_0 : \mu \le -\gamma$ versus $H_1 : \mu \ge \gamma$, the corresponding posterior loss is

$$L_\gamma(t,w) = t + \Phi(-|w|t^{-1/2} - \gamma t^{1/2}). \tag{3.45}$$

Thus, (3.44) can be regarded as a special case of (3.45) with $\gamma = 0$.

Lai (1988) derived asymptotic expansions, as $t \to 0$ and $t \to \infty$, for the optimal stopping boundaries of the continuous-time optimal stopping problems for the Wiener process with loss functions (3.44) and (3.45), respectively. He also computed the optimal stopping boundaries numerically and thereby derived closed-form approximations to the optimal stopping boundaries $\pm h_\gamma(t)$ for $\gamma = 0$, $0 < \gamma \le 20$ and $\gamma > 20$ and over different ranges of $t$. In particular, he showed that

$$h_\gamma(t) = \left\{2t\left[\log t^{-1} + \tfrac{1}{2}\log\log t^{-1} - \tfrac{1}{2}\log 4\pi + o(1)\right]\right\}^{1/2} \tag{3.46}$$

as $t \to 0$, for fixed $\gamma \ge 0$. He then considered the problem of testing (a) $H_0' : \theta < 0$ versus $H_1' : \theta > 0$ and (b) $H_0 : \theta \le -\triangle$ versus $H_1 : \theta \ge \triangle$ for the mean $\theta$ of i.i.d. normal random variables $X_1, X_2, \ldots$. Define

$$t = cn, \quad w(t) = c^{1/2}S_n, \quad \mu = c^{-1/2}\theta, \quad \gamma = c^{-1/2}\triangle.$$

Since $c^{1/2}\theta n = \mu t$, $w(t)$ is a Wiener process with drift coefficient $\mu$ and with $t$ restricted to the set $I_c = \{c, 2c, \ldots\}$. As $c \to 0$, $I_c$ becomes dense in $[0,\infty)$. Moreover, for any prior distribution $G$ on $\theta$ such that $G$ has a positive continuous density $G'$, the density function $\pi_c$ of $\mu = c^{-1/2}\theta$ is

$$\pi_c(x) = c^{1/2}G'(c^{1/2}x) \sim c^{1/2}G'(0) \qquad \text{as } c \to 0,$$

and thereby the family of probability measures with densities $\pi_c$ converges to Lebesgue measure (flat prior). This suggests using the flat-prior continuous-time Bayes stopping boundaries $h_0$ and $h_\gamma$ as approximations to the Bayes tests of $H_0'$ versus $H_1'$ and of $H_0$ versus $H_1$, respectively.

Lai (1988) also extended these approximations from the normal distribution to the exponential family $f_\theta(x) = \exp(\theta x - \psi(\theta))$. Of particular importance in the extension is the Kullback–Leibler information number $I(\theta,\lambda)$ in (3.16). Note that in the case of a normal distribution with mean $\theta$ and variance 1, $I(\theta,\lambda) = (\theta-\lambda)^2/2$, and $\hat\theta_n = \bar X_n$ (when $A = (-\infty,\infty)$). Moreover, letting $g_\gamma(t) = (h_\gamma(t) + \gamma t)^2/(2t)$ for $\gamma \ge 0$, it follows that

$$|w(t)| \ge h_\gamma(t) \Longleftrightarrow \frac{(|w(t)| + \gamma t)^2}{2t} \ge g_\gamma(t)$$

$$\Longleftrightarrow \frac{(|S_n| + \triangle n)^2}{2n} \geq g_\gamma(cn)$$
$$\Longleftrightarrow \max\left\{I\left(\hat{\theta}_n, \theta_0\right), I\left(\hat{\theta}_n, \theta_1\right)\right\} \geq n^{-1} g_\gamma(cn),$$

where $\theta_1 = \triangle = -\theta_0$ in the case $H_0 : \theta \leq -\triangle$ versus $H_1 : \theta \geq \triangle$ and $\theta_1 = 0 = -\theta_0$ in the case $H_0' : \theta < 0$ versus $H_1' : \theta > 0$. This explains the stopping rule (3.13) in which $g = g_\gamma$, with $\gamma \geq 0$. Note that in the exponential family $f_\theta(x) = e^{\theta x - \psi(\theta)}$, $\psi'(\hat{\theta}_n) = \bar{X}_n$ and the logarithm of the GLR statistic can therefore be expressed in terms of the Kullback–Leibler information number:

$$\sum_{i=1}^{n} \log\left(f_{\hat{\theta}_n}(X_i)/f_{\theta_0}(X_i)\right) = nI\left(\hat{\theta}_n, \theta_0\right).$$

### 3.7.3 Laplace's Asymptotic Formula and Approximations to Operating Characteristics

In the preceding section we consider the optimal stopping problem associated with a Bayes test of one-sided composite hypotheses and have developed approximations to solve that problem. As pointed out in the derivation of (3.41), a key idea is the approximation of the integrals in (3.39) using Laplace's method. We begin with a review of Laplace's method for asymptotic evaluation of the integral $\int_{-\infty}^{\infty} u(\theta) e^{av(\theta)} d\theta$ as $a \to \infty$, where $u$ and $v$ are continuous functions on $\mathbb{R}$ such that $v$ has unique maximum at $\theta^*$ and is twice continuously differentiable in some neighborhood of $\theta^*$, $\limsup_{|\theta| \to \infty} v(\theta) < \min\{v(\theta^*), 0\}$ and $\limsup_{|\theta| \to \infty} |u(\theta)| e^{Av(\theta)} < \infty$ for some $A > 0$. Since $v'(\theta^*) = 0$, $v''(\theta^*) < 0$ and

$$e^{av(\theta)} = e^{av(\theta^*)} \exp\left\{a\left[v''(\theta^*) + o(1)\right](\theta - \theta^*)^2/2\right\} \qquad \text{as } \theta \to \theta^*, \tag{3.47}$$

and since the assumptions on $u$ and $v$ imply that for every $\varepsilon > 0$, there exists $\eta_\varepsilon > 0$ such that as $a \to \infty$,

$$\left(\int_{-\infty}^{\theta^* - \varepsilon} + \int_{\theta^* + \varepsilon}^{\infty}\right) u(\theta) e^{av(\theta)} d\theta = O\left(\exp\left(a\left[v(\theta^*) - \eta_\varepsilon\right]\right)\right),$$

it follows that

$$\int_{-\infty}^{\infty} u(\theta) e^{av(\theta)} d\theta \sim u(\theta^*) e^{av(\theta^*)} \left(-av''(\theta^*)\right)^{-1/2} \int_{-\infty}^{\infty} e^{-t^2/2} dt$$
$$= \sqrt{\frac{2\pi}{a|v''(\theta^*)|}} u(\theta^*) e^{av(\theta^*)} \tag{3.48}$$

as $a \to \infty$, using the change of variables $t = (-av''(\theta^*))^{1/2}(\theta - \theta^*)$. Laplace's asymptotic formula (3.48) relates the integral $\int_{-\infty}^{\infty} u(\theta)e^{av(\theta)}d\theta$ to the maximum of $e^{av(\theta)}$ over $\theta$, which explains why the posterior loss in (3.39) can be approximated by GLR statistics. Laplace's formula also holds in $\mathbb{R}^d$, as in (2.9), and for more general regions $\Theta$ for which it involves a tubular neighborhood of the maximizing set; see Chan and Lai (2000).

Note that Laplace's asymptotic formula involves $|v''(\theta^*)|$, which it assumes implicitly to be bounded away from 0 and $\infty$. The likelihood function $\exp\{n(\bar{X}_n - \psi(\theta))\}$ is maximized at $\hat{\theta}_n$ with $\psi'(\hat{\theta}_n) = \bar{X}_n$. Since Laplace's asymptotic formula involves $\psi''$, Lai (1988) assumes that $\theta$ is known to belong to an open interval $A \subset \Theta$ such that $\psi''$ is bounded away from 0 and $\infty$ and is uniformly continuous on $A$ and that $\theta_0$ and $\theta_1$ belong to $A$ and $G$ is a probability distribution on $A$. The maximum likelihood estimator in this case is $\tilde{\theta}_n := (\hat{\theta}_n \vee a_1) \wedge a_2$, where $-\infty \le a_1 < a_2 \le \infty$ are the boundaries of $A$. Applying Laplace's method to the likelihood ratio identity (3.3) with suitably chosen $P$ and $Q$ and making use of Hoeffding's lower bound (3.10), Lai (1988) proved the following asymptotic approximations to the operating characteristics of the sequential GLR test with stopping rule (3.13), in which $\hat{\theta}_n$ is replaced by $\tilde{\theta}_n$, and with terminal decision rule that rejects $H_0$ if $\tilde{\theta}_{N(g,c)} > \theta^*$, where $\theta^* \in (\theta_0, \theta_1)$ is such that $I(\theta^*, \theta_0) = I(\theta^*, \theta_1)$. Note that if $\tilde{\theta}_{N(g,c)} > \theta^*$, then $I(\tilde{\theta}_{N(g,c)}, \theta_0) > I(\tilde{\theta}_{N(g,c)}, \theta_1)$ and $\tilde{\theta}_{N(g,c)} > \theta_0$.

**Lemma 3.2.** *Let $g$ be a nonnegative function on $(0,\infty)$ such that*

$$g(t) \sim \log t^{-1} \quad \textit{and} \quad g(t) \ge \log t^{-1} + \xi \log\log t^{-1} + O(1) \qquad \textit{as } t \to \infty, \tag{3.49}$$

*for some real number $\xi$. Let $\alpha = P_{\theta_0}\{\tilde{\theta}_{N(g,c)} > \theta^*\}$ and $\beta = P_{\theta_1}\{\tilde{\theta}_{N(g,c)} \le \theta^*\}$, and let $\mathscr{T}(\alpha,\beta)$ denote the class of sequential tests of $H_0 : \theta \le \theta_0$ versus $H_1 : \theta \ge \theta_1$ such that $P_\theta$(Reject $H_0$) $\le \alpha$ for $\theta \le \theta_0$ and $P_\theta$(Reject $H_1$) $\le \beta$ for $\theta \ge \theta_1$.*

*(i) For fixed $\theta_0$ and $\theta_1$, as $c \to 0$, $\log\alpha \sim \log\beta \sim \log c$, and for every bounded subset $B$ of $A$,*

$$E_\theta N(g,c) \sim \frac{|\log c|}{J(\theta)} \sim \inf_{(T,\delta)\in\mathscr{T}(\alpha,\beta)} E_\theta T \textit{ uniformly in } \theta \in B, \tag{3.50}$$

*where $J(\theta) = max\{I(\theta,\theta_0), I(\theta,\theta_1)\}$.*

*(ii) As $c \to 0$ and $\theta_1 \to \theta_0$ such that $(\theta_1 - \theta_0)^2/c \to \infty$,*

$$\log\alpha \sim \log\beta \sim \log\left(\frac{c}{d^2}\right), \qquad \textit{where } d = \theta_1 - \theta_0, \tag{3.51}$$

$$\sup_\theta E_\theta N(g,c) \sim \frac{8d^{-2}(\log c^{-1}d^2)}{\psi''(\theta_0)} \sim \inf_{(T,\delta)\in\mathscr{T}(\alpha,\beta)} \sup_\theta E_\theta T. \tag{3.52}$$

*Moreover, for every distribution function $G$ on $A$ having a positive continuous derivative $G'$ in some neighborhood of $\theta_0$,*

$$\begin{aligned}\int E_\theta N(g,c)\,dG(\theta) &\sim \left(\frac{8G'(\theta_0)}{\psi''(\theta_0)}\right) d^{-1}\log\left(\frac{d^2}{c}\right)\\ &\sim \inf_{(T,\delta)\in\mathscr{T}(\alpha,\beta)}\int E_\theta T\,dG(\theta).\end{aligned} \tag{3.53}$$

Lai (1988) makes use of Lemma 3.2 to prove that the GLR test is asymptotically Bayes risk efficient as the cost $c$ per observation approaches 0. This is the content of the following.

**Theorem 3.5.** *Let $G$ be a prior distribution on $A$, and let $r(T,\delta)$ be the Bayes risk (3.35) of a test of $H_0:\theta\le\theta_0$ versus $H_1:\theta\ge\theta_1$, with cost $c$ per observation and loss $l(\theta)$ for the wrong decision such that $\int l(\theta)\,dG(\theta)<\infty$ and*

$$l(\theta)\ge\triangle \qquad \textit{for all } \theta\notin(\theta_0,\theta_1) \textit{ and some } \triangle>0. \tag{3.54}$$

*Let $g$ be a nonnegative function on $(0,\infty)$ such that (3.49) holds for some $\xi>-1/2$.*

*(i) Assume that $G([\theta_0-t,\theta_0])>0$ and $G([\theta_1,\theta_1+t])>0$ for all $t>0$ and that for some $\rho>0$ and $\varepsilon>0$,*

$$G([x,y])\le\rho(y-x) \qquad \textit{for all } x,y\in[\theta_0-\varepsilon,\theta_0]\cup[\theta_1,\theta_1+\varepsilon] \textit{ with } x<y.$$

*Then for fixed $\theta_0$ and $\theta_1$, as $c\to0$,*

$$r(N(g,c),\delta^*)\sim c|\log c|\int_A\frac{dG(\theta)}{J(\theta)}\sim\inf_{(T,\delta)} r(T,\delta).$$

*(ii) Assume that $G$ has a positive continuous density $G'$ in some neighborhood of $\theta_0$. Then as $c\to0$ and $\theta_1\to\theta_0$ such that $(\theta_1-\theta_0)^2/c\to\infty$,*

$$r(N(g,c),\delta^*)\sim\frac{8G'(\theta)}{\psi''(\theta)}c(\theta_1-\theta_0)^{-1}\log\left[\frac{(\theta_1-\theta_0)^2}{c}\right]\sim\inf_{(T,\delta)} r(T,\delta).$$

*(iii) Suppose that $l(\theta)\to1$ as $(\theta-\theta_0)I_{\{\theta<\theta_0\}}+(\theta-\theta_1)I_{\{\theta>\theta_1\}}\to0$. Let $0\le\gamma<\infty$ and $g_\gamma(t)=(h_\gamma(t)+\gamma t)^2/(2t)$. Then $g_\gamma$ satisfies condition (3.49) with $\xi=1/2$. Assume that $G$ has a positive continuous density $G'$ in some neighborhood of $\theta_0$. Then as $c\to0$ and $\theta_1\to\theta_0$ such that $(\theta_1-\theta_0)/(2c^{1/2})\to\gamma$,*

$$\begin{aligned}\inf_{(T,\delta)} r(T,\delta)\sim r(N(g_\gamma,c),\delta^*)\sim\frac{c^{1/2}G'(\theta_0)}{(\psi''(\theta_0))^{1/2}}\Big\{\int_{-\infty}^{\infty}E(\tau_\gamma|\mu)\,d\mu\\ +\int_{-\infty}^{-\gamma}P\{w(\tau_\gamma)>0|\mu\}\,d\mu+\int_{\gamma}^{\infty}P\{w(\tau_\gamma)<0|\mu\}\,d\mu\Big\},\end{aligned}$$

*where $w(t)$, $t \geq 0$, denotes the Wiener process with drift $\mu$ under $P(\cdot|\mu)$ and $\tau_\gamma = \inf\{t > 0 : |w(t)| \geq h_\gamma(t)\}$, in which $h_\gamma(\cdot)$ is introduced in (3.46).*

Lemma 3.2 reveals an interesting connection between the sequential GLR test with stopping rule $N(g,c)$ and the 2-SPRT discussed in Sect. 3.7.1. Since the ideal value $\theta^*$ in the 2-SPRT is the true parameter $\theta$ that is unknown, it is natural to try replacing $\theta$ by its maximum likelihood estimator $\tilde{\theta}_n$. The accuracy of $\tilde{\theta}_n$ as an estimate of $\theta$ varies with $n$, and the stopping rule $N(g,c)$ takes this into account by using simple time-varying boundary $g(cn)$. Thus, the sequential GLR test with stopping rule $N(g,c)$ can be viewed as an adaptive 2-SPRT, with the value of $\theta^*$ being chosen adaptively and with a corresponding adjustment of the stopping boundary to account for the uncertainty in the estimate $\tilde{\theta}_n$. Lemma 3.2 shows that this idea still leads to a first-order asymptotically optimal solution in ignorance of $\theta$ for fixed $\theta_0$ and $\theta_1$, although the conclusion is weaker than the higher-order asymptotically optimum character given by Theorem 3.4 for the 2-SPRT with $\theta^*$ equal to the true parameter $\theta$. Moreover, even as $\theta_1 \to \theta_0$, Lemma 3.2 shows that if $(\theta_1 - \theta_0)^2/c \to \infty$, then the test $(N(g,c), \delta^*)$ asymptotically minimizes not only the maximal expected sample size $\sup_\theta E_\theta T$ but also $\int E_\theta T dG\theta$ for a large class of prior distributions $G$, among all tests that satisfy the prescribed error constraints.

This first-order asymptotic optimality of the sequential GLR translates easily into its asymptotic Bayes risk efficiency, by integrating the error probabilities and expected sample size at given $\theta$ over the prior distribution, yielding Theorem 3.5(i) and (ii). Only in the case $\theta_1 - \theta_0 = O(\sqrt{c})$ considered in Theorem 3.5(iii) do we need to solve the Bayes problem directly, instead of relying on asymptotic approximations to the error probabilities and expected sample size and applying Hoeffding's lower bound. But even that case can be approximated by an optimal stopping problem for a Wiener process, which in turn can be solved by analytic and numerical methods. Not only does this approach provide an approximate solution to the Bayes problem of testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta \geq \theta_1$ with an indifference zone but it also gives a unified method to treat the one-sided hypotheses $H_0' : \theta \leq \theta_0$ versus $H_1' : \theta > \theta_0$, which Lai (1988) used to prove the following.

**Theorem 3.6.** *Let $G$ be a prior distribution on $A$ such that $G$ has a positive continuous density $G'$ in some neighborhood of $\theta_0$ $(\in A)$. For a sequential test $(T,\delta)$ of $H_0' : \theta \leq \theta_0$ versus $H_1' : \theta > \theta_0$, define the Bayes risk (3.35) in which $\int_{\theta \geq \theta_1}$ is replaced by $\int_{\theta > \theta_0}$ and $l(\theta) = 1$ for the wrong decision. Let $T_c = \inf\{n \geq 1 : nI(\tilde{\theta}_n, \theta_0) \geq g_0(cn)\}$, where $g_0 = h_0^2/(2t)$ and $h_0(\cdot)$ is introduced in (3.46). Let $\delta^*$ denote the terminal decision rule that accepts $H_0'$ or $H_1'$ according as $\tilde{\theta}_{T_c} < \theta_0$ or $\tilde{\theta}_{T_c} > \theta_0$. Then as $c \to 0$,*

$$\inf_{(T,\delta)} r(T,\delta) \sim r(T_c^*, \delta^*) \sim \frac{c^{1/2} G'(\theta_0)}{(\psi''(\theta_0))^{1/2}} \left\{ \int_{-\infty}^{\infty} E(\tau_0|\mu)\, d\mu \right.$$
$$\left. + \int_{-\infty}^{0} P[w(\tau_0) > 0|\mu]\, d\mu + \int_{0}^{\infty} P[w(\tau_0) < 0|\mu]\, d\mu \right\},$$

*where $\tau_0 = \inf\{t > 0 : |w(t)| \geq h_0(t)\}$.*

Lai (1997) has further extended this approach to more general loss functions and prior distributions. Lai and Zhang (1994) and Chan and Lai (2000) have also provided extensions to sequential GLR tests in multiparameter exponential formulas.

## 3.8 ADP and Applications to Phase I Cancer Trial Designs

Section 3.6 has introduced dynamic programming (DP) to solve sequential stochastic optimization problems. In particular, we used DP to prove the optimality of Wald's SPRT for simple hypotheses based on i.i.d. observations and to formulate the optimal stopping problems associated with Bayes sequential tests of composite hypotheses. The latter problems are much harder to solve by directly applying DP, and we have resorted to approximations. Some systematic methods to approximate DP solutions have been developed in the past 2 decades under the rubric of *ADP*. In this section we give an introduction to some ADP methods and describe how Bartroff and Lai (2010) made use of them to resolve the dilemma between two conflicting goals in a Phase I cancer trial: (a) determination of the MTD for a future Phase II trial and (b) safe treatment of current patients in the trial, preferably at doses near the unknown MTD, and thereby to improve the Phase I cancer trial designs in Sect. 2.5. We begin by formulating the stochastic optimization problem that incorporates this dilemma. We then introduce some basic ADP techniques and apply them to address the stochastic optimization problem.

### *3.8.1 A Stochastic Optimization Problem Related to the Treatment Versus Experimentation Dilemma*

While the widely used 3+3 dose finding schemes seem reasonable for an initial group of patients that are the first human subjects to ever receive the treatment, they are very inefficient designs for estimating the MTD to be used in a subsequent Phase II trial. Moreover, even if the investigators should be content with getting IRB approval to try the treatment on human patients and thereby obtain some data and experience, there is the ethical dilemma that patients in the trial are treated at subtherapeutic albeit safe doses. Bartroff and Lai (2010) recommend beginning with a 3+3 design that is used to initialize the trial before switching to the Bayes rule in a two-stage design, the second stage of which is related to minimizing a global risk function defined below.

Following Babb et al. (1998), Bartroff and Lai (2010) specify the prior distribution on $\boldsymbol{\theta} := (\alpha, \beta)$ in the logistic regression model (2.28) by first specifying a range $[x_{\min}, x_{\max}]$ of possible dose values believed to contain the MTD, with $x_{\min}$

believed to be a conservative starting value. Rather than directly specifying the prior distribution $\pi$ for the unknown parameter $\boldsymbol{\theta}$ of the working model to be used in the second stage, which may be hard for investigators to do in practice, an upper bound $q > 0$ on the probability $\rho = F_{\boldsymbol{\theta}}(x_{\min})$ of toxicity at $x_{\min}$ can be elicited from investigators; uniform distributions over $[x_{\min}, x_{\max}]$ and $[0, q]$ are then taken as the prior distributions for the MTD and $F_\theta(x_{\min})$, respectively. Let $\mathscr{F}_k$ denote the information set generated by the first $k$ doses and responses, that is, by $(x_1, y_1), \ldots, (x_k, y_k)$. Letting $\eta$ denote the MTD, it is convenient to transform from the unknown parameters $(\alpha, \beta)$ in the two-parameter logistic model (2.28) to $(\rho, \eta)$ via the formulas

$$\alpha = \frac{x_{\min} \log(1/p - 1) - \eta \log(1/\rho - 1)}{\eta - x_{\min}} \tag{3.55}$$

$$\beta = \frac{\log(1/\rho - 1) - \log(1/p - 1)}{\eta - x_{\min}} \tag{3.56}$$

giving

$$\alpha + \beta x = \frac{(x - \eta) \log(1/\rho - 1) - (x - x_{\min}) \log(1/p - 1)}{\eta - x_{\min}} = \psi(x, \rho, \eta). \tag{3.57}$$

Assuming that the joint prior distribution of $(\rho, \eta)$ has density $\pi(\rho, \eta)$ with support on $[0, q] \times [x_{\min}, x_{\max}]$, the $\mathscr{F}_k$-posterior distribution of $(\rho, \eta)$ has density

$$f(\rho, \eta | \mathscr{F}_k) = C^{-1} \prod_{i=1}^{k} \left[ \frac{1}{1 + e^{-\psi(x_i, \rho, \eta)}} \right]^{y_i} \left[ \frac{1}{1 + e^{\psi(x_i, \rho, \eta)}} \right]^{1 - y_i} \pi(\rho, \eta) \tag{3.58}$$

where

$$C = \int_{x_{\min}}^{x_{\max}} \int_0^q \prod_{i=1}^{k} \left[ \frac{1}{1 + e^{-\psi(x_i, \rho, \eta)}} \right]^{y_i} \left[ \frac{1}{1 + e^{\psi(x_i, \rho, \eta)}} \right]^{1 - y_i} \pi(\rho, \eta) \, d\rho \, d\eta$$

is the normalizing constant. The marginal $\mathscr{F}_k$-posterior distribution of $\eta$ is then

$$f(\eta | \mathscr{F}_k) = \int_0^q f(\rho, \eta | \mathscr{F}_k) \, d\rho. \tag{3.59}$$

The CRM and EWOC doses based on $\mathscr{F}_k$ described in Sect. 2.5.2 are the mean and the $\omega$-quantile of (3.59), respectively.

Note that using EWOC or CRM amounts to the "myopic" policy of dosing the $(k+1)$th patient at the dose $x_{k+1} = x$ that minimizes $E[h(x, \eta) | \mathscr{F}_k]$, in which

$$h(x,\eta) = \begin{cases} (x-\eta)^2 & \text{for CRM} \\ \omega(\eta - x)^+ + (1-\omega)(x-\eta)^+ & \text{for EWOC} \end{cases} \tag{3.60}$$

where $x^+ = \max(x,0)$ and

$$E[h(x,\eta)|\mathscr{F}_k] = \int_{x_{\min}}^{x_{\max}} h(x,\eta) f(\eta|\mathscr{F}_k)\, d\eta.$$

Since the information about the dose–toxicity relationship gained from $x_{k+1}$ and the response $y_{k+1}$ affects the ability to safely and effectively dose the other patients $k+2, k+3, \ldots, n$, one potential weakness of these myopic policies is that they may be inadequate in generating information on $\theta$ for treating the rest of the patients, as well as the post-experimental estimate of the MTD to be used for future patients. To incorporate these considerations in a Phase I trial, $x_1, x_2, \ldots, x_n$ should be chosen sequentially in such a way as to minimize the *global risk*

$$E\left[\sum_{i=1}^{n} h(x_i,\eta) + g\left(\hat{\eta},\eta\right)\right], \tag{3.61}$$

in which the expectation is taken over the joint distribution of $(\rho,\eta; x_1, y_1, \ldots, x_n, y_n)$. Note that (3.61) measures the effect of the dose $x_k$ on the $k$th patient through $h(x_k,\eta)$, its effect on future patients in the trial through $\sum_{i=k+1}^{n} h(x_i,\eta)$, and its effect on the posttrial estimate $\hat{\eta}$ through $g(\hat{\eta},\eta)$. It can therefore be used to address the dilemma between safe treatment of current patients in the study and efficient experimentation to gather information about $\eta$ for future patients.

Dynamic programming is a standard approach to the finite-horizon problem of minimizing the global risk (3.61), which can in principle be solved by backward induction. Specifically, define

$$h_k(x) = \begin{cases} E\left[h(x,\eta)|\mathscr{F}_k\right] & 0 \le k < n-1 \\ E\left[h(x,\eta) + g\left(\hat{\eta}(x_1, \ldots, x_{n-1}, x), \eta\right)|\mathscr{F}_{n-1}\right] & k = n-1. \end{cases} \tag{3.62}$$

To minimize (3.61), dynamic programming solves for the optimal design $x_1^*, \ldots, x_n^*$ by backward induction that determines $x_k^*$ by minimizing

$$h_{k-1}(x) + E\left[\sum_{i=k+1}^{n} h_{i-1}(x_i^*) \,\middle|\, \mathscr{F}_{k-1}, x_k = x\right] \tag{3.63}$$

after determining the future dose levels $x_{k+1}^*, \ldots, x_n^*$. Note that (3.63) involves computing the conditional expectation of $\sum_{i=k+1}^{n} h_{i-1}(x_i^*, \eta)$ given the dose $x$ at stage $k$ and the information set $\mathscr{F}_{k-1}$ and that $x_k^*$ is determined by minimizing such conditional expectation over all $x$. For $i \ge k+1$, since $x_i^*$ is a complicated nonlinear function of the past observations and of $y_k, x^*_{k+1}, y_{k+1}, \ldots, x^*_{i-1}, y_{i-1}$ that are not yet observed, evaluation of the aforementioned conditional expectation is a formidable task. To overcome this difficulty, Bartroff and Lai (2010) use ADP techniques introduced below to tackle the problem of minimizing the global risk (3.61) via a hybrid design.

### 3.8.2 Some ADP Methods

**Rollout**

To begin with, consider the problem of minimizing (3.61) with $g = 0$ and $h(x; \alpha, \beta) = (\alpha + \beta x - y^*)^2$ in the linear regression model $y_k = \alpha + \beta x_k + \varepsilon_k$ with i.i.d. normal errors $\varepsilon_i$ having mean 0. Assuming a normal prior distribution of $(\alpha, \beta)$, the posterior distribution of $(\alpha, \beta)$ given $\mathscr{F}_{i-1}$ is also bivariate normal with parameters $E_{i-1}(\alpha)$, $E_{i-1}(\beta)$, $E_{i-1}(\alpha^2)$, $E_{i-1}(\beta^2)$, $E_{i-1}(\alpha\beta)$, in which $E_{i-1}$ denotes conditional expectation given $\mathscr{F}_{i-1}$. These conditional moments have explicit recursive formulas; see Sect. 4 of Han et al. (2006). The myopic policy that chooses $x$ at stage $i$ to minimize $E[(\alpha + \beta x - y^*)^2 | \mathscr{F}_{i-1}]$ is given explicitly by

$$\hat{x}_i = E_{i-1}\left\{(y^* - \alpha)\beta\right\}/E_{i-1}(\beta^2) = \left\{y^* E_{i-1}(\beta) - E_{i-1}(\alpha\beta)\right\}/E_{i-1}(\beta^2).$$

Although the myopic policy is suboptimal for the global risk function (3.61), Han et al. (2006) use it as a substitute for the intractable $x^*_i$ for $k+1 \le i \le n$ in (3.63), in which the conditional expectation can then be evaluated by Monte Carlo simulation. This method is called *rollout* in ADP. The idea is to approximate the optimal policy $x^*_k$ by minimizing (3.63) with $x^*_{k+1}, \ldots, x^*_n$ replaced by some known *base policy* $\hat{x}_{k+1}, \ldots, \hat{x}_n$, which ideally is some easily computed policy that is not far from the optimum. Specifically, given a base policy $\hat{\boldsymbol{x}} = (\hat{x}_1, \ldots, \hat{x}_n)$, let $\hat{x}^{(1)}_k$ be the $x$ that minimizes

$$h_{k-1}(x) + E\left[\sum_{i=k+1}^{n} h_{i-1}(\hat{x}_i) \,\middle|\, \mathscr{F}_{k-1}, \hat{x}_k = x\right], \tag{3.64}$$

and the expectation in the second term in (3.64) is typically evaluated by Monte Carlo simulation. The policy $\hat{\boldsymbol{x}}^{(1)} = (\hat{x}^{(1)}_1, \ldots, \hat{x}^{(1)}_n)$ is called the *rollout* of $\hat{\boldsymbol{x}}$ and has been used for stochastic control problems arising in a variety of applications; see Sect. 2.1 of Han et al. (2006). The rollout $\hat{\boldsymbol{x}}^{(1)}$ may itself be used as a base policy, yielding $\hat{\boldsymbol{x}}^{(2)}$, and in theory, this process may be repeated an arbitrary number of times, yielding $\hat{\boldsymbol{x}}^{(1)}, \hat{\boldsymbol{x}}^{(2)}, \hat{\boldsymbol{x}}^{(3)}, \ldots$. Letting $R(\boldsymbol{x}) = E[\sum_{i=1}^{n} h_{i-1}(x_i)]$, Bayard (1991) showed that, regardless of the base policy, rolling out $n$ times yields the optimal design and that rolling out always improves the base design, that is, that

$$R(\hat{\boldsymbol{x}}) \geq R(\hat{\boldsymbol{x}}^{(1)}) \geq R(\hat{\boldsymbol{x}}^{(2)}) \geq \cdots \geq R(\hat{\boldsymbol{x}}^{(n)}) = R(\boldsymbol{x}^*) \tag{3.65}$$

for any policy $\hat{\boldsymbol{x}}$, where $\boldsymbol{x}^*$ denotes the optimal policy.

For the global risk function (3.61) associated with Phase I designs, with $h$ given by (3.60), one can use the myopic design EWOC or CRM as the base design in the rollout procedure. In contrast with the explicit formula for the case of a linear regression model with normal errors $\varepsilon_t$, the posterior distribution with density function (3.58) does not have finite-dimensional sufficient statistics, and the myopic design involves (a) bivariate numerical integration to evaluate $E[h_i(x_{i+1})|\mathscr{F}_{k-1},\ x_k = x]$ for $i \geq k$ and (b) minimization of the conditional expectation over $x$. Although (3.65) says that rolling out a base design can improve it and rolling out $n$ times yields the dynamic programming solution, in practice it is difficult to use a rollout (which is defined by a backward induction algorithm that involves Monte Carlo simulations followed by numerical optimization at every stage) as the base policy for another rollout. To overcome this difficulty, we need a tractable representation of successive rollouts, which we develop by using other ideas from ADP.

### Combining Least Squares with Monte Carlo

The conditional expectation in (3.63), as a function of $x$, is called the *cost-to-go function* in dynamic programming. An ADP method, which grew out of the machine learning literature, is based on two statistical concepts concerning the conditional expectation. First, for given $x$ and the past information $\mathscr{F}_{k-1}$, the conditional expectation is an expectation and therefore can be evaluated by Monte Carlo simulations, if one knows how $h_k(x^*_{k+1}),\ldots,h_{n-1}(x^*_n)$ are generated. The second concept is that, by (3.62), $h_i(x_{i+1})$ is a conditional expectation given $\mathscr{F}_i$, which is a regression function (or minimum-variance prediction) of $h_i(x_{i+1})$, with regressors (or predictors) generated from $\mathscr{F}_i$. Based on a large sample (generated by Monte Carlo), the regression function can be estimated by least squares using basis function approximations, as is typically done in nonparametric regression. Combining least squares (LS) regression with Monte Carlo (MC) simulations yields the following LS-MC method for Markov decision problems. Let $\{s_t, t \geq 0\}$ be a Markov chain whose transition probabilities from state $s_t$ to $s_{t+1}$ depend on the action $x_t$ at time $t$, and let $f_t(s,x)$ denote the cost function at time $t$, incurred when the state is $s$ and the action $x$ is taken. Consider the statistical decision problem of choosing $x$ at each stage $k$ to minimize the cost-to-go function

$$Q_k(s,x) = E\left\{ f_k(s,x) + \sum_{t=k+1}^{n} f_t(s_t,x_t) \,\middle|\, s_k = s, x_k = x \right\}, \tag{3.66}$$

assuming that $x_{k+1},\ldots,x_n$ have been determined. Let

$$V_k(s) = \min_x Q_k(s,x), \quad x^*_k = \arg\min_x Q_k(s,x). \tag{3.67}$$

These functions can be evaluated by the backward induction algorithm of dynamic programming: $V_n(s) = \min_x f_n(s,x)$, and for $n > k \geq 1$,

$$V_k(s) = \min_x \{f_k(s,x) + E[V_{k+1}(s_{k+1})|s_k = s, x_k = x]\}, \tag{3.68}$$

in which the minimizer yields $x_k^*$. Assuming the state space to be finite-dimensional (e.g., $\mathbb{R}^n$), the LS-MC method uses basis functions $\phi_j$, $1 \leq j \leq J$, to approximate $V_{k+1}$ by $\hat{V}_{k+1} = \sum_{j=1}^J a_{k+1,j}\phi_j$, and uses this approximation together with $B$ Monte Carlo simulations to approximate

$$E\left[V_{k+1}(s_{k+1})|s_k = s,\ x_k = x\right]$$

for every $x$ in a grid of representative values. This yields an approximation $\tilde{V}_k$ to $V_k$ and also $\hat{x}_k$ to $x_k^*$. Moreover, using the sample

$$\left\{(s_{k,b}, \tilde{V}_k(s_{k,b}), 1 \leq b \leq B\right\} \tag{3.69}$$

generated by the control action $\hat{x}_k$, we can perform least squares regression of $\tilde{V}_k(s_{k,b})$ on $(\phi_1(s_{k,b}), \ldots, \phi_J(s_{k,b}))$ to approximate $\tilde{V}_k$ by $\hat{V}_k = \sum_{j=1}^J a_{k,j}\phi_j$. Further details of this approach can be found in Chap. 6 of Bertsekas (2007).

### Approximation in Policy Space

Although the problem (3.63) can be viewed as a Markov decision problem with the $\mathscr{F}_{t+1}$-posterior distribution being the state $s_t$, the state space of the Markov chain at hand is infinite-dimensional, consisting of all bivariate posterior distributions of the unknown parameter vector $(\alpha, \beta)$. Unlike the preceding paragraph for finite-dimensional state space, in the infinite-dimensional case, there is no such simple choice of basis functions of posterior distributions, which are the states. As pointed out in Sect. 6.7 of Bertsekas (2007), an alternative to approximating the value functions $V_k$ as in the preceding paragraph, called *approximation in value space*, is to approximate the optimal policy by a parametric family of policies so that the total cost can be optimized over the parameter vector. This approach is called *approximation in policy space*, and most of its literature has focused on finite-state Markov decision problems and gradient-type optimization methods that approximate the derivatives of the costs, as functions of the parameter vector, by simulation.

Bartroff and Lai (2010) have introduced a new method for approximation in policy space, which uses iterated rollouts to optimize the parameters in a suitably chosen parametric family of policies. The choice of the family of policies should involve domain knowledge and reflect the kind of policies that one would like to use for the actual application. One would therefore start with a set of real-valued basis functions of the state $s_t$ of the Markov chain with general, possibly infinite-dimensional, state space, on which the family of chosen policies will be based. The control policies in this family can be represented by $\pi_t(\phi_1(s_t),\ldots,\phi_m(s_t);\boldsymbol{\beta})$, which is the action taken at time $t$ (after $s_t$ has been observed and the basis functions $\phi_1(s_t),\ldots,\phi_m(s_t)$ have been evaluated) and in which $\boldsymbol{\beta}$ is a parameter to be chosen iteratively by using successive rollouts, with

$$\left\{\pi_t(\phi_1(s_t),\ldots,\phi_m(s_t);\boldsymbol{\beta}^{(j)}),\ 1\le t\le n\right\}$$

being the base policy for the rollout $\boldsymbol{x}^{(j+1)}$. Using the simulated sample

$$\left\{(s_{k,b},x_{k,b}^{(j+1)}),\ 1\le b\le B\right\},$$

in which $s_{k,b}$ denotes the $b$th simulated replicate of $s_k$, least squares regression of $x_{k,b}^{(j+1)}$ on $\pi_k(\phi_1(s_{k,b}),\ldots,\phi_m(s_{k,b});\boldsymbol{\beta})$ is performed to estimate $\boldsymbol{\beta}$ by $\boldsymbol{\beta}^{(j+1)}$; nonlinear least squares is used if $\pi_k$ is nonlinear in $\boldsymbol{\beta}$. In view of (3.65), each iteration is expected to provide improvements over the preceding one. A concrete example of this method in a prototypical Phase I setting is given in the next section, where linear regression splines are used in iterated rollouts. In this setting the state variable $s_t$ represents the complete treatment history up to time $t$ in the trial—all prior distributions, doses, and responses up to that time—and the cost function $f_t(s_t,x)$ will be replaced by $h_t(x)$ given by (3.62).

### 3.8.3 *Hybrid Designs Derived by ADP*

As pointed out in Sect. 3.8.1, the objective function of the dynamic programming problem (3.61) involves both experimentation (for estimating the MTD) and treatment (for the patients in the study). Consider the $k$th patient in a trial of length $n$ $(\ge k)$. If the $k$th patient were the last patient to be treated in the trial ($n=k$), the best dose to give him/her would be the myopic dose $m_k$ that minimizes the future risk $h_{k-1}(x_k)$. On the other hand, early on in the trial, especially if $n-k$ is relatively large, one expects the optimal dose to be perturbed from $m_k$ in the direction of a dose that provides more information about the dose–response model, for the relatively large number of doses that will have to be set for the future patients. Since the optimal design theory for learning the MTD under overdose constraints, developed by Haines et al. (2003), yields a $c$- or $D$-optimal design $\ell_k$ (see Sect. 2.3.3), Bartroff and Lai (2010) propose to use the following *hybrid design* representation of the optimal dose sequence:

$$x_k^* = (1-\varepsilon_k)m_k + \varepsilon_k\ell_k \tag{3.70}$$

where $\ell_k$ is the chosen "learning design." Of course any dosing policy admits the representation (3.70) with

$$\varepsilon_k = \frac{x_k^* - m_k}{\ell_k - m_k} \cdot \mathbf{1}_{\{\ell_k \neq m_k\}}.$$

However, we will show that it is possible to use rollouts to choose $\varepsilon_k$ of a simple form, not depending on $x_k^*$, such that the resulting hybrid design given by right-hand side of (3.70) is highly efficient.

### Choice of the Information-Driven Design

The theory of optimal experimental designs in generalized linear (in particular, logistic) regression models is concerned with choosing the design levels to give the estimate with the smallest determinant or some other function of the asymptotic matrix at the end of an experiment whose objective is to generate information about (or "learn") the unknown parameters; see Sects. 2.3 and 2.6. Since the asymptotic covariance matrix involves the unknown regression parameters for nonlinear models, one has to use adaptive or sequential designs to achieve optimality. Bartroff and Lai (2010) use the Bayesian $c$-optimal design with $c = (0,1)'$ as the learning design $\ell_k$ in (3.70), giving $c'\theta = c'(\alpha,\beta)' = \beta$, which is optimal for learning about $\beta$ or, equivalently, about the slope

$$\frac{\partial}{\partial x} E(y|x)\bigg|_{x=\eta} = \frac{\partial}{\partial x}\left(\frac{1}{1+e^{-(\alpha+\beta x)}}\right)\bigg|_{x=\eta} = \beta p(1-p)$$

of the dose response curve (2.28) at the MTD, with $p = 1/3$.

### Using Iterated Rollouts in ADP to Determine $\varepsilon_k$

Since the treatment versus experimentation dilemma discussed in Sect. 3.8.1 stems from the uncertainty in the current estimate of the MTD $\eta$, it is natural to expect that the amount of perturbation from the myopic dose $m_k$ depends on the degree of such uncertainty, using little perturbation when the posterior distribution of $\eta$ is peaked and much more perturbation when it is spread out. This suggests choosing $\varepsilon_k$ as a function of the posterior variance $v_{k-1}^2 = \mathrm{Var}(\eta|\mathscr{F}_{k-1})$, whose reciprocal is called the *precision* of $E(\eta|\mathscr{F}_{k-1})$ in Bayesian parlance. Bartroff and Lai (2010) use functions of $s_k = v_{k-1}/v_0$ as basic features of the posterior distribution of $\eta$ to approximate the $\varepsilon_k$ in (3.70). They use the rollout algorithm in ADP to determine the functions $\varepsilon_k = \varepsilon_k(s_k)$.

The idea behind the rollout algorithm is iterative policy improvement, beginning with a base policy $\hat{\boldsymbol{x}} = (\hat{x}_1, \ldots, \hat{x}_n)$, for which Bartroff and Lai (2010) choose EWOC.

Let $\hat{x}_k^{(1)}$ be the $x$ that minimizes

$$h_{k-1}(x) + E\left[\sum_{i=k+1}^{n} h_{i-1}(\hat{x}_i) \middle| \mathscr{F}_{k-1}, \hat{x}_k = x\right] \tag{3.71}$$

and the expectation in the second term in (3.71) is typically evaluated by Monte Carlo simulation. The policy $\hat{\boldsymbol{x}}^{(1)} = (\hat{x}_1^{(1)}, \ldots, \hat{x}_n^{(1)})$ is called the *rollout* of $\hat{\boldsymbol{x}}$. Thus, Monte Carlo simulations are performed to obtain the rollout $\boldsymbol{x}^{(1)}$ of EWOC, yielding a simulated sample $\{(e_{k,b}, s_{k,b}),\ 1 \le b \le B\}$, where $e_{k,b}$ is the $b$th simulated replicate of

$$e_k = \frac{x_k^{(1)} - m_k}{\ell_k - m_k} \cdot \mathbf{1}_{\{\ell_k \neq m_k\}}, \tag{3.72}$$

which is essentially the same as (3.70) with $(x_k^*, \varepsilon_k)$ replaced by $(x_k^{(1)}, e_k)$.

Another technique in ADP that Bartroff and Lai (2010) use is based on two statistical concepts concerning conditional expectations. First, for given $x$ and the past information $\mathscr{F}_{k-1}$, the conditional expectation is an expectation and therefore can be evaluated by Monte Carlo simulations, if one knows how $h_k(x_{k+1}^*), \ldots, h_{n-1}(x_n^*)$ are generated. The second concept is that, by (3.62), $h_i(x_{i+1})$ is a conditional expectation given $\mathscr{F}_i$, which is a regression function (or minimum-variance prediction) of $h_i(x_{i+1})$, with regressors (or predictors) generated from $\mathscr{F}_i$. Based on a large sample (generated by Monte Carlo), the regression function can be estimated by least squares using basis function approximations, The basis function approximation used by Bartroff and Lai (2010) is a truncated linear function

$$f_1(s) = \min\left\{1, \left(\beta_k^{(0)} + \beta_k^{(1)} s\right)^+\right\}, \qquad \text{for } s_* \le s \le s^*, \tag{3.73}$$

where $s_*$ and $s^*$ are the minimum and maximum of the sample values $s_{k,b}$, $1 \le b \le B$, which they extend beyond the range $[s_*, s^*]$ by

$$f_1(s) = \begin{cases} s f_1(s_*)/s_* & 0 \le s \le s_* \\ f_1(s^*) & s \ge s^*. \end{cases} \tag{3.74}$$

This agrees with the constraint $f_1(0) = 0$ and ensures that the weight assigned to experimentation does not exceed $f_1(s^*)$. A further simplification is to group the data into $K$ blocks so that $\varepsilon_k = \varepsilon_k(s)$ does not vary with $k$ within each block, since it is expected that the amount of experimentation for the initial stages depends mostly on the uncertainty about $\eta$ while for the final stages, experimentation would only benefit the posttrial estimate of $\eta$. Regressing $e_{k,b}$ on $s_{k,b}$ yields the estimated regression function $f_1$. Letting $\hat{e}_k = f_1(s_k)$, the hybrid design $x_k = (1 - \hat{e}_k)m_k + \hat{e}_k \ell_k$ can then be used as the base policy to form the rollout $\boldsymbol{x}^{(2)}$, and this procedure can be repeated to obtain the iterated rollouts $\boldsymbol{x}^{(3)}, \boldsymbol{x}^{(4)}, \ldots$.

**Table 3.1** Risk, bias, and RMSE of the final MTD estimate, DLT rate, and overdose (OD) rate of EWOC, rollout of EWOC, and 1st and 2nd hybrid approximations

| Design | Risk | Bias | RMSE | DLT | OD |
|---|---|---|---|---|---|
| EWOC | 0.84 (0.01) | −0.20 (0.010) | 0.31 (0.04) | 29.8% (0.7%) | 21.9% (0.6%) |
| Rollout | 0.75 (0.01) | −0.04 (0.009) | 0.22 (0.03) | 33.0% (0.7%) | 31.2% (0.7%) |
| Hybrid 1 | 0.75 (0.02) | −0.14 (0.012) | 0.29 (0.06) | 33.5% (1.5%) | 37.5% (1.5%) |
| Hybrid 2 | 0.71 (0.01) | −0.04 (0.005) | 0.22 (0.04) | 31.24% (0.9%) | 27.8% (0.9%) |

*Example 3.1.* Bartroff and Lai (2010) illustrate the preceding method with the following example, in which $n = 10$ and $[x_{\min}, x_{\max}]$ is transformed to $[0,1]$ by location and scale changes. Independent uniform priors on $[0,q]$ and $[0,1]$ are used for $\rho = F_\theta(x_{\min})$ and the MTD $\eta$, respectively. In this example, $q = 1/3$ and the EWOC loss is used with $\omega = 1/4$ in (3.60), and the squared error loss $g(\hat{\eta}, \eta) = (\hat{\eta} - \eta)^2$ is used in (3.61). Since $n$ is relatively small, $(\beta_k^{(0)}, \beta_k^{(1)})$ in (3.73) can be assumed not to vary with $k$ so that the common $(\beta^{(0)}, \beta^{(1)})$ can be estimated by applying least squares regression to the sample

$$\left\{(e_{k,b}, s_{k,b}) : 1 \le k \le n,\ 1 \le b \le B\right\};$$

moreover, (3.73) is used for all $s$ without performing the extrapolation beyond $[s_*, s^*]$. Rolling out EWOC as the base design and using $B = 2000$ simulations, the preceding method yielded $(\beta^{(0)}, \beta^{(1)}) = (0.096, 0.02)$. Putting

$$\varepsilon_k = \min\left\{1, (0.096 + 0.02 v_{k-1}/v_0)^+\right\} \tag{3.75}$$

in the hybrid design $x_k^{(1)} = (1 - \varepsilon_k) m_k + \varepsilon_k \ell_k$, Bartroff and Lai (2010) used $\boldsymbol{x}^{(1)}$ as the base policy of a second rollout, for which the preceding procedure yielded $(\beta^{(0)}, \beta^{(1)}) = (-0.72, 0.94)$. Here we used the sequential $c$-optimal design with $c = [0,1]'$ as the learning design $\ell_k$. Table 3.1 contains the operating characteristics, explained below, of EWOC and its rollout, the first hybrid design $\boldsymbol{x}^{(1)}$ with $\varepsilon_k$ given by (3.75) and the second hybrid design $\boldsymbol{x}^{(2)}$ in which $(0.096, 0.02)$ in (3.75) is replaced by $(-0.72, 0.94)$. Each result is based on 2000 simulation runs. Figure 3.2 plots the cumulative risk $R_k = \sum_{i=1}^{k} E[h_{i-1}(x_i)]$ of the EWOC, rollout, and hybrid designs for $k = 1, \ldots, n (= 10)$. The operating characteristics in Table 3.1 are the Monte Carlo estimates of overall risk $R_{10}$, the bias and root mean squared error (RMSE) of the terminal MTD estimate $\hat{\eta}_{10}$, the DLT rate $P(y = 1)$, and the overdose (OD) rate, which is the expected proportion of patients treated at doses higher than $\eta$. Standard errors are given in parentheses.

The first hybrid design, which is an approximation to the rollout design, provides more than 10% improvement in terminal risk $R_{10}$ over the myopic policy.

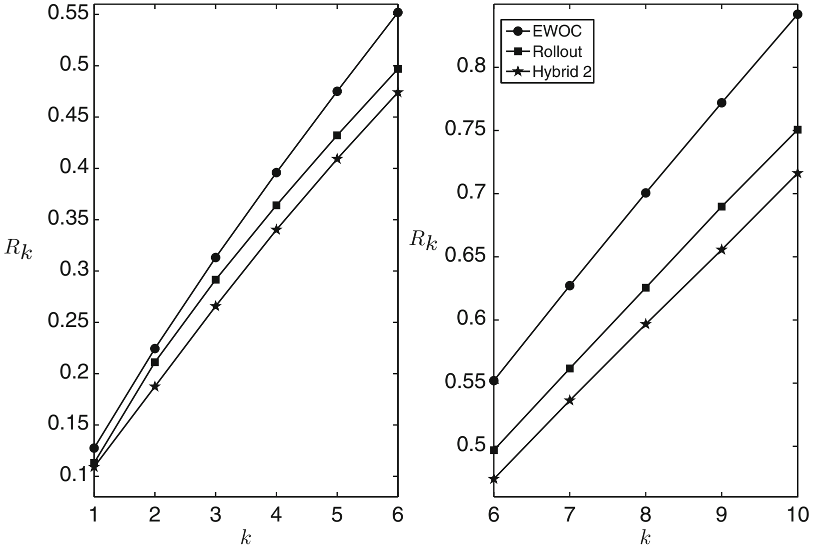


**Fig. 3.2** Cumulative risk $R_k$ for EWOC, rollout of EWOC, and hybrid design

The second hybrid design provides an additional 5% improvement in the terminal risk $R_{10}$ and also smaller values of the DLT and OD rates than the rollout design. The Monte Carlo simulations used to evaluate the operating characteristics and to fit the hybrid designs were performed by using rejection sampling to simulate from the posterior distribution. At each stage, the posterior distribution of $(\rho, \eta)$ is continuous and supported on the compact set $[0,q] \times [x_{\min}, x_{\max}]$; hence, the joint uniform distribution on $[0,q] \times [x_{\min}, x_{\max}]$ is a natural candidate for the instrumental distribution in rejection sampling.

### 3.8.4 A Two-Stage Modification

When one may have concerns about the validity of the Bayesian parametric model in this model-based approach, one can readily incorporate the hybrid designs as the second stage of a two-stage design. For the first stage of such a the two-stage design, the first stage of which escalates the doses cautiously using the traditional 3+3 design. For the batches of three in the traditional 3+3 design, we propose to combine the nonparametric approach with a parametric model-based dose determining scheme, thereby checking the parametric model to be used for model-based escalation in the second stage. This modification of the traditional 3+3 design uses a specified set of dose levels

$$x_{\min} = \lambda_1 < \lambda_2 < \cdots \leq x_{\max}. \tag{3.76}$$

Set $d_1 = \lambda_1 = x_{\min}$. In the $k$th group of three patients, two patients will be treated at the same dose $d_k = \lambda_j$ and one patient at the EWOC dose $m_k$, computed given the doses and responses of the previous $3(k-1)$ patients. If no DLTs occur in the group of three patients, $d_{k+1}$ is increased to $\lambda_{j+1}$. If one DLT occurs, $d_{k+1}$ stays the same at $d_k = \lambda_j$. Otherwise, 2 or 3 DLTS have occurred so the trial is stopped if $d_k = x_{\min}$ and otherwise continued with $d_{k+1}$ lowered to $\lambda_{j-1}$. (Alternatively, it may be desired to stop when 3 toxicities occur, regardless of what $d_k$ was.) Otherwise, the EWOC dose $m_{k+1}$ is updated and the process is repeated with next group of three patients. This process repeats until a certain fraction of the total number $n$ of patients has been treated, provided the first stage has not been stopped for excess toxicities. Simulation studies have shown that switchover points around $n/3$ or $n/4$ seem to strike a balance between enough time for conservative dose escalation and model checking during the first stage, while leaving enough time for efficient dose escalation in the second stage. The benefit of a first stage of conservative dose escalation occurs when the prior model is misspecified, as shown in the following example of Bartroff and Lai (2010).

*Example 3.2.* Babb et al. (1998) used EWOC to design a Phase I trial to determine the MTD, with $p = 1/3$, of the antimetabolite 5-fluorouracil (5-FU) for the treatment of solid tumors in the colon, when taken in conjunction with fixed levels of the agents leucovorin (20 mg/m$^2$) and topotecan (0.5 mg/m$^2$). In this setting, a toxicity is considered a grade 4 hematologic or grade 3 or 4 non-hematologic toxicity within 2 weeks. As mentioned above, EWOC involves specifying pretrial a set of possible dose values (3.76) believed to contain the MTD, where $x_{\min}$ is taken as the starting value. Based on preliminary studies of 5-FU given in conjunction with topotecan, a dose of $x_{\min} = 140$ mg/m$^2$ of 5-FU was believed to be safe when given with 0.5 mg/m$^2$ of topotecan. Also, a previous trial concluded that the MTD of 5-FU was 425 mg/m$^2$ when administered without topotecan, so $x_{\max}$ was taken to be 425 mg/m$^2$ since 5-FU has been observed to be more toxic when given with topotecan than alone. The two-parameter logistic model (2.28) was chosen based on previous experience with the agents, and uniform prior distributions over $[x_{\min}, x_{\max}]$ and $[0, 0.2]$ were chosen for the MTD and the probability $F_\theta(x_{\min})$, respectively. A feasibility bound of $\omega = 0.25$ was chosen for EWOC and $p = 1/3$.

Bartroff and Lai (2010) first compare previous designs, which are single stage, with the rollout (abbreviated by ROLL) of the EWOC design, assuming that $m = 0$ in ROLL in a single-stage trial of length $n = 24$. Besides EWOC and its rollout ROLL, the Bayesian designs they consider include CRM, the constrained $D$-optimal design (abbreviated by $D$-opt) of Haines et al. (2003) with constraint $\varepsilon = 0.05$ and the unconstrained sequential Bayesian $c$-optimal design (abbreviated by $c$-opt) with $c$ being the vector $(0,1)^T$. The prior density is assumed to be uniform:

$$\pi(\rho,\eta) = [q(x_{\max} - x_{\min})]^{-1} \cdot 1\{(\rho,\eta) \in [0,q] \times [x_{\min}, x_{\max}]\} \tag{3.77}$$

**Table 3.2** Risk, bias, and RMSE of the final MTD estimate, DLT rate, and MTD overdose (OD) rate, with SEs in parentheses, of various designs with the MTD fixed at the lower 15th percentile of the misspecified prior

| Design | | Risk | Bias | RMSE | DLT | OD |
|---|---|---|---|---|---|---|
| ROLL | (a) | 1.64 (0.02) | −0.031 (0.003) | 0.142 (0.025) | 30.32% (1.03%) | 39.38% (1.09%) |
| | (b) | 1.39 (0.02) | −0.025 (0.003) | 0.145 (0.026) | 27.41% (1.00%) | 33.39% (1.05%) |
| Hybrid 1 | (a) | 1.82 (0.05) | −0.032 (0.002) | 0.151 (0.027) | 36.90% (1.52%) | 42.31% (1.56%) |
| | (b) | 1.69 (0.04) | −0.027 (0.003) | 0.131 (0.036) | 35.70% (1.51%) | 41.11% (1.56%) |
| EWOC | | 2.29 (0.02) | −0.034 (0.003) | 0.155 (0.028) | 35.33% (1.07%) | 45.98% (1.11%) |
| CRM | | 3.83 (0.02) | 0.037 (0.004) | 0.179 (0.032) | 44.18% (1.11%) | 65.12% (1.07%) |
| 3+3$_{10}$ | | 1.87 (0.01) | 0.060 (0.003) | 0.138 (0.024) | 17.06% (0.84%) | 0.85% (0.21%) |

with $q = 0.2$. The performance of these designs is first evaluated in terms of the global risk (3.61), in which the squared error $g(\hat{\eta}, \eta) = (\hat{\eta} - \eta)^2$ for the MTD estimate $\hat{\eta} = \hat{\eta}(x_i, y_i, \ldots, x_n, y_n)$. We then evaluate performance exclusively in terms of the bias and RMSE of $\hat{\eta}$ without taking into consideration the risk to current patients, noting that the $c$- and $D$-optimal designs focus on errors of posttrial parameter estimates. Finally, since safety of the patients in the trial is the primary concern of traditional 3+3 designs, performance is also evaluated in terms of the DLT. Their results show that the effects of considering the "future" patients is large, with ROLL and Hybrid 1 substantially reducing the global risk from previous designs in the literature. Moreover, ROLL and Hybrid 1 have DLT rates of 27.68% and 24.68%, respectively, well below 33% in the simulation study.

If the true MTD falls in the left tail of the prior distribution of $\eta$, then the prior information about the MTD is biased upward, which can cause overdoses. In this situation, including an initial stage of modified dose escalation, like the modified 3+3 scheme, provides additional safety by refining the prior to be more accurate when it begins to be used in the second stage. Focusing on the EWOC, CRM, ROLL, and Hybrid 1 designs, Table 3.2 contains the results of a simulation study that zeroes in on a situation such as this, where the true MTD is the lower 15th percentile of the MTD's nominal uniform prior distribution on $[x_{\min}, x_{\max}]$. That is, the data are generated with $\eta$ fixed at the 15th percentile of $[x_{\min}, x_{\max}]$ and $\rho$ uniformly distributed over $[0, q]$, with $q = .2$. The nominal prior for $(\rho, \eta)$ used by the Bayesian procedures in Table 3.2 is (3.77). To see the effects of the first stage of more conservative dose escalation, the operating characteristics of ROLL are recomputed using a first stage of length $n/4 = 6$; the dose levels (3.76) used by the modified 3+3 design, described in the first paragraph of this section and denoted 3+3$_{10}$, are ten uniformly spaced levels in $[x_{\min}, x_{\max}] = [140, 425]$. Adding this first stage to ROLL and Hybrid 1 substantially reduces the risk, DLT, and overdose rates, as shown in Table 3.2, in which (a) refers to the case of $n = 24$ dose levels without the modified 3+3 first stage and (b) refers to the two-stage design using a first stage of length $n/4 = 6$ consisting of the modified 3+3 design using 10 uniformly spaced dose levels in $[x_{\min}, x_{\max}] = [140, 425]$.

## 3.9 Supplements and Problems

1. *Overshoots, renewal theory, and boundary crossing probabilities.*
Wald (1945) ignored the overshoots $(L_N/B)I_{\{L_N \geq B\}}$ and $(A/L_N)I_{\{L_N \leq A\}}$ in (3.4) to arrive at the approximations (3.5) to the error probabilities. Since $P_i(N < \infty) = 1$ for $i = 0, 1$, the likelihood ratio identity in fact gives

$$P_0\{L_N \geq B\} = E_1(L_N^{-1} I_{\{L_N \geq B\}}) = B^{-1} E_1 e^{-(l_N - b)} I_{\{l_N \geq b\}},$$

where $l_n = \log L_n = \sum_{i=1}^n \log(q_i(X_i|X_1, \ldots, X_{i-1})/p_i(X_i|X_1, \ldots, X_{i-1}))$ and $b = \log B$. In Wald's setting of i.i.d. $X_i$, $l_n$ is a random walk and renewal theory gives the limiting distribution of $l_N - b$ as $b \to \infty$. Siegmund (1985) introduces this kind of methods to develop analytic approximations to error probabilities for sequential tests based on random walks with linear or curved boundaries; the curved boundaries require nonlinear renewal theory described in his Chap. IX. Lai (2004, Sect. 4) shows how the likelihood ratio identity can be used to develop (a) asymptotic approximations to error probabilities in the dependent case and (b) importance sampling techniques for numerical computation of boundary crossing probabilities by importance sampling methods.
2. Prove (3.28) and (3.33).
3. *Stationary policies and a two-step look-ahead rule.*
As shown in Sect. 3.6.2, the optimal policy in an infinite-horizon discounted cost problem is a stationary policy, which is a time-invariant function of the state at every time $t$. In the context of model-based Phase I cancer trials, the state is the posterior distribution of the parameter $\boldsymbol{\theta} = (\alpha, \beta)$ in the logistic regression model (2.28). Note that the dose $x_n$ for the $n$th patient in CRM or EWOC depends only on the posterior distribution $\Pi_{n-1}$, that is, $x_n$ is a functional $f(\Pi_{n-1})$ of $\Pi_{n-1}$. This functional defines $\{\Pi_k : k \geq 0\}$ as a Markov chain whose states are distributions on the parameter space $\Theta$ and whose state transitions are given by the following: Given current state $\Pi$ (which is a prior distribution of $\boldsymbol{\theta}$), let $x = f(\Pi)$ and generate first $\boldsymbol{\theta}$ from $\Pi$ and then $y \sim \text{Bern}(F_{\boldsymbol{\theta}}(x))$. The new state is the posterior distribution of $\boldsymbol{\theta}$ given $(x, y)$. The functional $x = f(\Pi)$ for CRM is $E_\Pi(\eta)$, which minimizes the expected squared error loss $E_\Pi[(\eta - x)^2]$. EWOC with feasibility bound $\omega$ uses the functional $x = x(\Pi)$ that minimizes the asymmetric loss function $E_\Pi[\ell(\eta, x)]$, where

$$\ell(\eta, x) = \begin{cases} \omega(\eta - x) & \text{if } x \leq \eta, \\ (1 - \omega)(x - \eta) & \text{if } x \geq \eta. \end{cases}$$

For the problem of minimizing the global risk (3.61), the optimal doses $x_i$ depend on $n-i$, where the horizon $n$ is the sample size of the trial, and therefore are not of the form $x_i = f(\Pi_{i-1})$. In terms of "individual" and "collective" ethics, note that (3.61) measures the individual effect of the dose $x_k$ on the $k$th patient through $h(\eta, x_k)$ and its collective effect on future patients through $\sum_{i>k} h(\eta, x_i) + g(\hat{\eta}_n, \eta)$. Bartroff and Lai (2011) note that by using a discounted infinite-horizon version of (3.61), one can still have solutions of the form $x_i = f(\Pi_{i-1})$ for some functional $f$ that only depends on $\Pi_{i-1}$. Specifically, take a discount factor $0 < \delta < 1$ and replace (3.61) by

$$E_{\Pi_0}\left[\sum_{i=1}^{\infty} h(\eta, x_i)\delta^{i-1}\right] \tag{3.78}$$

as the definition of global risk. This global risk measures the individual effect of the dose $x_k$ on the $k$th patient through $h(\eta, x_k)$ and its collective effect on future patients through $\sum_{i>k} h(\eta, x_i)\delta^{i-k}$. Hence, the myopic dose $x_k$ that minimizes $E_{\Pi_{k-1}}[h(\eta, x)]$ for treating the $k$th patient has to be perturbed such that it also helps to create a more informative posterior distribution $\Pi_k$ that is used for dosing future patients. Note that (3.78) does not have the term $g(\hat{\eta}_n, \eta)$ appearing in the finite-horizon problem (3.61), but even without this term, the global risk (3.78) still captures the collective effect of the doses, as indicated above. The dynamic programming equation (3.24) now becomes

$$V(\Pi) = \inf_x E_\Pi\left\{h(\eta, x) + \delta E_\Pi V(\Pi_{+\{x\}})\right\}, \tag{3.79}$$

where $\Pi_{+\{x\}}$ is the new posterior distribution of $\theta$ after $(x, y)$ is observed, with $y \sim \text{Bern}(F_{\boldsymbol{\theta}}(x))$ and $\boldsymbol{\theta} \sim \Pi$, by the Bayesian updating scheme.

The main complexity of the infinite-horizon problem is that the dose $x$ for the next patient involves also consideration for future patients who will receive optimal doses themselves; these future doses depend on the future posterior distributions. To reduce the complexity, Bartroff and Lai (2011) consider two (instead of infinitely many) future patients. This amounts to choosing the next dose $x$ to minimize $E_\Pi \ell(\eta, x; \Pi)$ when the current distribution of $\theta$ is $\Pi$, where

$$\ell(\eta, x; \Pi) = h(\eta, x) + \lambda E_\Pi\left\{E_{\Pi}[h(\eta', x')|x_1 = x, y_1]\right\}, \tag{3.80}$$

in which $\eta' = F_{\boldsymbol{\theta}'}^{-1}(p)$ with $\boldsymbol{\theta}' \sim \Pi'$, and $\Pi'$ and $x'$ are defined below. The first summand in (3.80) measures the (toxicity) effect of the dose $x$ on the patient receiving it. The second summand considers the patient who follows and receives a myopic dose $x'$ that minimizes the patient's posterior loss; the myopic dose is optimal because there are no more patients involved in (3.80). The effect of $x$ on this second patient is through the posterior distribution $\Pi'$ that updates $\Pi$ after observing $(x_1, y_1)$, with $x_1 = x$. Since $y_1$ is not yet observed, the expectation outside the curly brackets is taken over $y_1 \sim \text{Bern}(F_{\boldsymbol{\theta}}(x))$, with $\boldsymbol{\theta} \sim \Pi$. Unlike

$0 < \delta < 1$ in the discounted infinite-horizon problem, the choice of $\lambda > 0$ in (3.80) can exceed 1 and reflects the balance between the collective ethics in generating information for future patients and the individual ethics for the patient receiving the dose. Even though a single patient is used to represent all patients following the one receiving the next dose, because the posterior distributions also change successively, the doses are functionals of these posterior distributions. Bartroff and Lai (2011) describe how to compute $E_{\Pi}\ell(\eta, x; \Pi)$ and its minimizer. They also provide a simulation study in the same setting as that in Example 3.2, showing that the proposed method performs better than EWOC and CRM.

4. *Dynamic programming to minimize global risk over discrete dose set.*
If one follows the traditional practice of choosing a finite set of doses before the trial and defines the MTD by (2.41), then dynamic programming, which is described in (3.62) and (3.63), can be carried out to minimize the global risk (3.61) when $n$ and the size of the dose set are relatively small. Azriel (2012) has done this to compare with various myopic model-based designs. His numerical results show only small improvement (within 3%) over the myopic designs, unlike those in Tables 3.1 and 3.2 for continuous doses. This suggests that fixing a small dose set in advance may severely limit learning and exploration even with a model-based approach and that it may be much better to determine the doses sequentially from the full dose range, at least for the second stage of the two-stage modification, described in Sect. 3.8.4, of the hybrid design.

# Chapter 4
# Group Sequential Design of Phase II and III Trials

In standard (nonsequential) designs of Phase II or III clinical trials, the sample size is determined by the power at a given alternative. In practice, especially for new treatments about which there is little information concerning the magnitude of the treatment effect before actual data are collected, it is often difficult for investigators to specify a realistic alternative at which sample size determination can be based. On the other hand, economic considerations related to funding and duration for the trial and administrative considerations related to other trials that compete for patients and investigators impose constraints on the sample size. A sequential design that can "self-tune" its sample size to the increasing information on the unknown parameters during the course of the trial, under prespecified constraints on the maximum sample size and type I error probability, can be used to resolve the difficulty in calculating a realistic sample size at the design stage for the trial. Unlike fixed sample size (FSS) trials which unblind the randomization and analyze the data after trial termination, fully sequential designs involve continuous examination of the data as they accumulate and often encounter administrative difficulties. A compromise between fully sequential and FSS designs is a group sequential design involving interim analyses of the data. As many later-phase trials have Data and Safety Monitoring Committees (DSMC) who conduct periodic reviews of the trial, interim analyses can be conveniently carried out by the DSMC.

As noted in Sect. 1.2, there has been steady growth of the methodology and applications of group sequential designs since the 1980s. The monographs by Jennison and Turnbull (2000) and Proschan et al. (2006) give an overview of the literature and describe a variety of group sequential methods, which are first developed for the "prototypical" problem of testing a normal mean when the variance is known and then are extended to more complicated problems by appealing to multivariate central limit theorems. We give a summary of these methods in Sect. 4.1 dealing with the prototypical problem. Section 4.2 describes a somewhat different approach, introduced by Lai and Shih (2004), that modifies the relatively complete theory of fully sequential tests of composite hypotheses for the group sequential setting, thereby deriving a class of flexible and efficient group sequential designs that can

J. Bartroff et al., *Sequential Experimentation in Clinical Trials,*
Springer Series in Statistics 298, DOI 10.1007/978-1-4614-6114-2_4,

self-tune to the unknown parameters. Section 4.3 describes implementation of group sequential tests, and Sect. 4.4 summarizes the simulation studies of Lai and Shih (2004) comparing different group sequential tests.

## 4.1 Group Sequential Tests for a Normal Mean

### *4.1.1 The Pocock, O'Brien–Fleming, and Wang–Tsiatis Group Sequential Boundaries*

Section 1.2 has introduced the Pocock and O'Brien–Fleming group sequential tests of $H_0 : \theta = 0$ for the mean $\theta$ of i.i.d. normal $X_i$ with known variance $\sigma^2$; see (1.2) and (1.3) which assume equal group sizes. Let $m$ be the group size so that $M = km$, where $k$ is the number of groups, and let $n_i = im$ be the total sample size at the $i$th interim analysis. Let $S_n = X_1 + \cdots + X_n$. As in FSS tests, the maximum sample size $M$ is so chosen that the test has power $1 - \beta$ at prespecified alternative $\theta_1$ or $-\theta_1$. Wang and Tsiatis (1987) introduced a more general class of tests that stop and reject $H_0$ as soon as

$$\left|\frac{S_{n_i}}{\sqrt{n_i}}\right| \geq \sigma b \left(\frac{i}{k}\right)^{\delta - \frac{1}{2}}, \tag{4.1}$$

where $0 \leq \delta \leq 0.7$. The special case $\delta = \frac{1}{2}$ corresponds to Pocock's test with the square-root boundary $|S_{n_i}| \geq b\sigma\sqrt{n_i}$, while the case $\delta = 0$ corresponds to the test of O'Brien and Fleming with horizontal boundary $|S_{n_i}| \geq b\sigma\sqrt{M}$, noting that $n_i = im$. Wang and Tsiatis (1987) recommended choosing $\delta$ to minimize the expected sample size $E_{\theta_1}(T)(= E_{-\theta_1}(T))$ in this class of tests. The minimization can be carried out by a grid search.

### *4.1.2 One-Sided Group Sequential Tests: Power Family and Triangular Tests*

The preceding two-sided tests, with symmetric stopping boundaries, reject $H_0 : \theta = 0$ if $|S_{n_i}|$ exceeds some threshold that depends on $i$. To test the one-sided hypothesis $H_0' : \theta \leq \theta_0$, group sequential designs have been developed to stop not only when $S_{n_i}$ exceeds an upper boundary (leading to rejection of $H_0$) but also when $S_{n_i}$ falls below a lower boundary (suggesting "futility" in continuing for eventual evidence against $H_0$). The futility boundary can be determined by considering an alternative $\theta_1 > \theta_0$. By shifting the origin, we shall assume without loss of generality that $\theta_0 = -\theta_1$.

Emerson and Fleming (1989) and Pampallona and Tsiatis (1994) have proposed the following *power family* of group sequential tests: stop sampling at stage $i \leq k-1$ if $S_{n_i} + \theta_1 n_i \geq b_i \sigma$, rejecting $H_0$, or if $S_{n_i} + \theta_1 n_i \leq a_i \sigma$, accepting $H_0$. If stopping does not occur before stage $k$, reject $H_0$ only when $S_{n_k} + \theta_1 n_k \geq b_k \sigma$. The lower and upper boundaries involve a parameter $0 \leq \delta \leq \frac{1}{2}$ and are defined by

$$b_i = c_1(\delta) i^{\delta} m^{1/2}, \qquad a_i = \left\{ 2i\theta_1/\sigma - c_2(\delta) i^{\delta} \right\} m^{1/2}, \tag{4.2}$$

where $c_1(\delta)$, $c_2(\delta)$, and $m$ are chosen to yield the prescribed error probabilities at $\theta_0$ and $\theta_1$ and to yield $a_k = b_k$. The case $\delta = 0$ corresponds to a modified version, incorporating a futility boundary, of the one-sided O'Brien–Fleming test, and $\delta = \frac{1}{2}$ corresponds to that of the one-sided Pocock test.

Letting $M$ be the smallest $mk$, that is, integral multiple of $k$, such that

$$mk \geq \left( \frac{\sigma}{\theta_1} \right)^2 \left[ \left\{ \frac{(0.583)^2}{k} + 2 \log \left( \frac{1}{2\alpha} \right) \right\}^{1/2} - \frac{0.583}{k^{1/2}} \right]^2, \tag{4.3}$$

Whitehead and Stratton (1983) proposed to stop at stage $i \leq k-1$ if $|S_{n_i}| \geq b_i \sigma$, where

$$b_i = \left( \frac{\sigma}{\theta_1} \right) \log \left( \frac{1}{2\alpha} \right) - 0.583 m^{1/2} - \frac{im\theta_1}{2\sigma}.$$

When stopping occurs at stage $i$ $(1 \leq i \leq k)$, reject $H_0$ if $S_{n_i} > 0$. These "triangular" tests are in fact a special case of Lorden's 2-SPRT described in Chap. 3; the term 0.583 in (4.3) arises from a diffusion approximation to the error probabilities.

The power family of tests can be easily extended to two-sided tests of $H_0 : \theta{=}0$, allowing early stopping not only to reject $H_0$ (demonstrating the treatment's efficacy) but also to accept $H_0$ (due to futility of continuing). Specifically, stop at stage $i \leq k-1$ if $|S_{n_i}| \geq b_i \sigma$, rejecting $H_0$, or if $|S_{n_i}| \leq a_i \sigma$, accepting $H_0$. If stopping does not occur before stage $k$, reject $H_0$ if $|S_{n_k}| \geq b_k \sigma$, recalling that $a_k = b_k$. The triangular test has also been extended in this way by Whitehead and Stratton (1983) as follows. Define $M$ as the smallest $mk$ satisfying (4.3) with $\theta_1$ replaced by $\tilde{\theta}_1/2$, where

$$\tilde{\theta}_1 = \frac{2\Phi^{-1}(1-\alpha/2)}{\Phi^{-1}(1-\alpha/2) + \Phi^{-1}(1-\beta)} \, \theta_1.$$

The stopping rule is the same as that of the power family, except that we now define

$$\begin{aligned} b_i &= \left( \sigma/\tilde{\theta}_1 \right) \log(1/\alpha) - 0.583 m^{1/2} + \left( \tilde{\theta}_1/4\sigma \right) n_i \\ a_i &= -\left( \sigma/\tilde{\theta}_1 \right) \log(1/\alpha) + 0.583 m^{1/2} + \left( 3\tilde{\theta}_1/4\sigma \right) n_i. \end{aligned} \tag{4.4}$$

### 4.1.3 The Lan–DeMets Error Spending Approach

The assumption of equal group sizes is too restrictive in practice since clinical trial protocols usually specify the calendar times of interim monitoring, for which the number $n_i$ of subjects available for the $i$th interim analysis is unknown in advance and the group sizes $n_i - n_{i-1}$ may be quite uneven. To address this difficulty, Lan and DeMets (1983) note that $(S_n/\sqrt{\sigma^2 M},\ 1 \le n \le M)$ has the same distribution as $(W(t),\ t \in \{1/M, \dots, 1\})$. Therefore, given any stopping rule $\tau$ associated with a sequential test of the drift of continuous-time Wiener process, one can obtain a corresponding stopping rule for testing the common mean $\theta$ of the $X_i$. In particular, for the null hypothesis that $W(t)$ has zero drift (corresponding to $\theta = 0$), Lan and DeMets (1983) regard $\pi(t) := P_0(\tau \le t)$ for $t < 1$ as the type I error spent, up to time $t$, in early stopping to reject the null hypothesis. Given an *error spending function* $\pi(t)$, they propose to transform it to stopping boundaries for $S_{n_i}$ via

$$P_0\left\{|S_{n_i}| \ge a_{n_i},\ \left|S_{n_j}\right| < a_{n_j} \text{ for } 1 \le j < i\right\} = \pi(n_i/M) - \pi(n_{i-1}/M) \qquad (4.5)$$

for $1 \le i < k-1$; the right-hand side of (4.5) can be regarded as how the type I error is spent at the $i$th interim analysis. Letting $\pi(1) = \alpha$, they extend (4.5) to cover the case $i = k$ as well, which means spending whatever is left in the $k$th analysis so that the overall type I error is $\alpha$. Kim and DeMets (1987) and Jennison and Turnbull (1990) suggest to use error spending functions of the form $\pi(t) = \alpha \min(t^\rho, 1)$, with $\rho > 0$.

## 4.2 Group Sequential Generalized Likelihood Ratio Tests with Modified Haybittle–Peto Boundaries

The Lan–DeMets approach provides a flexible method to modify a continuous-time stopping rule for the group sequential setting. In view of the relatively complete theory of fully sequential tests described in Chap. 3, a more direct approach is to modify this theory for the group sequential setting, taking into consideration its two distinguishing features, namely, maximum sample size and uneven group sizes. This approach was used by Lai and Shih (2004) who developed a corresponding theory for group sequential tests, first in a one-parameter exponential family

$$f_\theta(x) = \exp(\theta x - \psi(\theta)) \qquad (4.6)$$

of densities, with respect to some measure on the real line, and then in multiparameter exponential families and more general situations. Although the normal family with known variance 1 and unknown mean $\theta$, which is a special case of (4.6) with $\psi(\theta) = \theta^2/2$, is usually chosen to be the prototype in the group sequential literature, Lai and Shih (2004) consider the more general model (4.6) because linearity of $\psi'$ in the normal case obscures the general form of nearly optimal test statistics and stopping boundaries. Moreover, (4.6) includes the binomial, Poisson, and other commonly used parametric families as special cases, and its generalization including covariates leads to generalized linear models and GLMM; see (2.37) and (2.38).

### *4.2.1 Maximum Sample Size and the Alternative Implied by It*

To test the one-sided hypothesis $H_0 : \theta \leq \theta_0$ in the exponential family (4.6), suppose the significance level is $\alpha$ and no more than $M$ observations are to be taken because of funding and administrative constraints on the trial. The FSS test that rejects $H_0$ if $S_M \geq c_\alpha$ has maximal power at any alternative $\theta > \theta_0$. Although funding and administrative considerations often play an important role in the choice of $M$, justification of this choice in clinical trial protocols is typically based on some prescribed power $1-\beta$ at an alternative $\theta(M)$ "implied" by $M$. The implied alternative is defined by that $M$ and can be derived from the prescribed power $1-\beta$ at $\theta(M)$. Lai and Shih (2004) use this implied alternative to construct the futility boundary in the group sequential test described below.

### *4.2.2 Group Sequential GLR Tests with Nearly Optimal Power and Expected Sample Size*

As pointed out in the introductory paragraph of this chapter, one often does not have much information, prior to a clinical trial, on which to guide the choice of a realistic alternative for determining the sample size of the trial. Under the resource constraint of $M$ on the sample size, it is desirable to adapt to the information on the actual $\theta$ gathered during the course of the trial, allowing early stopping at times of interim analysis, so that the test has nearly optimal power and expected sample size properties. To achieve these goals in a group sequential test with $k$ groups and group sizes $n_1, n_2 - n_1, \ldots,$ $n_k - n_{k-1}$ so that $n_k = M$, we use a rejection region of the form

$$S_{n_k} \geq c \tag{4.7}$$

at the $k$th analysis, where $c > c_\alpha$ but $c$ does not differ much from $c_\alpha$. As in the Lan–DeMets error spending approach, the group sizes can be uneven and need not be known in advance.

Clearly efficiency of a group sequential test depends not only on the choice of the stopping rule but also on the test statistics used. For the prototypical normal family with known variance $\sigma^2$, $S_n^2/(2\sigma^2 n)$ is the logarithm of the generalized likelihood ratio (GLR) statistic for testing $H_0 : \theta = 0$, and the sample mean $\bar{X}_n = S_n/n$ is the maximum likelihood estimate (MLE) of $\theta$. The GLR statistic for testing $\theta$ in the exponential family (4.6) is $nI(\hat{\theta}_n, \theta)$, where $\hat{\theta}_n$ is the MLE of $\theta$ given by $\psi'(\hat{\theta}_n) = \bar{X}_n$, and $I(\gamma, \theta)$ is the Kullback–Leibler information number

$$I(\gamma,\theta) = E_\gamma\left[\log\left\{f_\gamma(X_i)/f_\theta(X_i)\right\}\right] = (\gamma-\theta)\psi'(\theta) - \{\psi(\gamma)-\psi(\theta)\}. \tag{4.8}$$

The theory of fully sequential tests in Chap. 3 shows that to test $H_0 : \theta \le \theta_0$ versus $H_1 : \theta \ge \theta(M)$, a sequential GLR test with stopping rule of the form (3.12) asymptotically minimizes the expected sample size at every $\theta$. When there is a sample size constraint that imposes an upper bound $M$ and a lower bound $m$, which is some fraction of $M$, the stopping boundaries in (3.12) can be chosen to be time-invariant (i.e., not varying with $n$), as noted in (3.14). This suggests the following group sequential test in Lai and Shih (2004, p. 511): for the $i$th interim analysis with $1 \le i \le k-1$, stop the trial if

$$\hat{\theta}_{n_i} > \theta_0, \quad n_i I(\hat{\theta}_{n_i}, \theta_0) \ge b, \tag{4.9}$$

or

$$\hat{\theta}_{n_i} < \theta(M), \quad n_i I\{\hat{\theta}_{n_i}, \theta(M)\} \ge \tilde{b}, \tag{4.10}$$

for $1 \le i \le k-1$. If (4.9) holds, reject $H_0$ upon stopping. If stopping occurs with (4.10), accept $H_0$. In case stopping does not occur in the first $k-1$ analyses, reject $H_0$ if $S_{n_k} \ge c$, as in (4.7). The thresholds $b$, $\tilde{b}$, and $c$ are so chosen that $P_{\theta_0}$ (test rejects $H_0$) $= \alpha$ and the power of the test at $\theta(M)$ does not differ much from its upper bound $1-\beta$.

To determine $b$, $\tilde{b}$, and $c$ satisfying these properties, Lai and Shih (2004) choose $0 < \varepsilon < \frac{1}{2}$ and define $\tilde{b}$ by the equation

$$P_{\theta(M)}\left\{\hat{\theta}_{n_i} < \theta(M) \text{ and } n_i I\left(\hat{\theta}_{n_i}, \theta(M)\right) \ge \tilde{b} \text{ for some } 1 \le i \le k-1\right\} = \varepsilon\beta.$$

After determining $\tilde{b}$, define $b$ and then $c$ by the equations

$$\begin{aligned}
&\sum_{j=1}^{k-1} P_{\theta_0}\Big\{\hat{\theta}_{n_j} > \theta_0 \text{ and } n_j I(\hat{\theta}_{n_j}, \theta_0) \ge b,\ n_i I\left(\hat{\theta}_{n_i}, \theta_0\right) 1_{\{\hat{\theta}_{n_i} > \theta_0\}} < b \text{ and} \\
&\qquad n_i I\left(\hat{\theta}_{n_i}, \theta(M)\right) 1_{\{\hat{\theta}_{n_i} < \theta(M)\}} < \tilde{b} \text{ for } i < j\Big\} = \varepsilon\alpha, \\
&P_{\theta_0}\Big\{S_{n_k} \ge c,\ n_i I\left(\hat{\theta}_{n_i}, \theta_0\right) 1_{\{\hat{\theta}_{n_i} > \theta_0\}} < b \text{ and} \\
&\qquad n_i I\left(\hat{\theta}_{n_i}, \theta(M)\right) 1_{\{\hat{\theta}_{n_i} < \theta(M)\}} < \tilde{b} \text{ for } i < k\Big\} = (1-\varepsilon)\alpha.
\end{aligned}$$

Because the threshold at the last analysis differs from that used in the previous $k-1$ analyses, the test is similar in spirit to that of Haybittle's modification of the repeated significance test for a normal mean described in Sect. 1.2. Actually Haybittle (1971) and later Peto et al. (1976) specifically proposed using $a = 3$ in (1.2) and conventional critical values of $c$ for the final test when the number $k$

of interim analyses is small. Whereas this does not ensure that the overall type I error is $\alpha$, the preceding *modified Haybittle–Peto* boundary chooses the thresholds $b$ and $c$ so that $P_{\theta_0}$ (test rejects $H_0$) $= \alpha$ and $P_{\theta(M)}$ (test rejects $H_0$) does not differ much from the power $1-\beta$ of the FSS test at $\theta(M)$.

Like Armitage's repeated significance test (1.2), the Haybittle–Peto test is actually a two-sided test of $H_0' : \theta = 0$, rejecting $H_0'$ if $|S_n|/\sqrt{\sigma^2 n_i}$ (or equivalently, if $S_n^2/(2\sigma^2 n_i)$) is too large. Here, we consider the one-sided hypothesis $H_0 : \theta \le \theta_0$, and the lower boundary that allows early stopping if (4.10) holds is actually a futility boundary. Note that (4.10) basically asserts enough evidence against the implied alternative $\theta(M)$ that the fixed sample test uses to determine the sample size $M$, and therefore leads to curtailing the test for futility in demonstrating the alternative hypothesis. Although smaller alternatives than $\theta(M)$ may still be compatible with the data, the maximum sample size $M$ suggests inadequate power to demonstrate these smaller alternatives. Theorem 3.3 in Chap. 3 shows that the above group sequential GLR test with modified Haybittle–Peto boundaries attains the asymptotically minimal value, at each fixed $\theta$, of the expected sample and also has power at $\theta(M)$ comparable to its upper bound $1-\beta$.

In Sect. 3.5, an asymptotic analog of Hoeffding's lower bound (3.15) is given in Theorem 3.1, and Lorden's 2-SPRT and its asymptotic optimality are established in Theorem 3.2. The ideal choice of $\theta$ in the 2-SPRT is the true parameter value, which is unknown in practice. By making use of Theorem 3.1, Lai and Shih (2004) have shown that the modified Haybittle–Peto test also attains the asymptotic lower bound in Theorem 3.1. Specifically, let $\tilde{\tau}$ denote the sample size of the modified Haybittle–Peto test and $p_\theta$ denote its power at $\theta > \theta_0$. Let $\alpha + \beta \to 0$ such that $\log \alpha \sim \log \beta$. Suppose that the $k$ group sizes satisfy (3.17) with $n_k = M \sim |\log \alpha|/I(\theta^*, \theta_0)$, where $\theta_0 < \theta^* < \theta(M)$ is defined by $I(\theta^*, \theta_0) = I(\theta^*, \theta(M))$. Lai and Shih (2004) have shown that for every fixed $\theta$,

$$E_\theta(\tilde{\tau}) \sim n_\nu + \rho(\theta)(n_{\nu+1} - n_\nu), \tag{4.11}$$

where $\nu$ and $\rho(\theta)$ are the same as in Theorem 3.2 with $\theta_1 = \theta(M)$. They have also shown that $p_{\theta(M)}$ is close to the power of the Neyman–Pearson test whose fixed sample size $M$ is chosen so that it has power $1-\beta$ at $\theta(M)$:

$$p_{\theta(M)} = 1 - \beta - (\kappa_\varepsilon + o(1))\beta, \tag{4.12}$$

where $\kappa_\varepsilon \sim \{1 + (\theta(M) - \theta^*)/(\theta^* - \theta_0)\}\varepsilon$ as $\varepsilon \to 0$.

Because stopping can only occur at a few sample sizes, the optimal stopping rule can be computed numerically by the backward induction algorithm of finite-horizon dynamic programming described in Sect. 3.6.1. For the relatively simple prototypical case of a normal mean $\theta$ when the variance is known, the optimal group sequential boundaries that minimize $\int E_\theta(T)dG(\theta)$ for some normal distribution (possibly degenerate) $G$ on the real line, subject to prescribed error probabilities at $\theta_0$ and $\theta_1$, can be transformed, via Lagrange multiplier, to a Bayes problem that has prior distribution which is a mixture of $G$ and degenerate distributions at $\theta_0$ and

$\theta_1$. Unlike the fully sequential case considered in Sect. 3.7.2, the Bayes problem for this group sequential setting is computationally tractable. This is similar to the global risk minimization problem with a discrete set of doses discussed at the end of Sect. 3.9 versus that with a continuous dose interval in Sect. 3.8. Eales and Jennison (1992, 1995) have carried this out for several choices of $G$. Section 4.4 gives a comparative study of the Eales–Jennison and Lai–Shih approaches.

### 4.2.3 Two-Sided Tests with or Without Futility Boundaries

To test $H_0 : \theta = \theta_0$ based on a sample of fixed size $M$, the GLR test rejects $H_0$ if $MI(\hat{\theta}_M, \theta_0)$ exceeds some threshold $c_\alpha$, where $P_{\theta_0}\{MI(\hat{\theta}_M, \theta_0) \geq c_\alpha\} = \alpha$. This test has power $1-\beta$ at the implied alternatives $\theta_+(M) > \theta_0$ and $\theta_-(M) < \theta_0$. A group sequential test, with $k$ groups and group sizes $n_1, n_2 - n_1, \ldots, n_k - n_{k-1}$ so that $n_k = M$, having power near $1-\beta$ at $\theta_+(M)$ and $\theta_-(M)$ and asymptotically optimal expected sample size at every given $\theta$, can be constructed by extending the ideas of the preceding section, as shown by Lai and Shih (2004, p. 512). The rejection region at the $k$th analysis has the form $n_k I(\hat{\theta}_{n_k}, \theta_0) \geq c$, and the stopping region during the first $k-1$ interim analyses has the form

$$n_i I\left(\hat{\theta}_{n_i}, \theta_0\right) \geq b \tag{4.13}$$

or

$$n_i I\left(\hat{\theta}_{n_i}, \theta_0\right) < b, \quad n_i I\left(\hat{\theta}_{n_i}, \theta_-(M)\right) \geq \tilde{b}_- \quad \text{and} \quad n_i I\left(\hat{\theta}_{n_i}, \theta_+(M)\right) \geq \tilde{b}_+, \tag{4.14}$$

where $c$, $b$, $\tilde{b}_-$, and $\tilde{b}_+$ will be specified below. If (4.13) occurs, reject $H_0$ upon stopping. If stopping occurs with (4.14), accept $H_0$.

As in the one-sided case of Sect. 4.2.2, let $0 < \varepsilon < \frac{1}{2}$ and define $\tilde{b}_+$, $\tilde{b}_-$ by

$$P_{\theta_+(M)}\left\{n_i I\left(\hat{\theta}_{n_i}, \theta_+(M)\right) \geq \tilde{b}_+ \text{ for some } 1 \leq i \leq k-1\right\} = \varepsilon\beta,$$
$$P_{\theta_-(M)}\left\{n_i I\left(\hat{\theta}_{n_i}, \theta_-(M)\right) \geq \tilde{b}_- \text{ for some } 1 \leq i \leq k-1\right\} = \varepsilon\beta.$$

After determining $\tilde{b}_+$ and $\tilde{b}_-$, define $b$ and then $c$ by the equations

$$\sum_{j=1}^{k-1} P_{\theta_0}\Big\{n_j I\left(\hat{\theta}_{n_j}, \theta_0\right) \geq b,\ n_i I\left(\hat{\theta}_{n_i}, \theta_0\right) < b \text{ and}$$
$$n_i I\left(\hat{\theta}_{n_i}, \theta_+(M)\right) 1_{\{n_i I(\hat{\theta}_{n_i}, \theta_-(M)) \geq \tilde{b}_-\}} < \tilde{b}_+ \text{ for } i < j\Big\} = \varepsilon\alpha,$$
$$P_{\theta_0}\Big\{n_k I(\hat{\theta}_{n_k}, \theta_0) \geq c,\ n_i I\left(\hat{\theta}_{n_i}, \theta_0\right) < b \text{ and}$$
$$n_i I\left(\hat{\theta}_{n_i}, \theta_+(M)\right) 1_{\{n_i I(\hat{\theta}_{n_i}, \theta_-(M)) \geq \tilde{b}_-\}} < \tilde{b}_+ \text{ for } i < k\Big\} = (1-\varepsilon)\alpha.$$

The inner wedge of the stopping region defined by (4.14) is targeted toward early stopping under the null hypothesis and nearby alternatives. Similar inner wedges have been proposed for normal $X_i$ by Whitehead and Stratton (1983) and Pampallona and Tsiatis (1994); see the last paragraph of Sect. 4.1.2. On the other hand, the Pocock, O'Brien–Fleming, and Wang–Tsiatis group sequential tests of $H_0 : \theta = 0$ in Sect. 4.1.1 do not have inner futility boundaries. If we remove the futility boundary (4.14) from the stopping rule, then during the first $k-1$ analyses, we stop and reject $H_0$ as soon as (4.13) holds. In this case, the threshold $b$ in (4.13) and the critical value $c$ for $n_k I(\hat{\theta}_{n_k}, \theta_0)$ are given by

$$\sum_{j=1}^{k-1} P_{\theta_0}\left\{n_j I\left(\hat{\theta}_{n_j}, \theta_0\right) \geq b \text{ and } n_i I\left(\hat{\theta}_{n_i}, \theta_0\right) < b \text{ for } i < j\right\} = \varepsilon\alpha,$$

$$P_{\theta_0}\left\{n_k I\left(\hat{\theta}_{n_k}, \theta_0\right) \geq c \text{ and } n_i I\left(\hat{\theta}_{n_i}, \theta_0\right) < b \text{ for } i < k\right\} = (1-\varepsilon)\alpha.$$

### *4.2.4 Extensions to Multiparameter and Multiarmed Problems*

The group sequential tests for the univariate exponential family can be readily extended to the multiparameter and multiarm settings, as shown by Lai and Shih (2004, pp. 513–514) in the context of the multiparameter exponential family of densities $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \exp(\boldsymbol{\theta}^T \boldsymbol{x} - \psi(\boldsymbol{\theta}))$ with respect to some measure $\nu$ on $\mathbb{R}^d$, in which $\boldsymbol{\theta}$ and $\boldsymbol{x}$ are $d \times 1$ vectors belonging to $\mathbb{R}^d$. To test the null hypothesis $H_0 : u(\boldsymbol{\theta}) = u_0$, where $u$ is a continuously differentiable real-valued function on the natural parameter space $\Theta := \{\boldsymbol{\theta} \in \mathbb{R}^d : \int \exp(\boldsymbol{\theta}^T \boldsymbol{x}) d\nu(\boldsymbol{x}) < \infty\}$, the GLR statistic based on $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{n_i}$ is

$$\Lambda_i = n_i \left\{\hat{\boldsymbol{\theta}}_{n_i}^T \bar{\boldsymbol{X}}_{n_i} - \psi\left(\hat{\boldsymbol{\theta}}_{n_i}\right)\right\} - \sup_{u(\boldsymbol{\theta})=u_0} n_i \left\{\boldsymbol{\theta}^T \bar{\boldsymbol{X}}_{n_i} - \psi(\boldsymbol{\theta})\right\} = \inf_{u(\boldsymbol{\theta})=u_0} n_i I\left(\hat{\boldsymbol{\theta}}_{n_i}, \boldsymbol{\theta}\right), \tag{4.15}$$

in which $I(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is given by (4.8) with $\psi'$ denoting the gradient vector $\nabla\psi$ of partial derivatives with respect to the components of $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}_{n_i}$ is the MLE of $\boldsymbol{\theta}$ given by $\nabla\psi(\hat{\boldsymbol{\theta}}_{n_i}) = \bar{\boldsymbol{X}}_{n_i}$. The modified Haybittle–Peto test of $H_0 : u(\boldsymbol{\theta}) = u_0$ without futility boundaries again has a stopping region of the form $\Lambda_i \geq b$ during the first $k-1$ interim analyses, and its rejection region at the $k$th analysis has the form $\Lambda_k \geq c$, where $b$ and $c$ are determined as in the last paragraph of Sect. 4.2.3.

To test the one-sided hypothesis $H_0 : u(\boldsymbol{\theta}) \leq u_0$, suppose $I(\boldsymbol{\gamma}, \boldsymbol{\theta})$ is increasing in $|u(\boldsymbol{\theta}) - u(\boldsymbol{\gamma})|$ for every fixed $\boldsymbol{\gamma}$. Then given the maximum sample size $n_k = M$, we can still define the alternative $u_M$ implied by $M$, that is, $u_M > u_0$ is the alternative at which the fixed sample size GLR test with type I error probability $\alpha$ and sample size $M$ has power $\inf_{\boldsymbol{\theta}:u(\boldsymbol{\theta})=u_M} P_{\boldsymbol{\theta}}\{\text{Reject } H_0\}$ equal to $1-\beta$. Therefore, the group sequential test in Sect. 4.2.2 can be readily extended to test $H_0 : u(\boldsymbol{\theta}) \leq u_0$.

A special case of particular interest in clinical trials involves $I$ independent populations having density functions $\exp\{\theta_i x - \tilde{\psi}(\theta_i)\}$ with respect to some measure on the real line, so that $\boldsymbol{\theta}^T \boldsymbol{x} - \psi(\boldsymbol{\theta}) = \sum\{\theta_i x_i - \tilde{\psi}(\theta_i)\}$. The null hypothesis of equality of population means can be represented by $u(\boldsymbol{\theta}) = 0$, where $u(\boldsymbol{\theta}) = \sum\{\tilde{\psi}'(\theta_i) - \tilde{\psi}'(\theta_1)\}^2$. In multiarmed clinical trials, for which different numbers of patients are assigned to different treatments, the GLR statistic $\tilde{\Lambda}_j$ at the $j$th interim analysis has the form

$$\tilde{\Lambda}_j = \sum_{i=1}^{I} n_{ij} \left\{ \hat{\theta}_{i,n_{ij}} \bar{X}_{i,n_{ij}} - \tilde{\psi}\left(\hat{\theta}_{i,n_{ij}}\right) \right\} - \sup_{u(\theta_1,\ldots,\theta_I)=u_0} \sum_{i=1}^{I} n_{ij} \left\{ \theta_i \bar{X}_{i,n_{ij}} - \tilde{\psi}(\theta_i) \right\}, \tag{4.16}$$

in which $n_{ij}$ is the total number of observations from the $i$th population up to the time of the $j$th interim analysis.

## 4.3 Implementation and Computational Methods

### 4.3.1 *Asymptotic Normality and Recursive Numerical Integration*

First, consider the prototypical model in which the $X_i$ are independent $N(\theta, 1)$, and let $\tau = \min\{i \le k : S_{n_i} \notin (a_i, b_i)\} \wedge k$. Let $f_i(x) = (d/dx) P_\theta\{\tau > i,\ S_{n_i} \le x\}$, and let $\phi$ and $\Phi$ denote the standard normal density and distribution function, respectively. Then $f_1(x) = \phi((x - n_1\theta)/\sqrt{n_1})$ for $a_1 < x < b_1$, and for $i > 1$ and $a_i < x < b_i$,

$$f_i(x) = \int_{a_{i-1}}^{b_{i-1}} f_{i-1}(y) \phi\left(\frac{x - y - \theta(n_i - n_{i-1})}{\sqrt{n_i - n_{i-1}}}\right) dy. \tag{4.17}$$

Moreover,

$$\begin{aligned} P(\tau = i) = \int_{a_{i-1}}^{b_{i-1}} f_{i-1}(y) \Bigg\{ &\Phi\left(\frac{a_i - y - \theta(n_i - n_{i-1})}{\sqrt{n_i - n_{i-1}}}\right) + 1 \\ &- \Phi\left(\frac{b_i - y - \theta(n_i - n_{i-1})}{\sqrt{n_i - n_{i-1}}}\right) \Bigg\} dy. \end{aligned}$$

The recursion is the essence of the recursive numerical integration algorithm of Armitage et al. (1969). It reduces direct multiple integration to evaluate $P(\tau = i)$ by an $i$-fold integral to univariate integrals whose integrand is calculated recursively. Numerical evaluation of the univariate integral involves a quadrature rule that replaces an integral by a weighted sum, and therefore $f_i(x)$ in (4.17) only needs to be computed on a grid of points. Details are given in Sect. 19.2 of Jennison and Turnbull (2000).

A major reason why a normal random walk is used as a prototypical case in the group sequential literature is that the multivariate distribution of many group sequential test statistics has a limiting normal distribution with independent increments; see Sect. 4.5. The adequacy, however, of this normal approximation may be questionable for many test statistics, even for $k = 1$. On the other hand, there is extensive numerical evidence that the signed-root likelihood ratio statistic

$$W_i = \operatorname{sign}\big(u(\hat{\boldsymbol{\theta}}_{n_i}) - u_0\big)\sqrt{2n_i\Lambda_i}, \tag{4.18}$$

where $\Lambda_i$ is defined in (4.15), is approximately normal with mean 0 and variance $n_i$ under $H_0 : u(\boldsymbol{\theta}) = u_0$ and that the increments $W_i - W_{i-1}$ are approximately independent under $H_0$. We can therefore approximate $W_i$ by a sum $S_{n_i}$ of independent standard normal random variables under $H_0$ and thereby determine the thresholds $b$, $\tilde{b}$, and $c$ in the modified Haybittle–Peto test in Sect. 4.2; see Sect. 4.5 for details.

### 4.3.2 Monte Carlo and the Bootstrap for Sequential GLR Statistics

An alternative to numerical quadrature is to use Monte Carlo simulations to evaluate boundary crossing probabilities. Although this is an obvious idea, it is far from being clear which distribution from a composite hypothesis should be chosen to simulate from. Bootstrap theory (see Sect. 7.2.1) provides an answer to this question. An important ingredient in this theory is an *approximate pivot*; a function of the data for which the sampling distribution does not depend on the unknown distribution generating the data is called a "pivot." Since the vector of GLR statistics to test a null hypothesis is an approximate pivot under that hypothesis, we can simulate from the estimated distribution under the assumed hypothesis. Further discussion is given in Sect. 4.5. This asymptotically pivotal property of sequential GLR statistics under the null hypothesis is related to the jointly normal asymptotic distribution of the statistics (4.18). Accordingly, we can simulate the joint distribution of the sequential GLR statistics under the parameter estimated by the constrained MLE satisfying the null hypothesis. The adaptive or group sequential Bayes tests considered in Sect. 1.5, however, do not make use of these frequentist principles and therefore cannot guarantee the prescribed type I error probability, which they try to control by simulating the error probabilities at certain parameter configurations belonging to the null hypothesis.

### 4.3.3 Design, Interim Analysis, and Nonparametric Extensions

At the design stage, in order to calculate the early stopping boundaries $b$ and $\tilde{b}$ for the modified Haybittle–Peto test, one needs to specify the number of interim analyses and the sample size $n_i$ at each interim analysis, considering the univariate one-sided hypothesis in Sect. 4.2.2 to fix the ideas. The actual $n_i$ during the course of the trial may differ substantially from those assumed at the design stage. This does not cause under- or over-spending of the prescribed type I error because the rejection threshold $c$ for the terminal analysis can be specified after the $k-1$ interim analyses, and we can use the *actual* (rather than assume a priori) $n_i$ for $1 \leq i \leq k-1$ to compute $c$. Thus, although the modified Haybittle–Peto test does not have the flexibility of the Lan–DeMets error spending approach that does not have to specify the $n_i$ at the design stage, it can still correct for the misspecification of the previous $n_i$ at the terminal analysis.

In fact, $c$ does not even have to be determined explicitly for the terminal analysis at maximum sample size $M$ if the trial has not stopped earlier. Let $\tilde{\tau}$ denote the sample size of the test, as in Sect. 4.2.2, and let $\boldsymbol{X}_{\tilde{\tau}} = (X_1, \ldots, X_{\tilde{\tau}})$. Let $\boldsymbol{x}_{\text{obs}}$ and $\hat{\theta}_{\text{obs}}$ denote the observed values of $\boldsymbol{X}_{\tilde{\tau}}$ and $\hat{\theta}_{\tilde{\tau}}$, respectively. Checking whether the observed value of $S_M = M\psi'(\hat{\theta}_M)$ exceeds the threshold $c$ if the trial has not stopped prior to $M$ is equivalent to checking

$$P_{\theta_0}\left\{\tilde{\tau} < M \text{ and } \hat{\theta}_{\tilde{\tau}} > \theta_0, \ \text{ or } \tilde{\tau} = M \text{ and } \hat{\theta}_M \geq \hat{\theta}_{\text{obs}}\right\} \leq \alpha, \tag{4.19}$$

in which $\hat{\theta}_{\text{obs}}$ is treated as a nonrandom constant in the probability calculation. In fact, the definition of $c$ can also be restated as

$$P_{\theta_0}\left\{\tilde{\tau} < M \text{ and } \hat{\theta}_{\tilde{\tau}} > \theta_0, \ \text{ or } \tilde{\tau} = M \text{ and } M\psi'(\hat{\theta}_M) \geq c\right\} = \alpha, \tag{4.20}$$

as the event in (4.20) can be decomposed as a union of two disjoint events, one with probability $\varepsilon\alpha$ and the other with probability $(1-\varepsilon)\alpha$. The probability in (4.19), with $\hat{\theta}_{\text{obs}}$ treated as nonrandom, can be computed in the same way as that in (4.20). The advantage of using (4.19) is that we do not have to first solve (4.20) for $c$ and then check $M\psi'(\hat{\theta}_M) \geq c$. Thus, the probability in (4.19) needs only to be computed once, whereas one has to compute that in (4.20) for many choices of $c$ to find the one for which the probability is $\alpha$. For the multiparameter case, the bootstrap method can be used to evaluate the extension of the left-hand side of (4.19) to the composite null hypothesis $u(\boldsymbol{\theta}) = u_0$. The second supplement in Sect. 4.5 provides further details on the bootstrap method.

He et al. (2012) have shown how the modified Haybittle–Peto test in the multiparameter exponential family can be extended to nonparametric group sequential tests of $H_0 : u(F,G) \leq 0$, where $F$ is the distribution function of the outcome of a new treatment and $G$ is that of the standard treatment (or placebo) and $u(F,F) = 0$. Let $n_i'$ be the sample size of the new treatment and $n_i''$ be that of the standard treatment at the $i$th interim analysis so that $n_i = n_i' + n_i''$, and let $X_1, \ldots, X_{n_i'}$ and $Y_1, \ldots, Y_{n_i''}$ be the corresponding outcomes. Let $\hat{F}_{n_i'}$ be the empirical distribution function of $X_1, \ldots, X_{n_i'}$, and $\hat{G}_{n_i''}$ be that of $Y_1, \ldots, Y_{n_i''}$. Commonly used two-sample nonparametric test statistics can be written in the form of a *generalized Chernoff–Savage statistic*

$$T_i = \int_{-\infty}^{\infty} J_i(\hat{F}_{n_i'}(x), \hat{G}_{n_i''}(x))\, d\hat{F}_{n_i'}(x), \tag{4.21}$$

where $J_i : \{0, 1/n_i', 2/n_i', \ldots, 1\} \times \{0, 1/n_i'', 2/n_i'', \ldots, 1\} \to \mathbb{R}$ satisfies

$$\frac{1}{n_i'} \sum_{l=1}^{n_i'} \sup_{y \in \{1/n_i'', \ldots, 1\}} \left| J_i\left(\frac{L}{n_i'}, y\right) - J\left(\frac{L}{n}, y\right) \right| \to 0$$

as $n_i' \to \infty$, and $J : [0,1] \times [0,1] \to \mathbb{R}$ is twice continuously differentiable except possibly at $(0,0)$ and $(1,1)$ and satisfies certain regularity conditions near $(0,0)$ and $(1,1)$. In this case, the function $u(F,G)$ in $H_0 : u(F,G) \le 0$ is given by

$$u(F,G) = \int_{-\infty}^{\infty} J(F(x), G(x))\, dF(x). \tag{4.22}$$

Since subjects are randomly assigned to the new or standard treatment,

$$n_i'/n_i \xrightarrow{p} 1/2, \text{ i.e., } n_i'' = n_i'(1 + o_p(1)). \tag{4.23}$$

Under (4.23), $T_i$ has the representation

$$T_i = u(F,G) + \frac{1}{n_i'} \sum_{l=1}^{n_i'} (\psi(X_1) - E\psi(X_1)) + \frac{1}{n_i''} \sum_{l=1}^{n_i''} (\psi^*(Y_1) - E\psi^*(Y_1)) + R_i, \tag{4.24}$$

where $R_i = o_p(1/\sqrt{n_i'})$ and

$$\psi(x) = J(F(x), G(x)) - \int_0^x \frac{\partial J}{\partial x}(F(t), G(t))dF(t),$$

$$\psi^*(y) = -\int_0^y \frac{\partial J}{\partial y}(F(t), G(t))dF(t); \tag{4.25}$$

see He et al. (2012). Therefore, under equal randomization to the two treatments, $n_i'T_i$ behaves asymptotically like a normal random walk under $H_0$ and under local alternatives, and the problem of testing $H_0 : u(F,G) \le 0$ versus $H_1 : u(F,G) \ge \delta$ can be approximated by that of group sequential testing of $H_0' : \mu \le 0$ versus $H_1' : \mu \ge \delta$ based on i.i.d. normal random variables $Z_1, Z_2, \ldots$ with mean $\mu = u(F,G)$ and variance

$$\sigma^2 = \text{Var}_{F=G}(\psi(X)) + \text{Var}_{F=G}(\psi^*(Y)), \tag{4.26}$$

where $\psi$ and $\psi^*$ are given by (4.25). Note that $F = G$ is the boundary case of $H_0$ and that $F(X)$ and $G(Y)$ are Uniform$(0,1)$ random variables. Under local alternatives, the asymptotic variance of $T_i$ is the same as that under $F = G$, and therefore, the variance formula (4.26) still holds for local alternatives. Therefore, similar to the signed-root likelihood ratio statistic (4.18), the modified Haybittle–Peto test can also be applied to the generalized Chernoff–Savage statistics (4.24); see Sect. 4.5.

## 4.4 Comparison of Group Sequential Tests

Lai and Shih (2004) have carried out extensive simulation studies to compare the performance of different group sequential tests for the prototypical problem of testing the mean $\theta$ of normal $X_i$ with known variance 1. Performance is measured in terms of power and expected sample size at various alternatives, besides the maximum sample size and expected sample size under the null hypothesis, subject to a type I error constraint $\alpha$.

### *4.4.1 One-Sided Tests*

Here, we consider one-sided group sequential tests of $H_0 : \theta \leq \theta_0$ with error probability $\alpha = 0.05$ and $\theta_0 < 0$, and with $k = 5$ groups of equal size $m$ so that the maximum size is $M = km$. The alternative chosen for the tests in Sect. 4.1.2 is taken to be $\theta_1 = |\theta_0|$, which we also choose to be the implied alternative $\theta(M)$, with $\tilde{\alpha} = 0.05$, for the group sequential test with modified Haybittle–Peto boundary in Sect. 4.2.2, denoted by ModHP. Besides these group sequential tests, we also consider the FSS test with sample size $n^*$ and the tests of Eales and Jennison (1992) mentioned in the last paragraph of Sect. 4.2.2. Letting

$$F_1 = E_0(T), \;\; F_2 = E_{\theta_1}(T), \;\; F_3 = E_{2\theta_1}(T), \;\; F_5 = \int E_\theta(T)dG(\theta),$$

where $G$ is normal with mean 0 and standard deviation $\theta_1$, Eales and Jennison (1992) considered the group sequential test $F_i^\dagger$, with five groups of equal size and maximum sample size $tn^*$, that minimizes $F_i$ subject to the prescribed error probabilities at $\theta_0$ and $\theta_1$, $i = 1,2,3,5$.

To compare the results of Eales and Jennison (1992, Table 2) on the values of $F_1/n^*$, $F_2/n^*$, $F_3/n^*$, and $F_5/n^*$ for their tests $F_1^\dagger$, $F_2^\dagger$, $F_3^\dagger$, and $F_5^\dagger$ that uses $t = 1.155$ with the corresponding values of ModHP, we adjust the $\varepsilon$ in Sect. 4.2.2 so that the ModHP has the same error probability $\tilde{\alpha}$ at $\theta_1$ and maximum sample size $M = tn^*$ for $t = 1.155$. Let $F_0 = E_{-\theta_1/2}(T)$ denote the expected sample size at $-\theta_1/2$, which is the midpoint between the null hypothesis $-\theta_1$ and 0, where $F_1$ is computed. Table 4.1 gives the values of $F_0$, $F_1$, $F_2$, $F_3$, and $F_5$, normalized by $n^*$, and the power functions of these tests. Also given in Table 4.1 for comparison are (a) the FSS Neyman–Pearson test, (b) ModHP with $t = 1$, for which $\varepsilon = 1/3$, (c) the triangular test of Whitehead and Stratton (see Sect. 4.1.2), and (d) tests in the power family (4.2) with $\delta = 0$, 0.2, 0.4, and 0.5. The results show that all the group sequential tests have power close to that of the FSS test at the alternatives considered and that the possibility of early termination, due to crossing either the efficacy or futility boundary, in the group sequential tests has led to substantial savings in expected sample size over the FSS test. With 15.5% inflation in maximum sample size over $n^*$, ModHP has expected sample sizes close to those of the optimal

**Table 4.1** Maximum and expected sample sizes and power functions of the test with fixed sample size $n^*$ and one-sided group sequential tests with maximum sample size $tn^*$

| Power (%) at $\eta\theta_1$ | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | $\eta$ | | | |
| Test | $t$ | $F_0/n^*$ | $F_1/n^*$ | $F_2/n^*$ | $F_3/n^*$ | $F_5/n^*$ | 0.154 | 0.512 | 0.779 | 1.000 |
| FSS | 1.0 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 60.0 | 80.0 | 90.0 | 95.0 |
| ModHP$_{\varepsilon=1/3}$ | 1.0 | 0.762 | 0.814 | 0.636 | 0.397 | 0.678 | 57.7 | 78.0 | 88.4 | 93.9 |
| Triangular | 1.336 | 0.728 | 0.783 | 0.612 | 0.420 | 0.656 | 60.1 | 80.3 | 90.1 | 94.9 |
| $F_1^\dagger$ | 1.155 | 0.736 | 0.792 | 0.616 | 0.419 | 0.662 | 60.0 | 80.2 | 90.1 | 95.0 |
| $F_2^\dagger$ | 1.155 | 0.740 | 0.801 | 0.607 | 0.391 | 0.658 | 60.0 | 80.3 | 90.1 | 95.0 |
| $F_3^\dagger$ | 1.155 | 0.777 | 0.845 | 0.627 | 0.378 | 0.683 | 60.1 | 80.4 | 90.2 | 95.0 |
| $F_5^\dagger$ | 1.155 | 0.737 | 0.796 | 0.607 | 0.396 | 0.657 | 60.0 | 80.3 | 90.1 | 95.0 |
| ModHP | 1.155 | 0.756 | 0.822 | 0.613 | 0.381 | 0.668 | 60.1 | 80.4 | 90.2 | 95.0 |
| PF$_{\delta=0.0}$ | 1.085 | 0.772 | 0.818 | 0.669 | 0.494 | 0.707 | 60.0 | 80.0 | 90.0 | 95.0 |
| PF$_{\delta=0.2}$ | 1.169 | 0.741 | 0.794 | 0.628 | 0.435 | 0.670 | 60.0 | 80.2 | 90.1 | 95.0 |
| PF$_{\delta=0.4}$ | 1.380 | 0.727 | 0.787 | 0.598 | 0.397 | 0.649 | 60.2 | 80.5 | 90.3 | 95.0 |
| PF$_{\delta=0.5}$ | 1.577 | 0.735 | 0.802 | 0.598 | 0.401 | 0.655 | 60.4 | 80.7 | 90.3 | 95.0 |

values given by $F_j^\dagger$. In comparison, the triangular test requires 33.6% inflation of the maximum sample size, while tests in the power family with $\delta = 0.4$, 0.5 require 38.0% and 57.7% inflation of the maximum sample size. Note that ModHP with $\varepsilon = 1/3$ requires no inflation in maximum sample size, that is $t = 1$, and has 94% power at $\theta_1$. Whereas the theory in (4.13) and (4.14) allows a lot of latitude in the choice of $\varepsilon$ to define modified Haybittle–Peto tests that are asymptotically optimal as $\alpha + \tilde{\alpha} \to 0$, there is a trade-off between power and expected sample size in choosing $\varepsilon$ for finite sample situations. Choosing small $\varepsilon$ increases power but decreases the chance of early stopping, making the test more like the FSS test. Simulation studies have shown that choosing $\varepsilon$ between $1/3$ and $1/2$ for ModHP strikes a good balance between power and expected sample size in practice.

### 4.4.2 Two-Sided Tests Without Futility Boundaries

Eales and Jennison (1995) developed tests that minimize $E_{\theta_1}(T)$ subject to error probabilities not exceeding $\alpha$ and $\tilde{\alpha}$ at 0 and $\theta_1$, respectively, assuming a maximum sample size of $tn^*$, where $t > 1$ and $n^*$ is the FSS of the uniformly most powerful invariant test having these error probabilities. Besides the $F_2^\dagger$ that minimizes $F_2 = E_{\theta_1}(T)$, they also consider the test $F_5^\dagger$ that minimizes $F_5 = \int E_\theta(T)dG(\theta)$, where $G$ is normal with mean 0 and standard deviation $\theta_1$, similarly to the corresponding one-sided tests in the preceding section. Table III of Eales and Jennison (1995) reports the values of $F_2/n^*$ and $F_5/n^*$ for their tests $F_2^\dagger$ and $F_5^\dagger$ in the case $k = 5$. These values, normalized by $n^*$, do not depend on $\theta_1$. Here, $\alpha = \tilde{\alpha} = 0.05$ and $t = 1.05$ or 1.10. To compare with corresponding values in the modified Haybittle–Peto test,

**Table 4.2** Comparison of various two-sided tests without futility boundaries, with $n^*$ being the fixed sample size that yields 95% power at $\theta_1$

| Power (%) at $\eta\theta_1$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | $\eta =$ | | | | | |
| Test | $t$ | $F_2/n^*$ | $F_5/n^*$ | 0.614 | 0.689 | 0.777 | 0.831 | 0.899 | 1.000 |
| $F_2^\dagger$ | 1.05 | 0.634 | 0.765 | 59.2 | 69.5 | 79.6 | 84.8 | 89.9 | 95.0 |
| $F_5^\dagger$ | 1.05 | 0.635 | 0.764 | 59.2 | 69.4 | 79.6 | 84.8 | 89.9 | 95.0 |
| $\text{WT}_{\delta=0.195}$ | 1.05 | 0.652 | 0.778 | 59.3 | 69.5 | 79.7 | 84.8 | 89.9 | 95.0 |
| ModHP | 1.05 | 0.646 | 0.769 | 59.2 | 69.3 | 79.7 | 84.8 | 89.9 | 95.0 |
| $F_2^\dagger$ | 1.1 | 0.611 | 0.771 | 58.4 | 68.8 | 79.3 | 84.5 | 89.8 | 95.0 |
| $F_5^\dagger$ | 1.1 | 0.612 | 0.770 | 58.4 | 68.8 | 79.3 | 84.5 | 89.8 | 95.0 |
| $\text{WT}_{\delta=0.195}$ | 1.1 | 0.617 | 0.774 | 58.6 | 68.9 | 79.3 | 84.6 | 89.8 | 95.0 |
| ModHP | 1.1 | 0.616 | 0.772 | 58.5 | 68.8 | 79.3 | 84.6 | 89.8 | 95.0 |
| OBF | 1.02 | 0.694 | 0.795 | 59.5 | 69.6 | 79.7 | 84.7 | 89.9 | 95.0 |
| Pocock | 1.19 | 0.603 | 0.800 | 57.1 | 67.7 | 78.7 | 84.1 | 89.6 | 95.0 |
| HP | 1.02 | 0.723 | 0.794 | 59.8 | 69.8 | 80.0 | 85.0 | 90.0 | 95.0 |
| $\text{WT}_{\delta=0.25}$ | 1.06 | 0.640 | 0.776 | 59.3 | 69.5 | 79.7 | 84.8 | 90.0 | 95.0 |
| $\text{ModHP}_{\varepsilon=1/3}$ | 1.04 | 0.668 | 0.774 | 59.5 | 69.6 | 79.8 | 84.9 | 90.0 | 95.0 |
| OBF | 1.0 | 0.685 | 0.782 | 58.6 | 68.7 | 78.9 | 84.0 | 89.2 | 94.5 |
| Pocock | 1.0 | 0.561 | 0.698 | 49.6 | 59.8 | 70.9 | 76.9 | 83.5 | 90.7 |
| HP | 1.0 | 0.718 | 0.784 | 59.2 | 69.2 | 79.4 | 84.5 | 89.6 | 94.7 |
| $\text{WT}_{\delta=0.25}$ | 1.0 | 0.619 | 0.737 | 56.6 | 66.7 | 77.1 | 82.5 | 88.0 | 93.7 |
| $\text{ModHP}_{\varepsilon=1/3}$ | 1.0 | 0.657 | 0.753 | 58.0 | 68.1 | 78.4 | 83.7 | 89.0 | 94.3 |

we adjust the $\varepsilon$ in the last paragraph of Sect. 4.2.3 so that the test has the same error probability $\tilde{\alpha}$ at $\theta_1$ and maximum sample size $tn^*$. The results are given in Table 4.2. Also given for comparison are the values of $F_2/n^*$ for the Pocock, O'Brien–Fleming (OBF), Haybittle–Peto (HP), and Wang–Tsiatis (WT) tests described in Sect. 4.1.1. The maximum sample sizes of these tests have values other than $1.05n^*$ and $1.1n^*$. In addition, we compute $F_5/n^*$ of these tests. Also included in Table 4.2 is the modified Haybittle–Peto test that uses $\varepsilon = 1/3$, giving a maximum sample size of $n^*$. Table 4.2 considers power not only at $\theta_1$ but also over a range of alternatives where the FSS test has a reasonable chance of rejection $H_0$. It can be seen that, when the maximum sample size is not inflated over the FSS, the power loss, as compared to the FSS test, is negligible for the O'Brien–Fleming and Haybittle–Peto tests, and also for $\text{ModHP}_{\varepsilon=1/3}$, but can be substantial for Pocock's test. When the maximum sample size is inflated by 5% or 10%, the modified Haybittle–Peto test performs nearly as well, in terms of expected sample size and power, as Eales and Jennison's optimal tests $F_2^\dagger$ and $F_5^\dagger$ and as the Wang–Tsiatis test. Furthermore, $\text{ModHP}_{\varepsilon=1/3}$ shows sample size savings over the O'Brien–Fleming test that has similar maximum sample size and power over the range of alternatives considered.

**Table 4.3** Power (%) and $E_\theta(T)/n^*$ (in parentheses) of various two-sided tests

| $\eta$ | FSS | ModHP$_{\varepsilon=1/3}$ | $\alpha_1^*$ | $\alpha_2^*$ | $\alpha_3^*$ | $\alpha_4^*$ | $\alpha_5^*$ |
|---|---|---|---|---|---|---|---|
| (a) Evenly spaced analyses | | | | | | | |
| 1.00 | 95 (1) | 94.3 (0.66) | 94.6 (0.70) | 91.1 (0.56) | 92.3 (0.58) | 93.4 (0.60) | 93.8 (0.62) |
| 0.90 | 90 (1) | 89.0 (0.72) | 89.4 (0.75) | 84.0 (0.63) | 85.9 (0.64) | 87.4 (0.66) | 88.3 (0.68) |
| 0.83 | 85 (1) | 83.7 (0.77) | 84.2 (0.78) | 77.6 (0.67) | 79.8 (0.69) | 81.3 (0.70) | 82.8 (0.72) |
| 0.78 | 80 (1) | 78.4 (0.80) | 79.1 (0.81) | 71.6 (0.71) | 74.1 (0.72) | 76.3 (0.74) | 77.5 (0.76) |
| 0.69 | 70 (1) | 68.1 (0.85) | 68.9 (0.85) | 60.5 (0.77) | 63.2 (0.78) | 65.7 (0.79) | 67.1 (0.81) |
| 0.61 | 60 (1) | 58.0 (0.88) | 58.9 (0.89) | 50.3 (0.81) | 53.0 (0.82) | 55.5 (0.83) | 57.0 (0.85) |
| (b) More frequent late analyses | | | | | | | |
| 1.00 | 95 (1) | 94.5 (0.67) | 94.4 (0.71) | 91.0 (0.58) | 92.2 (0.59) | 93.2 (0.62) | 93.7 (0.64) |
| 0.90 | 90 (1) | 89.3 (0.73) | 89.0 (0.75) | 83.9 (0.64) | 85.7 (0.65) | 87.2 (0.67) | 88.0 (0.70) |
| 0.83 | 85 (1) | 84.1 (0.78) | 83.8 (0.78) | 77.5 (0.68) | 79.6 (0.69) | 81.5 (0.71) | 82.5 (0.73) |
| 0.78 | 80 (1) | 79.0 (0.81) | 78.5 (0.81) | 71.5 (0.71) | 73.9 (0.72) | 76.0 (0.74) | 77.2 (0.76) |
| 0.69 | 70 (1) | 68.7 (0.85) | 68.3 (0.85) | 60.5 (0.77) | 63.0 (0.78) | 65.3 (0.79) | 66.7 (0.81) |
| 0.61 | 60 (1) | 58.6 (0.89) | 58.2 (0.88) | 50.5 (0.81) | 53.0 (0.82) | 55.3 (0.83) | 56.5 (0.85) |
| (c) More frequent early analyses | | | | | | | |
| 1.00 | 95 (1) | 94.0 (0.69) | 94.8 (0.77) | 91.9 (0.61) | 93.0 (0.62) | 93.4 (0.65) | 94.3 (0.68) |
| 0.90 | 90 (1) | 88.3 (0.76) | 89.8 (0.82) | 85.1 (0.67) | 86.7 (0.69) | 88.2 (0.71) | 88.9 (0.74) |
| 0.83 | 85 (1) | 82.8 (0.80) | 84.7 (0.85) | 78.8 (0.72) | 80.8 (0.73) | 82.7 (0.75) | 83.6 (0.78) |
| 0.78 | 80 (1) | 77.5 (0.81) | 79.6 (0.87) | 72.9 (0.75) | 75.3 (0.76) | 77.4 (0.79) | 78.4 (0.81) |
| 0.69 | 70 (1) | 67.0 (0.87) | 69.6 (0.91) | 61.9 (0.80) | 64.5 (0.81) | 66.9 (0.83) | 68.1 (0.85) |
| 0.61 | 60 (1) | 56.9 (0.90) | 59.6 (0.93) | 51.7 (0.84) | 54.3 (0.85) | 56.9 (0.87) | 58.1 (0.89) |

Here, each test has maximum sample size $n^*$, with $n^*$ being the fixed sample size that yields 95% power at $\theta_1$. The alternatives considered are $\eta\,\theta_1$

### 4.4.3 Unequal Group Sizes That Are Not Prespecified

When the group sizes, not necessarily equal, are unknown at the beginning of the trial, the modified Haybittle–Peto test can still be applied directly, whereas $F_5^\dagger$ and the O'Brien–Fleming and Pocock tests have to be implemented via error-spending functions; see Sect. 4.1.3. Lai and Shih (2004, pp. 519–520, 522–523) have shown that not only is the modified Haybittle–Peto test more convenient in this case but it can also outperform the error-spending approach. Consider group sequential trials for which the maximum sample size is the same as that of the FSS test but the group sizes are unknown at the beginning of the trial. Kim and DeMets (1987) considered group sequential tests, generated by five error-spending functions, for three analysis plans: (a) evenly spaced analyses at $n = (0.2, 0.4, 0.6, 0.8, 1.0)n^*$; (b) more frequent late analyses at $n = (0.3, 0.6, 0.8, 0.9, 1.0)n^*$; and (c) more frequent early analyses at $n = (0.1, 0.2, 0.3, 0.6,\ 1.0)n^*$, where $n^*$ is the sample size of the FSS test that has 95% power at $\theta_1$. The first two error-spending functions, $\alpha_1^*$ and $\alpha_2^*$, generate boundaries similar to those of the O'Brien–Fleming and Pocock tests, respectively. The other three error-spending functions $(\alpha_3^*, \alpha_4^*, \alpha_5^*)$ lie between $\alpha_1^*$ and $\alpha_2^*$. Table 4.3 gives the power functions and expected sample sizes of these five tests over the region where the FSS test has power ranging from 0.6 to 0.95. Also given for comparison are the values for ModHP$_{\varepsilon=1/3}$. Each of these results is based on 10000 simulations. For all three analysis plans, ModHP$_{\varepsilon=1/3}$ has power comparable to that of the $\alpha_1^*$ test, while its expected sample size, in parentheses and normalized by $n^*$, is smaller than that of the $\alpha_1^*$ test. The $\alpha_2^*$ test has the smallest expected sample size for all alternatives, but its power at small alternatives can be lower than that of the FSS test by as much as 10%.

Lai and Shih (2004) have also applied ModHP to the problem of testing $H_0 : p_1 = p_2$ concerning the success probabilities $p_1$ and $p_2$ of two treatments in a randomized two-armed trial, in which the sample size $n_{ij}$ for the two treatment arms can be different at the $j$th interim analysis and the group sizes $n_j = n_{1j} + n_{2j}$ can vary over $j$. Consider the case of known and equal group sizes, for which $n_j = (j/k)n_k$. Let $\hat{p}_{ij}$ be the proportion of successes for the $i$th treatment at the $j$th interim analysis, and let

$$\hat{p}_j = (n_{1j}\hat{p}_{1j} + n_{2j}\hat{p}_{2j})/n_j, \quad \hat{\sigma}_j^2 = \hat{p}_j(1-\hat{p}_j)(n_{1j}^{-1} + n_{2j}^{-1}), \quad Z_j = (\hat{p}_{1j} - \hat{p}_{2j})/\hat{\sigma}_j. \tag{4.27}$$

Suppose each subject is randomly allocated to either treatment. Then $n_{ij} = n_j/2 + O_p(\sqrt{n_j})$ and therefore $\hat{\sigma}_j^2 = n_j^{-1}\{4p(1-p) + O_p(n_j^{-1/2})\}$ under $H_0 : p_1 = p_2 = p$ and under the local alternatives $p_1 = p$, $p_2 = p + O_p(n_k^{-1/2})$. Hence, $\{Z_j/\hat{\sigma}_j, 1 \le j \le k\}$ has asymptotically normal increments under $H_0$ and under the local alternatives as the common group size becomes infinite, and therefore, group sequential tests of $\theta = 0$ for the mean $\theta$ of a normal distribution can be extended to test $H_0 : p_1 = p_2$; see Sect. 4.3.1. In particular, the Pocock and O'Brien–Fleming tests, with maximum sample size $n^*$, can be used to test $p_1 = p_2$ by replacing $S_{n_j}/(\sigma\sqrt{n_j})$ by $Z_j$. The equal-probability assignment can also be extended to the following adaptive treatment allocation scheme that attempts to reduce the expected sample size from the inferior treatment. Take $1/2 < q < 1$. Start with equal-probability assignment to either treatment. At the $j$th interim analysis $(1 \le j \le k-1)$, call the population with the larger $\hat{p}_{ij}$ the "leading population" and assign it with probability $q$ to subjects between the $j$th and $(j+1)$st analyses. When there is no leading population, that is $\hat{p}_{1j} = \hat{p}_{2j}$, use equal-probability assignment between the $j$th and $(j+1)$st analyses. For this adaptive allocation scheme, we still have $n_{ij} = n_j/2 + O_p(\sqrt{n_j})$ under $H_0 : p_1 = p_2$, and therefore, the preceding group sequential tests using the normal approximation to control the type I error probability can still be applied to test $H_0$ with the adaptive treatment allocation scheme.

Suppose the group sizes are unknown at the beginning of the trial. The modified Haybittle–Peto test can still be applied directly, but the Pocock and O'Brien–Fleming group sequential tests have to be implemented via their error-spending functions. The error-spending approach requires specification of the maximum information, which in the present case is $1/\hat{\sigma}_k^2$, where $\hat{\sigma}_j^2$ is defined in (4.27). Since $\hat{\sigma}_k^2$ is not available until the $k$th analysis, we need to replace it by some approximation. Noting that $1/\hat{\sigma}_k^2 = n_k/\{4\hat{p}_j(1-\hat{p}_j) + O_p(n_k^{-1/2})\}$ under $H_0$ and the assumption $\liminf n_1/n_k > 0$, we can implement the error-spending approach by using $n_k/\{4\hat{p}_j(1-\hat{p}_j)\}$ at the $j$th interim analysis as an approximation to the maximum information under $H_0$. With this approximation, the proportion of maximum information at the $j$th interim analysis is $4n_{1j}n_{2j}/(n_j n_k) \le n_j/n_k$. Table 4.4 assumes $k = 5$ groups and $M = 100$. Table 4.4a assumes equal group sizes and adaptive treatment allocation. Table 4.4b considers equal-probability assignment to either treatment but takes 15, 15, 20, 25, 25 for the five group sizes $n_1$, $n_2 - n_1, \dots, n_5 - n_4$. Table 4.4c uses the same group sizes as those in Table 4.4b and adaptive treatment allocation as in Table 4.4a. The Pocock and O'Brien–Fleming tests are implemented by the preceding error-spending approach. Each result is based on 5000 simulations. One reason for the substantial inflation of the type I error probability for the error-spending methods in Table 4.4 is the difficulty in estimating the maximum information, which is not yet observable at the time of the $j$th interim analysis but is needed in the error-spending approach. Although under $H_0: p_1 = p_2$ random treatment allocation in Table 4.4b suggests approximating $n_{1k}$ and $n_{2k}$ both by $n_k/2$, sampling fluctuations may result in substantial difference between $n_{1k}^{-1} + n_{2k}^{-1}$ and $4/n_k$. Moreover, despite their asymptotic equivalence under $H_0$, $4/n_k$ is no longer asymptotically equivalent to $n_{1k}^{-1} + n_{2k}^{-1}$ under alternatives because of adaptive treatment allocation. This explains why ModHP, which does not require estimation of the maximum information, outperforms the error-spending approach to implement the Pocock and O'Brien–Fleming methods ($\alpha_1^*$ and $\alpha_2^*$, respectively) because of unknown group sizes at the beginning of the trial.

## 4.5 Supplements and Problems

1. *Limiting joint distributions of group sequential GLR and nonparametric statistics.*

   For samples of fixed size $n_i$, the asymptotic normality of the signed-root likelihood ratio statistic (4.18) under $u(\boldsymbol{\theta}) = u_0$ is a standard result in likelihood theory. It is related to the asymptotic normality of the constrained maximum likelihood estimator and can be proved by linearization of $f_{\boldsymbol{\theta}}$ around $u(\boldsymbol{\theta}) = u_0$ and approximating the $(d-1)$-dimensional hypersurface $u(\boldsymbol{\theta}) = u_0$ by a hyperplane at every point belonging to the hypersurface. This linearization argument shows that (4.18) can be written as a sum of i.i.d. zero-mean and unit-variance random variables plus a remainder term that is of order $o_p(\sqrt{n_i})$. Another related consequence of this linearization argument is Wilks' theorem that the GLR statistic has a limiting $\chi_1^2$-distribution. For the sequential GLR statistics $W_1, \dots, W_k$, their joint asymptotic normality follows from the random walk approximation, with $W_i$ depending on $\boldsymbol{X}_1, \dots, \boldsymbol{X}_{n_i}$. Since the $\boldsymbol{X}_i$ are i.i.d., the independent increments property of the jointly normal limiting distribution of $(W_1, \dots, W_k)$ follows. Working with $W_i$ provides a simpler and more transparent derivation of the joint limiting distribution than working directly with the group sequential MLEs in Jennison and Turnbull (2000, Sect. 3.1 and Chap. 11).

**Table 4.4** Power (%) and expected sample size (in parentheses) of two-sided group sequential tests of $H_0 : p_2 - p_1 = 0$ without futility boundaries

| | $p_2 - p_1$ | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 0.15 | 0.2 | 0.24 | 0.27 | 0.30 |
| (a) Equal group sizes, adaptive treatment allocation | | | | | | |
| $\alpha_1^*$ | 6.1 (99.0) | 34.4 (93.7) | 52.2 (89.6) | 69.0 (84.2) | 78.1 (80.0) | 85.5 (75.6) |
| $\alpha_2^*$ | 7.1 (97.3) | 29.2 (90.1) | 45.9 (83.4) | 59.4 (77.4) | 68.6 (73.3) | 78.5 (66.0) |
| ModHP | 5.5 (98.9) | 36.1 (93.1) | 55.9 (89.0) | 69.6 (83.7) | 79.4 (78.6) | 86.2 (74.1) |
| (b) Unequal group sizes, even treatment allocation | | | | | | |
| $\alpha_1^*$ | 5.0 (99.4) | 36.8 (94.0) | 59.0 (89.0) | 72.6 (84.8) | 81.3 (80.7) | 87.2 (76.9) |
| $\alpha_2^*$ | 7.2 (97.3) | 33.3 (88.9) | 51.0 (81.3) | 64.7 (76.1) | 76.7 (70.6) | 82.8 (65.7) |
| ModHP | 5.3 (98.7) | 38.2 (93.0) | 58.1 (88.0) | 72.9 (82.5) | 80.9 (77.7) | 88.4 (72.5) |
| (c) Unequal group sizes, adaptive treatment allocation | | | | | | |
| $\alpha_1^*$ | 5.6 (99.3) | 35.3 (94.2) | 55.0 (89.7) | 70.6 (85.0) | 79.1 (81.5) | 86.2 (77.5) |
| $\alpha_2^*$ | 6.9 (97.1) | 27.9 (90.6) | 45.1 (84.4) | 59.8 (77.6) | 69.5 (73.7) | 78.9 (67.9) |
| ModHP | 5.6 (98.6) | 35.2 (93.0) | 55.3 (87.9) | 68.5 (84.0) | 77.8 (78.8) | 86.7 (74.1) |

For nonparametric test statistics of the type (4.21), a similar linearization argument together with some technical arguments involving the limiting distribution of $\sqrt{n_i'}(\hat{F}_{n_i'} - F)$ can be used to prove the random walk approximation (4.24).

2. *Bootstrap method for the multivariate extension of left-hand side of* (4.19).
The probability in (4.19) under the parameter value $\theta_0$, which is the largest value in the null hypothesis $\theta \leq \theta_0$, can be easily computed by Monte Carlo if one does not want to use normal approximations. For the multivariate null hypothesis $H_0 : u(\boldsymbol{\theta}) \leq u_0$, bootstrap theory for tests and confidence intervals, which will be described more fully in Chap. 7, can be applied to show that the Monte Carlo simulations can be carried out under $P_{\tilde{\boldsymbol{\theta}}_M}$, where $\tilde{\boldsymbol{\theta}}_M$ is the constrained maximum likelihood estimator of $\boldsymbol{\theta}$ under the constraint $u(\boldsymbol{\theta}) = u_0$. Chapter 7 also extends this approach to nonparametric and semiparametric tests in more general models and points out the importance of using appropriate pivots in the bootstrap approach to tests and confidence intervals.
3. *Local alternatives assumed at the design stage.*
Finding the implied alternative can be quite difficult in multivariate exponential families or for nonparametric hypotheses concerning $(F, G)$. As the maximum sample size for typical confirmatory trials is seldom small, the power calculations to determine the implied alternative at the design stage can assume local alternatives under which the test statistics are still approximately normal random walks. Using normal random walk approximations under both the null and the implied (local) alternative reduces the problem to that of a *locally asymptotically normal* (LAN) family, in which the relationship between the sample size $M$ and the implied alternative $\boldsymbol{\theta}(M)$ is relatively simple. Section 6.5.2 will explain further and illustrate this point in the context of censored rank statistics.
4. Show that the Wilcoxon statistic is a special case of generalized Chernoff–Savage statistics and write a program to carry out a group sequential Wilcoxon test, with $k = 5$ groups, of $H_0 : u(F, G) \leq 0.5$ using modified Haybittle–Peto boundaries and using equal randomization to the $X$ and $Y$ groups, where $u(F, G) = P_{F,G}(X < Y)$, subject to 5% type I error probability and a maximum total size of $M = 600$.
5. *Efficacy and futility stopping.*
Early stopping, prior to reaching the maximum sample size $M$, in the modified Haybittle–Peto test, is related to either crossing an upper boundary and rejecting $H_0$, or crossing a lower boundary and accepting $H_0$. The former amounts to efficacy stopping, claiming that the trial has shown the new treatment to be efficacious, and the latter is tantamount to futility stopping, making a no-go decision as there is sufficient evidence that the treatment effect is not large enough for it to lead to a successful claim at the scheduled end of the trial with $M$ observations. Although futility stopping does not inflate the type I error, care must be taken so that there is little loss of power compared to the FSS test that takes $M$ observations based on the power calculations at the design stage. In Sect. 4.2, we have assumed the futility stopping to be "binding" in the sense that the futility stopping boundary is strictly followed. This explains the "coupled" inequalities for not crossing either boundary in the probabilities defining $b$, $\tilde{b}$, and $c$ in Sect. 4.2.

6. A major reason why we have used a binding stopping rule in Sect. 4.2 is to provide a group sequential analog of the fully sequential GLR test (3.12). In practice, one may want to have the flexibility to override the futility stopping rule during the course of the trial since it does not inflate the type I error. Explain how you would modify the probability specifications for $b$, $\tilde{b}$, and $c$ in Sect. 4.2 to make futility stopping nonbinding.
7. *Group sequential versus adaptive designs.*
   Note that the $n_i$ are fixed in advance in a group sequential design, whereas they are determined adaptively from the data in an adaptive design to which we have referred in Sect. 1.5. For more complex group sequential designs that are considered in Chaps. 6 and 7, the $n_i$ need not be the sample size but may represent total information, such as total number of events, up to the $i$th interim analysis. Note that the information used to represent $n_i$ does not include the treatment difference between the two groups. It can be regarded as information under the null hypothesis that the treatment and control groups have the same effect. The error probability calculation, therefore, is more complicated for adaptive designs in which the $n_i$ can depend on the observed outcomes up to $n_{i-1}$, but still makes use of the Markov property of the test statistics to carry out recursive numerical integration or can be evaluated by the bootstrap method that uses Monte Carlo simulations.
8. *Group sequential designs in Phase II cancer trials.*
   Section 2.5 has introduced typical designs for Phase I cancer trials whose goal is to find the maximum tolerated dose (MTD). Because the tumor response rate for a cytotoxic treatment increases with dose, the goal of Phase II trials is to test if the MTD is effective in treating the tumor. The clinically definitive endpoint for cancer treatments, however, is survival rather than tumor response and is tested in Phase III trials. As pointed out by Vickers et al. (2007, p. 927), in a typical Phase II study of a novel cancer treatment, "a cohort of patients is treated, and the outcomes are compared to the prespecified target or bar. If the results meet or exceed the target, the treatment is declared worthy of further study; otherwise, further development is stopped. This has been referred to as the 'go/no-go' decision. Most often, the outcome specified is a measure of tumor response, e.g., complete or partial response using Response Evaluation Criteria in Solid Tumors, expressed as a proportion of the total number of patients. Response can also be defined in terms of the proportion who have not progressed or who are alive at a predetermined time (e.g., 1 year) after treatment is started." Note that the no-go decision corresponds to futility stopping.

   The most widely used designs for these single-arm Phase II trials are Simon's (1989) optimal two-stage designs, which allow early stopping of the trial if the treatment, using the dose chosen at the end of the Phase I trial, has not shown beneficial effect that is measured by a Bernoulli proportion. These designs are optimal in the sense of minimizing the expected sample size under the null hypothesis of no viable treatment effect, subject to type I and type II error probability bounds. Given a maximum sample size $M$, Simon considered the optimal stopping problem involving rules of the form that stops for futility after

$m < M$ patients if the number of patients exhibiting positive treatment effect is $r_1 (\leq m)$ or fewer, and otherwise treats an additional $M - m$ patients and rejects the treatment if and only if the number of patients exhibiting positive treatment effect is $r_2 (\leq M)$ or fewer. Simon's designs require that a null proportion $p_0$, representing some "uninteresting" level of positive treatment effect, and an alternative $p_1 > p_0$ be specified. The null hypothesis is $H_0 : p \leq p_0$, where $p$ denotes the probability of positive treatment effect. The type I and II error probabilities $P_{p_0}\{\text{Reject } H_0\}$, $P_{p_1}\{\text{Accept } H_0\}$ and the expected sample size $E_{p_0}N$ can be computed for any design of this form, which can be represented by the parameter vector $(m, M, r_1, r_2)$. Using computer search over these integer-valued parameters, Simon (1989) tabulated the optimal designs in his Tables 4.1 and 4.2 for different values of $(p_0, p_1)$. Note that Simon's designs are *group sequential designs with two groups* and early stopping only for futility; there is no early stopping for efficacy.

Whether the new treatment is declared promising in a Phase II trial depends strongly on the prescribed $p_0$ and $p_1$. In their systematic review of 134 papers reporting Phase II trials in *J. Clin. Oncology* or *Cancer*, Vickers et al. (2007) found 70 papers referring to historical data for their choice of the null or alternative response rate, and that nearly half (i.e., 32) of these papers did not cite the source of the historical data used, while only nine gave clearly a single historical estimate of their choice of $p_0$. Moreover, no study "incorporated any statistical method to account for the possibility of sampling error or for differences in case mix between the Phase II sample and the historical cohort." The modified Haybittle–Peto test applied to this setting chooses $p_1$ to be the alternative where the FSS test, with type I error probability $\alpha$ at $p_0$, has power $1 - \beta$, that is, choosing $p_1$ to be the solution of $F_{M,p_1}(F^{-1}_{M,p_0}(1-\alpha)) = \beta$, where $F_{M,p}$ is the distribution function of the $\text{Bin}(M, p)$ distribution. The GLR statistic at the $i$th stage is

$$n_i \left[ \hat{p}_{n_i} \log \left( \frac{\hat{p}_{n_i}}{p_j} \right) + (1 - \hat{p}_{n_i}) \log \left( \frac{1 - \hat{p}_{n_i}}{1 - p_j} \right) \right], \quad j = 0, 1.$$

Because the maximum sample size for a typical Phase II cancer trial is small, the number $k$ of groups in the modified Haybittle–Peto test is necessarily small. Bartroff and Lai (2008a) proposed to use $k = 3$ and determine $n_2$ adaptively after the first interim analysis. The enhancement of the modified Haybittle–Peto design is a special case of adaptive designs introduced in Sect. 8.2.1. Its stopping rule can be stated in terms of the number of cumulative successes $S_{n_i}$ at the $i$th stage, $i = 1, 2$. Table 4.5 describes the adaptive design (denoted by ADAPT) and Simon's (1989) optimal two-stage design (denoted by Sim2) for two choices of $m$, $M$, $\alpha$, $\beta$, and $p_0$, and Table 4.6 contains their operating characteristics, computed exactly using the $\text{Bin}(n, p)$ distribution. ADAPT has expected sample size close to Sim2 for $p$ near $p_0$, and smaller sample size when $p$ is roughly midway between $p_0$ and $p_1$ or is larger; $p_1 = 0.3$ in the top panel of Table 4.6 and $p_1 = 0.44$ in the bottom panel. The expected number of stages shows a similar

**Table 4.5** Description of ADAPT and Sim2 for two cases

| $S_m$ | ADAPT | Sim2 ($p_1 = 0.3$ or 0.44) |
|---|---|---|
| (a) $m = 10, M = 29, p_0 = 0.1, \alpha = 0.05, \beta = 0.2$ | | |
| $\le 1$ | Accept $H_0$ | Accept $H_0$ |
| 2 | $n_2 = M$; reject $H_0$ if $S_{n_2} \ge 6$ | $n_2 = M$ and |
| 3 | $n_2 = 20$ | reject $H_0$ if $S_M \ge 6$ |
| | (i) If $S_{n_2} \le 3$, accept $H_0$ | |
| | (ii) If $S_{n_2} \ge 6$, reject $H_0$ | |
| | (iii) If $4 \le S_{n_2} \le 5$ and $S_M \ge 6$, reject $H_0$ | |
| $\ge 4$ | Reject $H_0$ | $n_2 = M$; rej. $H_0$ if $S_M \ge 6$ |
| (b) $m = 30, M = 82, p_0 = 0.3, \alpha = \beta = 0.1$ | | |
| $\le 8$ | Accept $H_0$ | Accept $H_0$ |
| 9 | $n_2 = 57$ | Accept $H_0$ |
| | (i) If $S_{n_2} \le 19$, accept $H_0$ | |
| | (ii) If $S_{n_2} \ge 24$, reject $H_0$ | |
| | (iii) If $20 \le S_{n_2} \le 23$ and $S_M \ge 32$, reject $H_0$ | |
| $10 - 13$ | $n_2 = M$; reject $H_0$ if $S_{n_2} \ge 31$ | $n_2 = M$ and |
| $\ge 14$ | Reject $H_0$ | reject $H_0$ if $S_M \ge 30$ |

**Table 4.6** Expected sample size, power (in parentheses), and expected number of stages (in brackets) of Phase II designs

| $p$ | ADAPT | | | Sim2 | | |
|---|---|---|---|---|---|---|
| (a) $m = 10, M = 29, p_0 = 0.1, \alpha = 0.05, \beta = 0.2$ | | | | | | |
| 0.05 | 11.6 | (0.3%) | [1.1] | 11.6 | (0.2%) | [1.1] |
| $p_0 = 0.1$ | 14.5 | (5.0%) | [1.3] | 15.0 | (4.7%) | [1.3] |
| 0.2 | 18.8 | (43.3%) | [1.6] | 21.9 | (43.1%) | [1.6] |
| $p_1 = 0.3$ | 18.1 | (79.4%) | [1.6] | 26.1 | (79.6%) | [1.8] |
| 0.4 | 14.8 | (94.9%) | [1.4] | 28.1 | (95.0%) | [2.0] |
| 0.5 | 12.1 | (98.9%) | [1.2] | 28.8 | (98.9%) | [2.0] |
| 0.6 | 10.1 | (99.9%) | [1.0] | 29.0 | (99.9%) | [2.0] |
| (b) $m = 30, M = 82, p_0 = 0.3, \alpha = \beta = 0.1$ | | | | | | |
| 0.2 | 34.9 | (0.3%) | [1.1] | 33.2 | (0.03%) | [1.1] |
| $p_0 = 0.3$ | 51.8 | (10.0%) | [1.5] | 51.4 | (10.0%) | [1.4] |
| 0.35 | 60.4 | (35.0%) | [1.7] | 63.4 | (36.2%) | [1.6] |
| $p_1 = 0.44$ | 52.9 | (88.7%) | [1.5] | 77.7 | (87.8%) | [1.9] |
| 0.5 | 42.4 | (98.4%) | [1.3] | 80.9 | (97.5%) | [2.0] |
| 0.6 | 31.9 | (99.9%) | [1.0] | 82.0 | (99.9%) | [2.0] |

pattern, while their power functions are nearly identical. Note that even though ADAPT has a maximum of three stages; its expected number of stages is less than 2 for all $p$ and usually close to 1.

# Chapter 5
# Sequential Methods for Vaccine Safety Evaluation and Surveillance in Public Health

In this chapter we describe the applications of sequential testing methodology to the problem of testing the incidence rates of adverse events in vaccine clinical trials and post-marketing safety evaluation. Section 5.1 describes typical design considerations for vaccine safety evaluation and the application of the SPRT and its other sequential tests that have been applied to test vaccine safety. It also reviews recent developments in vaccine safety evaluation and the interest in sequential methods spurred by these developments. Section 5.2 describes the work of Shih et al. (2010) who introduced a new class of sequential generalized likelihood ratio (GLR) tests, a key ingredient of which is an exponential family representation of the rare event sequence under the commonly assumed model of Poisson arrivals of adverse events. Section 5.3 gives an illustrative example from the Rotavirus Efficacy and Safety Trial (REST). Section 5.4 describes post-marketing surveillance for vaccine and drug safety. In Sect. 5.5 we move beyond pharmacovigilance and describe change-point detection methods in public health surveillance. In particular, we show how the sequential GLR tests of Chaps. 3 and 4 can be modified into moving window sequential GLR detection schemes for quick detection of changes, subject to a constraint on the false-alarm rate.

## 5.1 Vaccine Safety Evaluation

Despite the significant public health impact seen from the introduction of vaccines, the safety of vaccines continues to receive considerable attention and has raised a variety of issues. First, the withdrawal of a rotavirus vaccine (a tetravalent rhesus–human reassortant rotavirus vaccine, RRV-TV) in 1999 has raised public concerns on vaccine safety and hence the balance of benefit and risk of a vaccine product; see Murphy et al. (2001). Second, unlike other therapeutic products, vaccines are typically given to healthy people and even to vulnerable populations such as infants and young children. In addition, many vaccines are universally recommended

J. Bartroff et al., *Sequential Experimentation in Clinical Trials,*
Springer Series in Statistics 298, DOI 10.1007/978-1-4614-6114-2_5,

and mandated for schooling and some special programs (e.g., military service), where the tolerance of vaccine risk is low. Hence, ensuring vaccine safety is important in public health activities and policies. Pre-licensure vaccine clinical trials usually involve selected populations who receive the vaccine according to a protocol-defined administration method and who are followed for a limited period after vaccination. Some commonly encountered adverse events, such as fever and injection site reaction, are easily observable and documented in vaccine clinical trials; however, a small number of extremely rare and sometimes potentially life-threatening adverse events may not be seen in such trials in spite of their large sample size. Hence, many regulatory agencies require that post-approval surveillance be implemented to monitor potential safety issues after the introduction of a new vaccine or vaccine component. Examples of post-approval vaccine safety surveillance include the Vaccine Adverse Event Reporting System (VAERS) and Vaccine Safety Datalink (VSD); see Ellenberg et al. (2005), Greene et al. (2011), and Nelson et al. (2012). Sequential safety monitoring is now common practice in vaccine clinical trials and post-licensure surveillance; see Davis et al. (2005), Lieu et al. (2007), and Kulldorff et al. (2011). The goal of sequential monitoring is quick detection of the association of adverse events that might be caused by the vaccine so that a confirmatory investigation and/or medical evaluation into the association can be launched.

### *5.1.1 Design Considerations for Clinical Trials to Test Vaccine Safety*

Safety profiles of vaccine candidates evolve throughout evaluations in laboratories, animals, phased human clinical trials, as well as post-marketing surveillance; see Chen et al. (2005). It is crucial to recognize that vaccines are different from most pharmaceutical products in many ways; understanding these differences is important in designing safety studies of vaccines. First, the safety standard is generally higher for vaccines than for drugs. Unlike therapeutic products, vaccines are usually administered to healthy populations, some of whom may be vulnerable children and infants. Some vaccines are universally recommended and as a result are administered to a large number of people. Hence, "first do no harm" is the widely accepted principle in public health, and a much lower risk tolerance is expected. Second, given that the duration of observation in prelicensure clinical trials is often less than 30 days (sometimes 42 days) after vaccination, the rarity of certain serious adverse events often necessitates a large sample size. For instance, with an incidence rate of 1 in 2000 person years and 30 days of postvaccination follow-up period after each of three-dose vaccinations, a sample of 60,000 subjects is required in order to observe approximately ten such events. Third, vaccines are biologically derived and variations in biological activities can occur. This is further complicated by variations in biological manufacturing processes such as formulation, fermentation, and virus sensitivity to storage condition, which together contribute to the variability of biologic activities. These factors may contribute to the adverse experience profile of the vaccine. In addition, many vaccines are combinations of multiple active biologic agents, and it is generally difficult, if not impossible, to attribute an adverse event to a particular agent. Finally, unlike drugs for which substitute therapies may be available, vaccines prevent significant morbidity and mortality and usually do not have many alternative options. Hence the decision to withdraw a vaccine should be made with extra care according to risk and benefit balance.

Safety assessment of a vaccine is an ongoing process throughout the product's life cycle. Statistical aspects of design and analysis play an important role in this process. The study design, such as the choice of endpoints, sample size, and study duration, is driven by the objectives, hypotheses, and prespecified criteria for success, which may vary depending on whether the vaccine is (a) the first vaccine for a particular disease or a vaccine for which a safety issue has been identified for similar vaccine products, or (b) for vulnerable populations, or (c) to be recommended for universal application. A continuous safety monitoring system is often used to detect increased risk of targeted adverse events as early as possible. In addition, it is widely recognized that the trial should have provisions for early termination due to unsafe outcomes associated with the vaccine during interim monitoring, which would minimize the risk to study participants.

### *5.1.2 Application of SPRT and Other Sequential Tests*

The design considerations for vaccine safety evaluation described in Sect. 5.1.1 pave the way for adopting fully sequential tests, beginning with the application of the SPRT by Davis et al. (2005) and the subsequent introduction of MaxSPRT by Lieu et al. (2007) and the conditional MaxSPRT by Li and Kulldorff (2010). Suppose $X_1, X_2, \dots$ are independent random variables with a common density $f$ and one is interested in testing $H_0 : f = f_0$ versus $H_1 : f = f_1$. Let $R_n = \prod_{i=1}^n f_1(X_i)/f_0(X_i)$ denote the likelihood ratio based on $X_1, \dots, X_n$. The sequential probability ratio test (SPRT) stops sampling at stage

$$T = \inf\{n \geq 1 : \ R_n \geq B \text{ or } R_n \leq A\} \tag{5.1}$$

and accepts $H_0$ (or $H_1$) if $R_T \leq A$ (or $R_T \geq B$), where $A$ and $B$ are chosen to satisfy the type I and type II error probability constraints $\alpha = P_0\{R_T \geq B\}$ and $\tilde{\alpha} = P_1\{R_T \leq A\}$. As shown in Sect. 1.2, the thresholds $A$ and $B$ can be approximated by using Wald's approximations to the error probabilities: $A \approx (\frac{\tilde{\alpha}}{1-\alpha})$ and $B \approx (\frac{1-\tilde{\alpha}}{\alpha})$. Dvoretzky et al. (1953) extended the SPRT to continuous-time processes with independent increments.

To apply the SPRT to vaccine safety testing, Lieu et al. (2007) and Kulldorff et al. (2011) assume that the number $N_t$ of adverse events within $d$ days following vaccination given to $m$ subjects in a clinical trial during the period $[0,t]$ follows a Poisson process with known mean $\mu_t$ for the population at risk. For subjects who have received the vaccine, they assume that the mean number of adverse events is still $\mu_t$ under $H_0$ but increases to $\rho\mu_t$ under $H_1$ with known $\rho > 1$. This is often called the "Poisson model" in the safety evaluation literature; see Greene et al. (2011, p. 584). The stopping rule of the continuous-time SPRT in this case is of the form $T = \inf\{t > 0 : R_t \geq B \text{ or } R_t \leq A\}$, where the likelihood ratio $R_t$ is the ratio of the density functions of $N_t$ under $H_i$ $(i = 0,1)$:

$$R_t = \frac{e^{-\rho\mu_t}(\rho\mu_t)^{N_t}/N_t!}{e^{-\mu_t}\mu_t^{N_t}/N_t!} = \rho^{N_t}e^{-(\rho-1)\mu_t}. \tag{5.2}$$

Because in practice it is often difficult to come up with an appropriate choice of $\rho$ for the alternative hypothesis, Lieu et al. (2007) maximize (5.2) over $\rho \geq 1$, yielding

$$\hat{R}_t = \sup_{\rho>1} \rho^{N_t}e^{-(\rho-1)\mu_t} = \exp\{-(\hat{\rho}_t - 1)\mu_t + N_t \log \hat{\rho}_t\}, \tag{5.3}$$

where $\hat{\rho}_t = \max(1, N_t/\mu_t)$ is the constrained MLE of $\rho$ $(\geq 1)$ at time $t$. They propose to use the stopping rule

$$\hat{T} = \inf\left\{t > 0 : \hat{R}_t \geq B\right\} \tag{5.4}$$

and to reject $H_0$ if $\hat{R}_{\hat{T}} \geq B$; see also Kulldorff et al. (2011). They call the test a MaxSPRT and use Monte Carlo simulations to determine its type I error probability and the power at various alternatives. Noting that the SPRT and the MaxSPRT do not have bounded stopping rules, Lieu et al. (2007) consider a variant of (5.4) that stops the trial at time $\tilde{T} = \min(\hat{T}, t^*)$ and rejects $H_0$ if $\hat{R}_{\tilde{T}} \geq B$, accepting $H_0$ otherwise.

Earlier, Davis et al. (2005) conducted a retrospective study that uses data submitted by the health maintenance organizations (HMOs) to the VSD from 1995 through 2000. The data are first segmented into weekly cohorts of vaccinated children. The weekly data are partitioned into a baseline period, which is defined as a period before the introduction of the new vaccine considered in the study, and the surveillance period beginning with the introduction of the vaccine. Each week's dataset is used to count the number of children receiving the vaccine for that week and the number diagnosed with adverse events within 30 days after the vaccination. Thus, $X_i$ in this case is binomial$(n_i, p)$, where $n_i$ is the number of children vaccinated in week $i$ and $X_i$ counts how many of them experience adverse events within the 30-day window. The null hypothesis is $H_0 : p = p_0$, where $p_0$ is determined from the event rate in the baseline period, and the alternative hypothesis is $H_1 : p = p_1$, where $p_1$ is based on the effect size that the study wants to detect, for example, $p_1 = 2p_0$. Davis et al. (2005) propose to use the SPRT to test $H_0$ versus $H_1$ based on the independent binomial random variables $X_i$ with density function $\binom{n_i}{X_i}p^{X_i}(1-p)^{n_i-X_i}$, which belong to an exponential family with natural parameter $\theta = \log(p/(1-p))$. This is often called the "binomial model"; see Greene et al. (2011, p. 585).

Instead of working with the Poisson process $N_t$, we can work with the interarrival times $X_i$ between successive adverse events. These are independent exponential random variables with means $\xi_i$. First, assume that $\mu_t = \lambda t$ and therefore all the $\xi_i$ are equal to $\xi = 1/\lambda$. The $X_i$ belong to the exponential family $f_\lambda(x) = \lambda e^{-\lambda x}$ with natural parameter $\theta = -\lambda$. The SPRT for testing $H_0 : \lambda \le \lambda_0$ versus $H_1 : \lambda \ge \lambda_1$ is

$$T = \inf\{n \ge 1 : \ R_n \ge b \text{ or } R_n \le a\}, \tag{5.5}$$

where $R_n = n\log(\lambda_1/\lambda_0) - (\lambda_1 - \lambda_0)S_n,\ S_n = \sum_{i=1}^n X_i$. Since the MLE is $\hat{\lambda}_n = n/S_n$, the stopping rule of the MaxSPRT is

$$\hat{T} = \inf\{n \ge 1 : \ \lambda_0 S_n - n - n\log(\lambda_0 S_n/n) \ge b\} \tag{5.6}$$

for $b > 0$. From the theory of sequential tests of one-sided hypotheses in Chap. 3, a more efficient extension of the SPRT is the sequential GLR test with stopping rule

$$\tau = \inf\left\{n \ge 1 : \ \max_{j=0,1}\left[\lambda_j S_n - n - n\log(\lambda_j S_n/n) - b_j\right] \ge 0\right\}, \tag{5.7}$$

rejecting $H_j$ upon stopping if $\lambda_j S_\tau - \tau - \tau\log(\lambda_j S_\tau/\tau) \ge b_j$. The truncated version of MaxSPRT can be regarded as a sequential GLR test without a lower boundary. The more general case in which $X_i \sim \text{Exp}(\lambda_i)$ have rates $\lambda_i$ varying with $i$, as in Lieu et al. (2007), can be converted back to the i.i.d. case by considering $X_i' = X_i\lambda_i \sim \text{Exp}(1)$. For more general $\mu_t$, we can still work with exponential interarrival times $X_i$ with known means $\xi_i$ for the population at risk and $\rho\xi_i$ for those who have received the vaccine. The hypotheses can be formulated as $H_0 : \rho \ge 1$ versus $H_1 : \rho \le 1 - \varepsilon$, with $\varepsilon > 0$. In this case, letting $\lambda_0 = 1$ and $\lambda_1 = 1/(1-\varepsilon)$, (5.7) can be generalized to

$$\tau = \inf\left\{n \ge 1 : \ \max_{j=0,1}\left[\lambda_j \sum_{i=1}^n \xi_i^{-1} X_i - n - n\log\left(\lambda_j \sum_{i=1}^n \xi_i^{-1} X_i/n\right) - b_j\right] \ge 0\right\},$$

and (5.5) and (5.6) can be generalized similarly.

Li and Kulldorff (2010, p. 286) note that in practice the $\mu_t$ in (5.2) are "usually estimated from historical data collected before the beginning of the surveillance among a cohort of subjects with no exposure to the vaccine" and that the random fluctuation in the estimates has not been considered in (5.4) or the MaxSPRT. For the MaxSPRT, they have found from a simulation study that estimating $\mu_t$ from a historical cohort with a small number of events can lead to inflation of the type I error probability. They therefore propose not to require a known baseline mean function $\mu_t$ for adverse events but to condition the maximized likelihood ratio $L_k$ (based on the interarrival times of the adverse events up to the time of the $k$th event in the surveillance group) on the total number of adverse events in the historical cohort during that surveillance period. The conditional maximized likelihood has an explicit form, but its sampling distribution under the null hypothesis of no increase in adverse event rate for the vaccinated group is complicated and has to be evaluated by Monte Carlo simulations; see Sect. 5.6 for further details.

## 5.2 Sequential GLR Tests in Prelicensure Randomized Clinical Trials

### *5.2.1 An Exponential Family Representation of Sequential GLR Tests*

Consider a clinical trial in which subjects are randomized to receiving vaccine or placebo. Assume that the arrivals of adverse events follow a Poisson process, with rate $\lambda_V$ for vaccine (V) and $\lambda_C$ for placebo (C) recipients. This assumption will be relaxed later by allowing the rates to vary with time. When an event occurs, it is associated with either V or C and

$$P(\mathrm{V}\,|\,\text{event occurs at time } t \text{ after previous one}) \\ = \frac{\lambda_V e^{-\lambda_V t}\cdot e^{-\lambda_C t}}{(\lambda_V+\lambda_C)e^{-(\lambda_V+\lambda_C)t}} = \frac{\lambda_V}{\lambda_V+\lambda_C}. \tag{5.8}$$

Suppose adverse events occur at times $T_1 < T_2 < \ldots$ and the event indicator at $T_i$ is $\delta_i = 1$ for V or 0 for C. Let $\tau_i = T_i - T_{i-1}$. Since the Poisson interarrival times are i.i.d. exponential, it follows from (5.8) that the likelihood function of $(\lambda_V, \lambda_C)$ based on the observations $(T_i, \delta_i),\ 1 \le i \le n$, is

$$\prod_{i=1}^{n}\left[\left(\frac{\lambda_V}{\lambda_V+\lambda_C}\right)^{\delta_i}\left(\frac{\lambda_C}{\lambda_V+\lambda_C}\right)^{1-\delta_i}(\lambda_V+\lambda_C)\,e^{-(\lambda_V+\lambda_C)\tau_i}\right]. \tag{5.9}$$

The goal of a prelicensure randomized clinical trial is to show that the vaccine product is safe. This can be formulated as testing $H_0 : \lambda_V/\lambda_C \le 1$ versus $H_1 : \lambda_V/\lambda_C \ge \gamma$, where $\gamma > 1$. Let $p = \frac{\lambda_V}{\lambda_V+\lambda_C}$. Then $\lambda_V/\lambda_C \ge \gamma$ if and only if $p \ge \frac{\gamma}{1+\gamma}$. Let $p_0 = 1/2$ and $p_1 = \gamma/(1+\gamma)$. In view of (5.9), the likelihood ratio statistic for testing $H_0$ versus $H_1$ is $\prod_{i=1}^{n}(\frac{p_1}{p_0})^{\delta_i}(\frac{1-p_1}{1-p_0})^{1-\delta_i}$. Hence there is no loss of information in working with the Bernoulli distribution; that is, the actual event times contain no additional information about $\lambda_V/\lambda_C$ beyond that provided by the type (V or C) of the events. This argument also applies to $\lambda_{V,i}$ and $\lambda_{C,i}$ that vary with $i$, since $\prod_{i=1}^{n}(\lambda_{V,i}+\lambda_{C,i})e^{-(\lambda_{V,i}+\lambda_{C,i})\tau_i}$ is cancelled out in the likelihood ratio statistic, as the $\delta_i$ are still independent Bernoulli random variables with means $\pi_i = \lambda_{V,i}/(\lambda_{V,i}+\lambda_{C,i})$. The appendix of Shih et al. (2010) describes an algorithm to implement the GLR test, for which a software package has been developed using R and is available at the book's website.

The SPRT for testing $H_0 : \pi_i \le p_0$ versus $H_1 : \pi_i \ge p_1$ (for all $i$) is

$$T = \inf\{n \ge 1 : l_n \ge b \text{ or } l_n \le a\}, \tag{5.10}$$

where $l_n = \sum_{i=1}^{n}\{\delta_i \log(\frac{p_1}{p_0}) + (1-\delta_i)\log(\frac{1-p_1}{1-p_0})\}$ and $a < 0 < b$. The SPRT does not have a bounded stopping rule. The GLR statistic for testing $p_j$ $(j = 0,1)$ has logarithm

$$l_{n,j} = \sum_{i=1}^{n} \left\{ \delta_i \log\left(\hat{p}_n/p_j\right) + (1-\delta_i)\log\left[(1-\hat{p}_n)/(1-p_j)\right] \right\}, \tag{5.11}$$

where $\hat{p}_n = (\sum_{i=1}^{n} \delta_i)/n$. The truncated MaxSPRT has stopping rule $\tilde{T} = \min\{\hat{T}, n^*\}$, where

$$\hat{T} = \inf\{n \geq 1 : l_{n,0} \geq b \hat{p}_n > p_0\}, \tag{5.12}$$

and rejects $H_0$ if $l_{\tilde{T},0} \geq b$.

A more efficient extension to composite hypotheses than the MaxSPRT is the sequential GLR test

$$\tau = \inf\{n \geq 1 : l_{n,0} \geq b_0 \quad \text{and} \quad \hat{p}_n > p_0, \quad \text{or} \quad l_{n,1} \geq b_1 \quad \text{and} \quad \hat{p}_n < p_1\}. \tag{5.13}$$

It is in fact asymptotically efficient for testing $H_0 : p \leq p_0$ versus $H_1 : p \geq p_1$; see the asymptotic theory of sequential GLRs for testing one-sided hypotheses in exponential families in Sects. 3.4 and 3.7.3. The stopping rule (5.13) is bounded above by $n^*$, where $n^*$ is the smallest integer $n$ such that $nI(p^*) \geq \max(b_0, b_1)$ and $p^* \in (p_0, p_1)$ is the solution of the equation

$$p^* \log\left(\frac{p^*}{p_0}\right) + (1-p^*)\log\left(\frac{1-p^*}{1-p_0}\right) = p^* \log\left(\frac{p^*}{p_1}\right) + (1-p^*)\log\left(\frac{1-p^*}{1-p_1}\right),$$

whose common value is denoted by $I(p^*)$. Note that (5.13) introduces a lower boundary into (5.12) to allow early stopping for "futility" in the sense that the vaccine is unlikely to be shown unsafe by the prescheduled end of the trial (after observing $n^*$ adverse events).

### 5.2.2 *Implementation and Example*

For a given type I error probability $\alpha$ and type II error probability $\tilde{\alpha}$, the thresholds in the stopping rule of the SPRT can be approximated by using Wald's approximations. The thresholds of the truncated MaxSPRT test and the sequential GLR test can be obtained by solving for the largest positive constants that satisfy the error probability constraints. For example, if $X_i \sim \mathrm{Exp}(\lambda)$, the threshold $b$ of the MaxSPRT truncated at $n^*$ is the solution of $\mathrm{P}_{\lambda_0}(l_{\tilde{T},0} \geq b) = \alpha$, where $\tilde{T} = \min(\hat{T}, n^*)$ and $\hat{T}$ is given by (5.12); the thresholds $b_0, b_1$ of the stopping rule (5.13) of the sequential GLR test are the solutions of $\mathrm{P}_{\lambda_0}(l_{\tau,0} \geq b_0) = \alpha$ and $\mathrm{P}_{\lambda_1}(l_{\tau,1} \geq b_1) = \tilde{\alpha}$. Because the $X_i$ are independent, these error probabilities can be computed by recursive numerical integration using the Markov property of the random walk $l_n$ or

**Table 5.1** Power and expected sample size for various sequential tests of $H_0 : \lambda_V/\lambda_C = 1$ versus $H_1 : \lambda_V/\lambda_C \geq 3$ in a two-armed prelicensure clinical trial

| | | $\text{SPRT}_1^{\text{b}}$ | | | $\text{SPRT}_2^{\text{b}}$ | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $\lambda_V/\lambda_C$ | GLR[a] | $\gamma = 2.0$ | 3.0 | 5.0 | $\gamma = 2.0$ | 3.0 | 5.0 | $\text{MaxSPRT}_1^{\text{c}}$ | $\text{MaxSPRT}_2^{\text{c}}$ |
| (a) Expected total number of events | | | | | | | | | |
| 1.0 | 17.4 | 37.0 | 16.2 | 8.3 | 35.8 | 16.2 | 8.3 | 957.4 | 96.5 |
| 2.0 | 29.4 | 45.2 | 27.6 | 14.4 | 43.4 | 27.4 | 14.4 | 63.8 | 49.2 |
| 3.0 | 21.8 | 26.2 | 20.3 | 14.2 | 26.2 | 20.3 | 14.2 | 28.2 | 24.5 |
| 4.0 | 16.5 | 20.3 | 15.9 | 12.4 | 20.3 | 15.9 | 12.4 | 19.3 | 17.1 |
| 5.0 | 13.6 | 17.6 | 13.7 | 11.0 | 17.6 | 13.7 | 11.0 | 15.4 | 13.9 |
| (b) Probability of rejecting $H_0$ | | | | | | | | | |
| 1.0 | 0.041 | 0.044 | 0.043 | 0.044 | 0.042 | 0.043 | 0.044 | 0.050 | 0.048 |
| 2.0 | 0.642 | 0.914 | 0.647 | 0.398 | 0.860 | 0.639 | 0.398 | 1.000 | 0.865 |
| 3.0 | 0.931 | 0.994 | 0.926 | 0.729 | 0.993 | 0.925 | 0.730 | 1.000 | 0.998 |
| 4.0 | 0.979 | 0.999 | 0.978 | 0.873 | 0.999 | 0.978 | 0.873 | 1.000 | 1.000 |
| 5.0 | 0.991 | 1.000 | 0.992 | 0.932 | 1.000 | 0.992 | 0.932 | 1.000 | 1.000 |

[a] The thresholds $b_0 = 3.466$ and $b_1 = 2.773$ are chosen such that $p_{\lambda_V/\lambda_C=1}(\text{reject } H_0) \leq 0.05$, $p_{\lambda_V/\lambda_C=3}(\text{accept } H_0) \leq 0.10$

[b] Truncated at $n^* = 1000$ ($\text{SPRT}_1$) or $n^* = 100$ ($\text{SPRT}_2$); $\gamma$ is the assumed alternative value of $\lambda_V/\lambda_C$ in the likelihood ratio statistic. The thresholds $a = -2.251$ and $b = 2.890$ are obtained by using Wald's approximations to boundary crossing probabilities

[c] Truncated at $n^* = 1000$ ($\text{MaxSPRT}_1$, $b = 4.130$) or $n^* = 100$ ($\text{MaxSPRT}_2$, $b = 3.466$). The threshold $b$ is chosen such that $p_{\lambda_V/\lambda_C=1}(\text{ reject } H_0) \leq 0.05$

$l_{n,j}$. If $X_i$ is discrete, the integration is replaced with summation. When $n^*$ is large, it is more convenient to use Monte Carlo simulations instead of recursive numerical integration to compute the error probabilities.

*Example 5.1.* Shih et al. (2010) have carried out a simulation study on a prelicensure randomized clinical trial to test $H_0 : \lambda_V/\lambda_C = 1$ versus $H_1 : \lambda_V/\lambda_C \geq 3$, with prescribed type I error probability $\alpha = 0.05$ and type II error probability $\tilde{\alpha} = 0.1$ at $\lambda_V/\lambda_C = 3$. Table 5.1 gives the expected sample size and power for the SPRT, MaxSPRT, and the sequential GLR test, whose stopping rules are given by (5.10), (5.12), and (5.13). The SPRT and MaxSPRT are truncated at 1000 or 100 events (2 cases); 100 is the maximum number of events for the sequential GLR test. To determine the thresholds of the stopping rules, the boundary crossing probabilities of the SPRT are obtained by Wald's approximations, and those of MaxSPRT and the sequential GLR test are computed by using recursive numerical summation. Table 5.1, whose results are computed by recursive numerical summation, shows the superior performance of the sequential GLR test in two-armed randomized trials.

### 5.2.3 *Discussion*

As noted in Sect. 1.2, although refinements and modifications of Wald's SPRT for the design of clinical trials had been developed in the 1950s, they received little attention from the biomedical community until the Beta-Blocker Heart Attack Trial (BHAT). The main reason for this lack of interest is that the sample size for a typical trial is too small to allow further reduction while still maintaining reasonable power at the alternatives of interest. The success of BHAT led to the development and increasing use of group sequential designs in Phase III clinical trials. The development of vaccine safety tests in the past decade seems to have given fully sequential methods a surge of interest that had been lacking in clinical trials since the 1950s. For rare adverse events following vaccination (V) or placebo (C) injection, the effective sample size is the total number of adverse events in the sample of a large number of subjects accrued over a number of years. Section 5.2.1 has shown how this effective sample size can be used to develop an efficient sequential test comparing the event rates of the V and C treatments in a prelicensure randomized clinical trial. An additional advantage of the information-based design is that one can adjust, without altering the type I and type II error probabilities, the total number of subjects accrued per year and the number of years as the trial progresses, based on the observed adverse event rate of the combined V and C groups as the trial progresses. Section 5.5.3 will discuss an important difference, in the formulation of the null hypothesis, between safety evaluation and efficacy testing for a new drug or vaccine, but the fully sequential and group sequential methodologies described in Chaps. 3 and 4 are applicable to both kinds of testing problems.

## 5.3 The Rotavirus Efficacy and Safety Trial (REST)

The REST is a blinded, placebo-controlled clinical trial conducted in 11 countries between 2001 and 2004, to assess the efficacy and safety of a pentavalent human–bovine reassortant rotavirus vaccine (RV5). Infants between 6 and 12 weeks of age were randomized at a 1:1 ratio to receive either three doses of RV5 or placebo. All infants were monitored for serious adverse events for at least 42 days after each dose, which is the typical time frame for observation of adverse events following live virus vaccines. The primary safety hypothesis was that RV5 would not increase the risk of intussusception, relative to placebo, within 42 days after any dose. This concern of potential increased risk of intussusception, which is a serious yet uncommon illness with a background incidence rate of 18–56 cases per 100,000 infant years during the first year of life, stems from the withdrawal of a tetravalent rhesus–human reassortant rotavirus vaccine (RRV-TV) in October 1999 when the post-licensure safety surveillance revealed a substantial short-term increase in the risk of intussusception among RRV-TV recipients, primarily in the exposure window 3–14 days after the first dose; see Murphy et al. (2001, 2003) and Heyse et al. (2008).

Assuming that the arrivals of intussusception cases follow a Poisson process, with rate $\lambda_V$ for vaccine and $\lambda_C$ for placebo recipients, as in Sect. 5.2, Heyse et al. (2008) made use of the fact that conditional on total number $n$ of cases from both groups, the number of vaccine cases is binomial$(n, p)$, where $p = \lambda_V/(\lambda_V + \lambda_C)$. They therefore applied a repeated significance test that terminates the study after $n$ intussusception cases are observed and declares the vaccine to be unsafe if

$$\mathrm{P}\{\text{Binomial}(n, p_0) \geq \#_n(V)\} \leq 0.025, \tag{5.14}$$

where $\#_n(V)$ denotes the number of vaccine cases among the $n$ cases and $p_0 = 1/2$. The study is also terminated and declares the vaccine to be safe if

$$\mathrm{P}\{\text{Binomial}(n, p_1) \leq \#_n(V)\} \leq 0.025, \tag{5.15}$$

where $p_1 = 10/11$, corresponding to a ten-fold increase in risk for the vaccine group. Although the nominal significance level of 0.025 in (5.14) or (5.15) does not adjust for repeated analysis of the accumulated data, Monte Carlo simulations (involving 10000 random sequences) showed that the probability for the study to stop with a positive conclusion regarding vaccine safety is 0.94 for a vaccine with no increased risk of intussusception, and the probability for the study to declare the vaccine to be unsafe is almost 1 for relative risks of 6 or greater; see Heyse et al. (2008). This conservative approach is appropriate given the nature of the safety evaluation. Section 5.2 provides a methodological innovation that leads to independent Bernoulli random variables without conditioning on the total number of events, thereby making conventional sequential tests directly applicable (to these independent Bernoulli observations).

During the study, all suspected cases of intussusception were promptly reported to, and adjudicated by, an independent, blinded adjudication committee. The study stopped enrollment upon the recommendation of the Data and Safety Monitoring Board (DSMB) when about 70,000 infants had completed their follow-up. At that time, there were 11 confirmed cases of intussusception, 6 in the vaccine group and 5 in the placebo group. Figure 5.1 summarizes the sequentially accumulated data and the boundaries of (a) the repeated significance test (5.14)–(5.15) and (b) the sequential GLR test (5.13). Here, $p_0 = 1/2$ and $p_1 = 10/11$. The lower boundary of the repeated significance test was crossed, confirming the safety of the vaccine, and therefore the DSMB recommended to stop the study since the predefined criteria were met. If the sequential GLR test (5.13) had been used instead, the lower boundary would also have been crossed at the same time.

In the REST study, the lower "safe" boundary actually used a group sequential design for the DSMB to conduct interim analysis, starting with a minimum of 60,000 infants and subsequent groups of 10,000 infants. Therefore, stopping at the

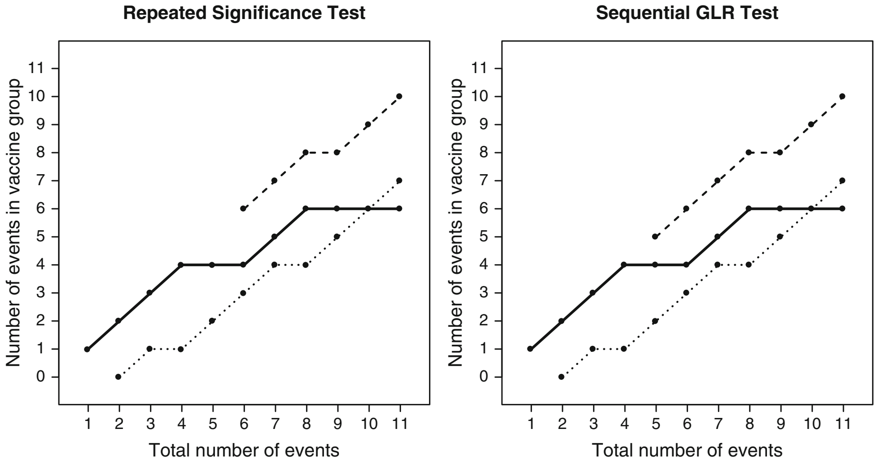


**Fig. 5.1** Stopping boundaries of the repeated significance test and sequential GLR test for the REST study, where the unsafe boundaries are in *dashed lines* and the safe boundaries in *dotted lines*. Also given are the observed data (*solid lines*)

lower boundary involves the total number of intussusception cases of the vaccine and placebo recipients up to the time of each interim analysis. The implementation methods described in Sect. 5.2 can be easily modified to handle this situation.

## 5.4 Post-marketing Surveillance of Drug Safety

Post-marketing drug or vaccine safety surveillance is essential for public health and safety as the limited sample size, study duration, and target population reflected in the inclusion/exclusion criteria of the preapproval clinical trials make it virtually impossible to detect all possible side effects. Pharmacovigilance systems have been established in different parts of the world. In the USA, such systems include VAERS (see the first paragraph of Sect. 5.1) and MedWatch that are passive surveillance systems accepting voluntary reports on adverse events associated with approved vaccines and drugs. Active surveillance systems using health plans' electronic medical claims data that contain information about both exposure and adverse events status have also been developed in the past two decades. An example is the VSD project mentioned in Sect. 5.1, which is a collaborative effort between the Centers for Disease Control and Prevention (CDC) and eight large health plans. Another example is the Post-Licensure Rapid Immunization Safety Monitoring (PRISM) program, established by the FDA in 2009 to monitor the safety of the H1N1 influenza vaccine using data from national health insurance plans and immunization registries; see Nguyen et al. (2012). PRISM is now integrated into the Mini-Sentinel program initiated by the FDA in 2009 as part of its Sentinel

Initiative. Mini-Sentinel is a collaborative effort between the FDA and 31 academic and private organizations to use routinely collected electronic healthcare data to perform active surveillance of the safety of marketed medical products including drugs, biologics, and medical devices.

Nelson et al. (2012, p. 63) point out that sequential testing offers a "particularly promising approach" to post-marketing surveillance evaluations of the safety of regulated medical products by monitoring electronic healthcare records that are routinely collected by insurance plans. They mention MaxSPRT as an example that has successfully detected an increased risk of seizure after receiving a new combination vaccine and Li's (2009) conditional sequential sampling procedure (CSSP) as a promising new tool designed specifically to handle confounding of observational data.

The CSSP is a group sequential method to test if a new drug $D$ leads to an elevated risk for an adverse event compared with an established drug $C$. The method is designed for prospective drug safety surveillance studies in which, for each considered drug, a summary table with the exposed person-times and the associated numbers of adverse events within each of the strata (defined by several potential confounders) is updated periodically using the health plans' administrative claims data. It was motivated by a retrospective study using the administrative claims data collected from 2000 to 2005 by nine large integrated healthcare systems. As noted in Sect. 5.1.2, the Poisson-based MaxSPRT method requires the availability of rich historical data to provide reliable estimates of the expected numbers $\mu_t$ of adverse events for drug $D$ under the null hypothesis. To apply the MaxSPRT in this retrospective study, all available data for the comparison drug $C$ collected during the entire study period are used as historical data to obtain an estimate of $\mu_t$. This requirement would not be satisfied in prospective studies in which both drugs $D$ and $C$ are new. To address this issue, CSSP assumes a semiparametric Poisson regression model for the numbers of adverse events within each stratum for each drug. The parameter of interest is the relative risk; the nuisance parameters in the model reflect possible temporal effect on event risks and population heterogeneity. Using the relative risk as the test statistic, the CSSP test is based on the conditional distribution of this test statistic given the sufficient statistics of the nuisance parameters, which are the numbers of adverse events within each stratum during each time period. By conditioning on sufficient statistics of the nuisance parameters, it preserves the overall type I error with any specified error spending function and adjusts for temporal trend and population heterogeneity across strata. The probability of having more "extreme" outcomes than the observed value is obtained via a sequential sampling procedure in which independent realizations of the test statistics are generated under the null hypothesis conditional on the values of the sufficient statistics of the nuisance parameters.

As pointed out by Platt et al. (2012), Mini-Sentinel is a pilot program that has developed policies, procedures, and technical specifications for developing and operating a secure distributed data system comprised of data covering enrollment, demographics, encounters, diagnosis, procedures, and ambulatory dispensing of prescription drug. Cook et al. (2012) have recently reviewed four sequential testing methods for monitoring post-marketing safety of a medical product in the Mini-Sentinel pilot program. These methods include (a) the Lan–DeMets error-spending approach for asymptotically normal test statistics (one such test statistic is the standardized relative risk estimate in Poisson regression to control for confounding), (b) MaxSPRT applied to binomial data as in the approach of Davis et al. (2005) described in Sect. 5.1.2 and using exposure matching to control for confounding, (c) the CSSP described in the preceding paragraph, and (d) a group sequential approach applied to the score statistics of a generalized estimating equation (GEE) that only requires specification of the mean model to adjust for confounding. Poisson regression is a special case of generalized linear models (Supplement 2 of Sect. 2.6), to which GEE is also closely related.

Sequential testing requires precise specification of the maximum sample size and study duration and the null hypothesis on the adverse event rate of subjects treated by the new medical product. This may be too restrictive in safety monitoring using continually updated electronic healthcare data. As Cook et al. (2012) point out, the current statistical methods for post-marketing safety surveillance in their review "represent a first step toward a general methodology appropriate for the signal refinement surveillance setting." Signal refinement refers to post-marketing evaluation of prespecified potential adverse events using a prospective observational design with existing electronic healthcare data to compare event rates in recipients of the new medical product to a comparable control cohort after adjusting for confounders.

## 5.5 Sequential Change-Point Detection Methods for Pharmacovigilance and Public Health Surveillance

There is an extensive literature in engineering and statistics on the subject of quick detection, with low false alarm rate, of faults or defects in a production or control system on the basis of sequential observations from the system; see Lai (1995, 2001). Sonesson and Bock (2003) have given a review and discussion of the applications of some of those methods to prospective surveillance in public health. In Sect. 5.5.1 we give an overview of the methodology of sequential change-point detection, highlighting its connections to sequential testing theory in Chap. 3. Section 5.5.2 then discusses modifications and extensions of the methodology for applications to surveillance in public health. In Sect. 5.5.3 we outline some related work to develop methods for pharmacovigilance.

### *5.5.1 Sequential Change-Point Detection Methodology: An Overview*

Whereas the theory of sequential testing began with Wald's SPRT in response to demands for more efficient testing of weaponry during World War II, the theory of sequential change-point detection began with Page's (1954) CUSUM chart and the Shiryaev–Roberts chart (Roberts 1966) that were developed to improve the traditional Shewhart chart in quality control by modifying Wald's SPRT. As noted by Lai (2001), the subject of statistical quality control is concerned with monitoring and evaluation of the quality of products from a continuous production process. Shewhart (1931) introduced (a) the fundamental concept of a "state of statistical control," in which the behavior of some suitably chosen quality characteristics at time $t$ has a given probability distribution, and (b) a process inspection scheme that takes samples of fixed size at regular intervals of time and computes from the sample at time $t$ a suitably chosen statistic $X_t$, which can be presented graphically in the form of a control chart. Shewhart's control chart, therefore, is a "single-sample" scheme whose decision depends solely on the current sample although the results of previous samples are available from the chart. To improve the sensitivity of the Shewhart chart, Page (1954) and subsequently Lorden (1971) considered a statistical model, denoted by $P^{(\nu)}$, that consists of a sequence of independent random variables $X_1, X_2, \ldots$ such that the $X_t$ have a common specified distribution $F_0$ for $t < \nu$, representing Shewhart's "state of statistical control," and such that the $X_t$ have another common distribution $F_1$ for $t \geq \nu$. Let $P_0$ denote the alternative model of perpetual statistical control (corresponding to $\nu = \infty$). Assuming that $F_0$ and $F_1$ have densities $f_0$ and $f_1$ with respect to some measure, Lorden noted that Page's CUSUM rule can be written as

$$N = \inf\left\{ n : \max_{1 \leq k \leq n} \sum_{i=k}^{n} \log(f_1(X_i)/f_0(X_i)) \geq c_\gamma \right\} \tag{5.16}$$

and showed that it minimizes asymptotically the worst-case detection delay in the following sense. Let $c_\gamma$ be so chosen that $E_0(N) = \gamma$, and let $\mathscr{F}_\gamma$ be the class of all monitoring schemes subject to the constraint $E_0(T) \geq \gamma$. Then $E_0(N) \geq \exp(c_\gamma)$ and that for $c_\gamma = \log\gamma$,

$$\begin{aligned} &\sup_{\nu \geq 1} \text{ess sup} E^{(\nu)}\left[(N - \nu + 1)^+ \,\middle|\, X_1, \ldots, X_{\nu-1}\right] \sim (\log\gamma)/I(f_1, f_0) \\ &\sim \inf_{T \in \mathscr{F}_\gamma} \left\{ \sup_{\nu \geq 1} \text{ess sup} E^{(\nu)}\left[(T - \nu + 1)^+ \,\middle|\, X_1, \ldots, X_{\nu-1}\right] \right\} \quad \text{as } \gamma \to \infty, \end{aligned} \tag{5.17}$$

where $I(f_1, f_0) = E_{f_1}\{\log(f_1(X_1)/f_0(X_1))\}$ denotes the Kullback–Leibler information number. Note that in (5.17), $\nu$ represents the change-time and $(N - \nu + 1)^+$ represents the detection delay so that $\sup_{\nu \geq 1}$ refers to the worst-case change-time. The conditional expectation $E^{(\nu)}[(N - \nu + 1)^+ | X_1, \ldots, X_{\nu-1}]$ is the conditional expected delay given all observations prior to the change-point; it is a random variable depending on these pre-change observations, and essential supremum again refers to the worst-case scenario over all possible values of $(X_1, \ldots, X_{\nu-1})$.

An important observation by Lorden (1971) is that the CUSUM rule (5.16) corresponds to stopping when a one-sided SPRT (without the lower boundary) based on $X_{\hat{k}}, X_{\hat{k}+1}, \ldots$ rejects the null hypothesis $H_0 : f = f_0$, where $\hat{k}$ is the maximum likelihood estimate of the change-time $\nu$. Thus, (5.16) can be expressed as

$$N = \min_{k \geq 1} (N_k + k - 1), \tag{5.18}$$

where $N_k$ is the stopping time of the one-sided SPRT applied to $X_k, X_{k+1}, \ldots$. Instead of the stopping rule of the one-sided SPRT, one can use other stopping rules. Lorden (1971) showed that if $X_1, X_2, \ldots$ are i.i.d. and $\tau$ is a stopping time with respect to $X_1, X_2, \ldots$ such that $P_0(\tau < \infty) \leq \alpha$, then letting $N_k$ be the stopping time obtained by applying $\tau$ to $X_k, X_{k+1}, \ldots$ and defining $N$ by (5.18), $E_0(N) \geq 1/\alpha$ and $N$ is a stopping time. Making use of Lorden's result with $\tau = m_\gamma$ if $\sum_{i=1}^{m_\gamma} \log(f_1(X_i)/f_0(X_i)) \geq \log \gamma$ and $\tau = \infty$ otherwise, Lai (1995) showed that the moving average scheme

$$N^* = \inf\left\{ n : \sum_{i=n-m_\gamma+1}^{n} \log(f_1(X_i)/f_0(X_i)) \geq \log \gamma \right\} \tag{5.19}$$

satisfies both $E_0(N^*) \geq \gamma$ and the asymptotic minimax property (5.17) (with $N$ replaced by $N^*$) if the fixed sample size $m_\gamma$ of the Neyman–Pearson test in $N^*$ is so chosen that

$$m_\gamma \sim (\log \gamma)/I(f_1, f_0) \quad \text{and} \quad \left\{ m_\gamma - (\log \gamma)/I(f_1, f_0) \right\} / (\log \gamma)^{1/2} \to \infty. \tag{5.20}$$

Hence the moving average rule (5.19) is asymptotically as efficient as the CUSUM rule as $\gamma \to \infty$ when the window size $m_\gamma$ satisfies (5.20).

In Sect. 3.6.4 we have considered Bayes sequential tests of a simple null hypothesis $f_0$ versus a simple alternative hypothesis $f_1$ and shown that they are SPRTs. In this simple versus simple case, the Bayesian problem of sequential change-point can be formulated by putting a prior distribution on $\nu$. In particular, if $\nu$ has a geometric prior distribution with $P(\nu = n) = p(1-p)^{n-1}$, $n = 1, 2, \ldots$, and there is a loss of 1 for a false alarm before $\nu$ and cost of $c$ for each observation taken after $\nu$, then the optimal stopping rule associated with the Bayes sequential detection problem has the explicit form

$$\begin{aligned} N_q(\gamma) &= \inf\left\{ n \geq 1 : P(\nu \leq n | X_1, \ldots, X_n) \geq \gamma/(\gamma + p^{-1}) \right\} \\ &= \inf\left\{ n \geq 1 : R_{q,n} \geq \gamma \right\}, \end{aligned} \tag{5.21}$$

where $q = 1 - p$ and $R_{q,n} = \sum_{k=1}^{n} \prod_{i=k}^{n} \{q^{-1} f_1(X_i)/f_0(X_i)\}$. Note that $P(\nu \leq n | X_1, \ldots, X_n) = R_{q,n}/(R_{q,n} + p^{-1})$. This was first derived by Shiryaev and subsequently modified by Roberts (1966) to

$$N(\gamma) = \inf\left\{ n \geq 1 : \lim_{q \to 1} R_{q,n} \geq \gamma \right\} = \inf\left\{ n \geq 1 : \sum_{k=1}^{n} \prod_{i=k}^{n} (f_1(X_i)/f_0(X_i)) \geq \gamma \right\}, \tag{5.22}$$

which Pollak (1985) proved to be asymptotically Bayes risk efficient as $p \to 0$ and also asymptotically minimax in the sense of (5.17) as $\gamma \to \infty$. Note that (5.21) involves the parameter $p$ in the prior geometric distribution for $\nu$. Letting $p \to 0$ corresponds to a flat prior and yields the Shiryaev–Roberts rule (5.22). Both the CUSUM and the Shiryaev–Roberts rules involve likelihood ratio statistics $\prod_{i=k}^{n}(f_1(X_i)/f_0(X_i))$, as in Wald's SPRT with $k=1$. On the other hand, whereas CUSUM uses the maximum over $1 \le k \le n$ for all possible change-times $k$ up to time $n$, the Bayesian approach (corresponding to a flat prior) sums over $1 \le k \le n$.

In Sects. 3.3 and 3.7.2, we have extended Wald's sequential testing theory to composite hypotheses. Lai (1995, Sect. 3.2) provides a similar extension for sequential change-point detection, assuming known baseline (in-control) parameters and unknown post-change parameters for the observations $X_t$ which he does not assume to be independent. A commonly used performance measure for quality control charts is the *average run length* (ARL), which is defined as $E_\theta(T)$ when the quality parameter remains at a fixed level $\theta$. This explains the background behind $\mathscr{F}_\gamma$ in (5.17), and Lai's first step toward this extension is to replace the ARL by more trackable and yet also sharper performance criteria. The ARL constraint $E_{\theta_0}(T) \ge \gamma$ stipulates a long expected duration to false alarm under the baseline parameter $\theta_0$. However, a large mean of $T$ does not necessarily imply that the probability of having a false alarm before some specified time $m$ is small. In fact, it is easy to construct positive integer-valued random variables $T$ with a large mean $\gamma$ and also having a high probability that $T=1$. This high probability of false alarm at the initial stage is clearly unacceptable, and the mean may to too crude a summary of the desired features of T under $P_{\theta_0}$. In practice, the system only fails after a very long in-control period, and we expect many false alarms before the first correct alarm. It is therefore much more relevant to consider:

(a) The probability of no false alarm during a typical (steady-state) segment of the baseline period
(b) The expected delay in signaling a correct alarm

instead of the ARL which is the mean duration to false alarm assuming a constant in-control or out-of-control parameter value.

It is straightforward to extend the CUSUM rule (5.16) to nonindependent and multivariate observations by simply replacing $f_j(X_i)$ in (5.16) by the conditional density $f_j(\boldsymbol{X}_i|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_{i-1})$ for $j=0,1$. In practice, these densities are usually modeled by parametric families. An obvious way to modify the CUSUM rule (5.16) for the case of $f_{\boldsymbol{\theta}}(\cdot|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{i-1})$ with unknown post-change parameter $\boldsymbol{\theta}$ is to estimate it by maximum likelihood, leading to the GLR rule

$$N_G = \inf\left\{ n : \max_{1\le k\le n} \sup_{\boldsymbol{\theta}\in\Theta} \sum_{i=k}^{n} \log \frac{f_{\boldsymbol{\theta}}(\boldsymbol{X}_i|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_{i-1})}{f_{\boldsymbol{\theta}_0}(\boldsymbol{X}_i|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_{i-1})} \ge c_\gamma \right\}. \tag{5.23}$$

For the problem of detecting shifts in the mean $\boldsymbol{\theta}$ of independent normal observations with known variance, this idea was proposed by Barnard (1959), but the statistical properties of the procedures remained a long-standing problem that was recently solved by Siegmund and Venkatraman (1995), whose asymptotic approximations to the ARL of the GLR rule under $\theta_0$ and under $\theta \neq \theta_0$ show that the GLR rule is asymptotically optimal in the sense of (5.17).

For practical implementation, the CUSUM rule (5.16) can be written in the recursive form $N = \inf\{n : \ell_n \geq c_\gamma\}$, where $\ell_n = \{\ell_{n-1} + \log(f_1(X_n)/f_0(X_n))\}^+$ with $\ell_0 = 0$. The GLR rule (5.23) does not have such convenient recursive forms, and the memory requirements and number of computations at time $n$ grow to infinity with $n$, which cannot be implemented for on-line fault detection in engineering systems. A natural modification to get around this difficulty is to replace $\max_{1 \leq k \leq n}$ in (5.23) by $\max_{n-M \leq k \leq n-\widetilde{M}}$; one may need a minimum sample size $\widetilde{M}$ for $n-k$ to be able to estimate the post-change parameter vector. Such window-limited GLR rules were first introduced by Willsky and Jones (1976) in the context of detecting changes in linear stochastic systems. Although these window-limited GLR rules have found widespread applications in fault detection of navigation and other control systems and in signal processing and tracking of maneuvering targets, how to choose $M$, $\widetilde{M}$, and $c_\gamma$ appropriately has remained a difficult open problem that was addressed by Lai (1995). Lai (1995) began by considering the simpler situation of detecting changes in the mean $\theta$ of independent normal observations $X_1, X_2, \ldots$ from a known baseline value $\theta = 0$. Here, the window-limited GLR rule has the form

$$N_W = \inf\left\{n : \max_{n-M \leq k \leq n} (X_k + \cdots + X_n)^2 / (2(n-k+1)) \geq c_\gamma\right\}, \tag{5.24}$$

and the methods of Siegmund and Venkatraman (1995) to analyze the GLR rule (5.23) in this independent normal case can be extended to (5.24). In particular, if we choose $M \sim \gamma$, then we have $E_0 N_W \sim E_0 N_G \sim K c_\gamma^{-1/2} e^{c_\gamma}$ as $c \to \infty$, where an explicit formula for $K$ is given in Siegmund and Venkatraman (1995). Therefore, choosing $c_\gamma = \log\gamma + \frac{1}{2}\log\log\gamma - \log K + o(1)$ gives $E_0 N_W \sim E_0 N_G \sim \gamma$. With this choice of $c_\gamma$, we also have $E_\theta N_W \sim E_\theta N_G \sim \min\{\gamma, (2\log\gamma)/\theta^2\}$ uniformly in $0 < |\theta| \leq (\log\gamma)^{1/2-\varepsilon}$ for every $\varepsilon > 0$. The choice $M = \gamma$ for the window size in (5.24) requires computation of $\gamma + 1$ quantities $(X_{n-i} + \cdots + X_n)^2/(i+1)$, $i = 0, \ldots, \gamma$, at every stage $n > \gamma$, and it is desirable to reduce the computational burden for large $\gamma$ by using a smaller window size. To develop efficient detection schemes that involve $O(\log\gamma)$ computations at every stage $n$, Lai (1995) used an idea similar to that in the theory of group sequential tests, which is to replace $\max_{0 \leq n-k \leq M}$ in (5.24) by $\max_{n-k+1 \in \mathscr{N}}$ where $\mathscr{N} = \{1, \ldots, M\} \cup \{[b^j M] : 1 \leq j \leq J\}$, with $M \sim a\log\gamma$, $b > 1$, and $J = \min\{j : [b^j M] \geq \gamma\} \sim (\log\gamma)/(\log b)$. Specifically, replacing $N_W$ by

$$\widetilde{N}_W = \inf\left\{n : \max_{k: n-k+1 \in \mathscr{N}} (X_k + \cdots + X_n)^2 / (2(n-k+1)) \geq c_\gamma\right\}, \tag{5.25}$$

Lai (1995) showed that $E_0(\widetilde{N}_W) \sim \widetilde{K} c_\gamma^{-1/2} e^{c_\gamma} \sim \gamma$ if $c_\gamma = \log\gamma + \frac{1}{2}\log\log\gamma - \log\widetilde{K} + o(1)$ and that $E_\theta(\widetilde{N}_W) \sim (2\log\gamma)/\theta^2$ if $|\theta| > \sqrt{2/a}$ while

$$E_\theta(\widetilde{N}_W) \leq (1+o(1))\min\{\gamma, (2b\log\gamma)/\theta^2\} \quad \text{uniformly in } 0 < |\theta| \leq \sqrt{2/a}.$$

Hence, choosing $b$ close to 1 (say $b$=1.1), there is little loss of efficiency in reducing the computational complexity of $N_G$ by its window-limited modification (5.25).

By using the likelihood ratio identity (3.3) and a change of measures, Lai (1998) proved an asymptotic lower-bound result for the detection delay subject to the ARL constraint $E_0(T) \geq \gamma$, similar to Hoeffding's lower bound for sequential testing, thereby proving the asymptotic optimality (5.17) of the CUSUM rule for nonindependent and multivariate observations, in which $I(f_1, f_0)$ is defined as the limit (which exists by the strong law for ergodic sequences) of the average of the log-likelihood ratio statistics. Instead of the ARL constraint $E_0(T) \geq \gamma$, he introduces a probability constraint of the form $\sup_{k\geq 1} P_0\{k \leq T < k+m\} \leq m/\gamma$ with $m = o(\gamma)$ but $m/\log\gamma \to \infty$. Then a similar change-of-measure argument gives an asymptotic lower bound for the detection delay of the form

$$E^{(k)}(T-k+1)^+ \geq \{P_0(T\geq k)/I(f_1,f_0) + o(1)\}\log\gamma, \text{ uniformly in } k \geq 1. \quad (5.26)$$

Lai (1995, 1998) has extended the window-limited GLR rule (5.25) from the normal case to general stochastic systems, but using the GLR statistics in (5.23) instead of those for the normal case in (5.25). Under certain stability assumptions on the stochastic systems, he has shown that these window-limited GLR rules attain the asymptotic lower bounds in the preceding paragraph and are therefore asymptotically efficient. Moreover, in Markov systems, these window-limited GLR rules $T$ satisfy

$$\sup_{k\geq 1} P_{\boldsymbol{\theta}_0}(k \leq T < k+m) \sim P_{\boldsymbol{\theta}_0}(T \leq m) \sim m/E_{\boldsymbol{\theta}_0}(T), \quad (5.27)$$

as $E_{\boldsymbol{\theta}_0}(T) \sim \gamma \to \infty$ and $m/\log\gamma \to \infty$ but $\log m = o(\log\gamma)$. Hence, to determine the threshold $c_\gamma$, the ARL constraint $E_{\boldsymbol{\theta}_0}(T) \doteq \gamma$ can be replaced by the probability constraint $P_{\boldsymbol{\theta}_0}(T \leq m) \doteq m/\gamma$, which is much more amenable to Monte Carlo computation since simulating $P_{\boldsymbol{\theta}_0}(T \leq m)$ involves far fewer random variables (no more than $m$ in each simulation run) than directly simulating $T$.

Lai and Xing (2010) have also extended window-limited rules to the case where pre- and post-change parameters are unknown in multiparameter exponential families and have established their asymptotic optimality. This provides a complete analog of sequential testing of composite hypotheses in Sect. 3.7.3 for sequential change-point detection, showing the important roles of GLR statistics and the asymptotic optimality of the associated detection rules.

### 5.5.2 Sequential Detection in Public Health Surveillance

Timely detection of adverse health events and adoption of public health policies to rectify the situation or prevent repeated occurrences are beneficial to the affected individuals and the society. This involves systematic collection, analysis, and interpretation of large amounts of outcome-specific data by national public health programs in different countries and international networks; see Sonesson and Bock (2003, pp. 5–6), who also refer to the literature on retrospective analysis of these data to estimate disease prevalence or to compare disease patterns in different regions. They point out, however, that there are many situations where the sequentially accumulated data can be used prospectively to detect quickly an increased incidence of a disease so that timely rectifying actions can be taken. They note that while much of the research on statistical surveillance originated from engineering applications, as we have reviewed in the preceding section, "the context of public health surveillance implies specific problems that are not generally present in the case of industrial production control." These include problems of seasonal effects and reporting delays, inherent differences among diseases (such as chronic conditions versus acute infections), and monitoring not only cases of disease but also risk factors. Of particular importance in their review of the sequential change-point detection literature for public health surveillance are (a) detection of a changed intensity in a Poisson process in their Sect. 5.4 and (b) multivariate surveillance methods and their modifications for spatial surveillance.

### 5.5.3 A Hybrid of Sequential Testing and Detection for Pharmacovigilance

Sections 5.1.2 and 5.4 have described the sequential testing approach to post-marketing surveillance of vaccine and drug safety. Note that the null hypothesis actually assumes the medical product is safe, and its rejection means that there is enough evidence against that assumption, hence also against the medical product. Such a formulation of the null hypothesis in fact extends to prelicensure safety testing as in Sects. 5.2 and 5.3. This is unlike the sequential testing for efficacy studied in Chap. 4, in which the null hypothesis assumes that the new treatment is no better than the control, and only rejection of the null hypothesis would result in regulatory approval of the new treatment. The approval allows beneficial claims of the drug or vaccine in its labeling; this explains why the null hypothesis $H_0$ takes the form that there is no such benefit so that rejection of $H_0$ suggests enough evidence to support the claim. For safety testing, the manufacturer of the medical product does not usually make a claim about its safety but is required by the regulatory agency, which does not take the position that it is unsafe, to collect data about potential adverse events. Prelicensure randomized clinical trials for a new vaccine have to first establish that the vaccine is efficacious and group sequential tests for efficacy are often used. If the efficacy bar is passed, the next stage is to test for safety, which may involve a very large sample size in order to detect rare serious adverse events. Therefore, the regulatory position is to presume safety of the vaccine unless proven otherwise in a large randomized trial. This position implicitly assumes that there is further post-marketing evaluation since acceptance of the null hypothesis does not have a prescribed probability guarantee that it (i.e., that the vaccine is safe) is indeed true.

Post-marketing safety monitoring of an approved medical product is a continual process in view of the active surveillance systems organizing and disseminating routinely collected electronic healthcare data. In this way it is similar to sequential detection that involves continual follow-up until faults are detected. On the other hand, unlike change-point detection, the change-time is known and does not need to be estimated because it is when the medical product is approved and put to the market. Lai, Shih, and Hock Peng Chan are developing a hybrid of sequential testing and sequential fault detection that entails continual safety monitoring of a vaccine or drug using observational data from VSD or Sentinel or other electronic healthcare databases. The methodology involves adjustments for confounding from observational data, GLR statistics for generalized linear models (including Poisson and logistic regressions), propensity scores and inverse probability weighting, and control of false alarm and false discovery rates. This work will be posted at the book's website when it is complete.

## 5.6 Supplements and Problems

1. *The CMaxSPRT test.*
   We give here more details of the CMaxSPRT described in Sect. 5.1.2, showing in particular the explicit form of the test statistic obtained by conditioning the maximized likelihood ratio on the cumulative person-time in the historical cohort. In the MaxSPRT defined by (5.3) and (5.4), the number of adverse events of the surveillance group is considered to be random while the cumulative person-time or the cumulative number of vaccinations is considered to be fixed. Hence the expected number of adverse events under the null is assumed to be a known function of the cumulative person-time and some potential confounders such as age, sex, and site. For the CMaxSPRT, Li and Kulldorff (2010) condition on the numbers of adverse events in the historical data and the surveillance group and regard the cumulative person-time taken to observe the given number of adverse events as the random variable. Specifically, let $c$ and $Q$ denote the total number of adverse events and the total person-time in the historical data, respectively, and let $T_k$ denote the cumulative person-time since the beginning of the surveillance until the $k$th adverse event. Here, $c$ and $k$ are fixed numbers while $Q$ and $T_k$ are random. The intuitive idea is simply to compare the event rates between the historical cohort and the surveillance group. To fix ideas, assume

the numbers of events over the cumulative person-time in the historical cohort and the surveillance group are homogeneous Poisson processes with the event rates denoted by $\lambda_h$ and $\lambda_v$, respectively. Instead of working with the Poisson process, Li and Kulldorff (2010) work with the exponential interarrival times, or the "inter-event person-times" as they call them, like what we have done in Sect. 5.1.2. Thus, the cumulative person-time $T_k$ in the surveillance group since the beginning of the surveillance until the $k$th event is a sum of $k$ i.i.d. random variables from the exponential distribution with rate $\lambda_v$ and therefore, has a gamma distribution with shape parameter $k$ and scale $1/\lambda_v$. Right after the $k$th adverse event in the surveillance group, the likelihood function is

$$L_k = \lambda_h^c e^{-\lambda_h Q} \lambda_v^k e^{-\lambda_v T_k} = \lambda_h^c \lambda_v^k e^{-(\lambda_h Q + \lambda_v T_k)}.$$

The null hypothesis $\lambda_h = \lambda_v$ is composite as the common value is unknown. Therefore the logarithm of the ratio of the maximized likelihood of the composite alternative to the composite null hypotheses is

$$\begin{aligned} U_k &= \log \left( \frac{\max_{\lambda_v \geq \lambda_h} e^{-\lambda_h Q - \lambda_v T_k} \lambda_h^c \lambda_v^k}{\max_\lambda e^{-\lambda(Q+T_k)} \lambda^{c+k}} \right) \\ &= I_{\{k/c > T_k/Q\}} \log \left( \frac{e^{-c}(c/Q)^c e^{-k}(k/T_k)^k}{e^{-(k+c)}[(c+k)/(Q+T_k)]^{c+k}} \right). \end{aligned}$$

(a) Show that conditional on $c$, the only random part of $U_k$ is the ratio $T_k/Q$.

(b) Show that under the composite null hypothesis $H_0 : \lambda_h = \lambda_v = \lambda$ with unknown $\lambda$, the conditional distribution of $T_k/Q$ given $c$ does not depend on $\lambda$. Hence, show that under $H_0$, the joint distribution of $(U_1, \ldots, U_k)$ depends only on $c$ and $k$.

2. *Implementation of CMaxSPRT.*
A maximum number $K$ of adverse events from the surveillance group is specified. If $K$ is reached, the test stops; this serves as a truncation bound for CMaxSPRT. The test is based on the conditional distribution of $(U_1, \ldots, U_K)$ given the total number $c$ of adverse events in the historical data, which does not depend on the unknown value of $\lambda_h$ and $\lambda_v$ under the null hypothesis and can be determined by Monte Carlo simulations. Tables of critical values are given in Li and Kulldorff (2010) for $\alpha = 0.05$ and different values of $c$ and $K$.

3. *Adjustments for confounding.*
Li and Kulldorff (2010, Sect. 4.5) note that the assumption of homogeneous Poisson arrivals with event rates $\lambda_h$ and $\lambda_v$ over the historical and surveillance cohorts may be overly restrictive in view of population heterogeneity; for example, men and women may have different event rates. To adjust for confounding, they propose to stratify the entire population into several subgroups that are likely to have different risks for adverse events and assign different weights to

the person-time from different subgroups; the weights are chosen "using either subject-matter expertise and/or published results from previous studies."

4. *Comparison with randomized clinical trials.*
In Sect. 5.2.1, we also use a conditioning argument for the interarrival times of adverse events for the exposed and unexposed cohorts, $V$ and $C$ in that case. The conditioning argument in (5.8) is much simpler and also transforms the testing problem for rare Poisson rates to that of testing a Bernoulli proportion. The setting of a randomized clinical trial is pivotal to that argument, in which we consider successive adverse events of the combined $V$ and $C$ groups.

The conditioning argument in CMaxSPRT, therefore, is very different from that in Sect. 5.2.1. On the other hand, it is similar in principle to that of Fisher's exact test in a $2 \times 2$ contingency table: Conditional on the margins of the table, the count in a cell (which is the test statistic used) has a hypergeometric distribution under the composite null hypothesis. Li and Kulldorff (2010) want to use the maximized likelihood ratio statistic $U_k$ as the test statistic of the composite null hypothesis $\lambda_h = \lambda_v$. By conditioning on the number of adverse events $c$ in the historical cohort up to the $k$th adverse event of the surveillance group, the distribution of $U_k$ is completely specified under the null hypothesis. In this connection, note that the conditioning argument in (5.8) also reduces the composite null hypothesis $\lambda_V = \lambda_C$ to a simple one: $p_0 = 1/2$.

# Chapter 6
# Time-Sequential Design of Clinical Trials with Failure-Time Endpoints

As mentioned in Sect. 1.3, sequential statistical methods received little attention from the biomedical community until the early termination of the Beta-Blocker Heart Attack Trial (BHAT) in 1981. In Sects. 6.2 and 6.3, we describe the design details, interim analyses, and final report of BHAT. As the primary endpoint of BHAT was survival following at least one heart attack, we give an overview of survival analysis in Sect. 6.1. However, because of the sequential nature of interim analyses, traditional survival analysis methodology reviewed in Sect. 6.1 has to be extended to the sequential setting. This sequential setting also differs from that in Chap. 4 in that it is *time-sequential* rather than group sequential. In fact, when BHAT stopped at the sixth interim analysis, all subjects had been accrued prior to the fourth analysis. Therefore, unlike the group sequential trials in Chap. 4, early stopping did not reduce the sample size; what was reduced was the study duration. Extension of conventional survival analysis to the time-sequential setting, therefore, involves two time scales to measure the covariance structure of the time-sequential test statistics. One is the amount of information, as in conventional survival analysis, accumulated by the time of an interim analysis. The other is calendar time, and the covariance structure depends on the calendar times at which the interim analyses are conducted. In Sects. 6.4 and 6.5, we describe several important developments in group sequential and time-sequential methods following BHAT. Some recent innovative designs are summarized as supplements in Sect. 6.7. An appendix on martingale theory that provides important tools for conventional survival analysis and its extensions to the time-sequential setting is given in Sect. 6.6.

## 6.1 An Overview of Survival Analysis

Survival analysis is concerned with the statistical analysis of the failure time $\tau$ of an individual from a homogeneous population, or with the comparison of the failure times of two populations (one receiving a new treatment and the other a standard

J. Bartroff et al., *Sequential Experimentation in Clinical Trials,*
Springer Series in Statistics 298, DOI 10.1007/978-1-4614-6114-2_6,

one or placebo), or with regression analysis of $\tau$ on covariates. An important feature of survival analysis is that some subjects in the study may not fail during the observation period or may have withdrawn from the study during the period. Thus, the data from these subjects are right censored. This makes the statistical analysis much more complicated than the case in which all failure times in the sample are fully observable. The time-sequential setting further increases the complexity considerably. A theoretical tool to simplify the analytic calculations is martingale theory. The overview takes advantage of this tool whose background is given in Sect. 6.6.

### 6.1.1 Nelson–Aalen and Kaplan–Meier Estimators

Let $\tau_1, \ldots, \tau_n$ be $n$ independent failure times with common distribution $F$. If $F$ is absolutely continuous with density function $f$, then

$$\lambda(t) = \lim_{h \to 0+} P(t \le \tau_i < t+h \,|\, \tau \ge t)/h = f(t)/(1-F(t))$$

is called the hazard (or intensity) function; its integral $\Lambda(t) = \int_0^t \lambda(s)ds$ is called the cumulative hazard function. More generally, the cumulative hazard function is defined by

$$\Lambda(t) = \int_0^t \frac{dF(s)}{1-F(s-)},$$

where $F(s-) = \lim_{t \to s-} F(t)$.

The $\tau_i$ are subject to right censoring, which may be due to withdrawal or the restricted length of the survival study. Thus, there are censoring variables $c_i$, representing the time in the study during the observation period. The observations, therefore, are $(T_i, \delta_i), i = 1, \cdots, n$, where $T_i = \min(\tau_i, c_i)$ and $\delta_i = I_{\{\tau_i \le c_i\}}$ is the censoring indicator that indicates whether $T_i$ is an actual failure time or is censored. Subject $i$ is "at risk" at time $s$ if $T_i \ge s$ (i.e., has not failed and not withdrawn prior to $s$). Let

$$Y(s) = \sum_{i=1}^n I_{\{T_i \ge s\}}, \quad N(s) = \sum_{i=1}^n I_{\{T_i \le s,\ \delta_i = 1\}}. \tag{6.1}$$

Note that $Y(s)$ is the risk set size and $\Delta N(s) = N(s) - N(s-)$ is the number of observed deaths at time $s$. The *Nelson–Aalen estimator* of the cumulative hazard function $\Lambda(t)$ is

$$\hat{\Lambda}(t) = \sum_{s \le t} \frac{\Delta N(s)}{Y(s)} = \int_0^t \frac{I_{\{Y(s)>0\}}}{Y(s)} dN(s), \tag{6.2}$$

where we use the convention $0/0 = 0$. It is shown in Sect. 6.6 that

$$\left\{ \int_0^t U(s)[dN(s) - Y(s)d\Lambda(s)],\ t \ge 0 \right\} \text{ is a martingale,} \tag{6.3}$$

for every left-continuous stochastic process $U(s)$; note that $Y(s)$ is left continuous. Suppose $F$ is continuous. It then follows from the martingale central limit theorem (CLT) that as $n \to \infty$,

$$(\hat{\Lambda}(t) - \Lambda(t)) \Big/ \left\{ \int_0^t \frac{I_{\{Y(s)>0\}}}{Y^2(s)} dN(s) \right\}^{1/2} \xrightarrow{\mathscr{D}} N(0,1), \tag{6.4}$$

where $\xrightarrow{\mathscr{D}}$ denotes convergence in distribution.

Partition time into disjoint intervals $I_1 = (0,t_1], I_2 = (t_1,t_2]$, etc. A life table in actuarial science summarizes the mortality results of a large cohort of $n$ subjects as follows:

$n_j =$ number of subjects alive at the beginning of $I_j$,
$d_j =$ number of deaths during $I_j$,
$l_j =$ number lost to follow-up during $I_j$.

It estimates $p_j = P(\text{died during } I_j \mid \text{alive at the beginning of } I_j)$ by $\hat{p}_j = d_j/(n_j - l_j)$. The actuarial (life-table) estimate of $P(\tau > t_k)$ is the product

$$\prod_{j=1}^{k}(1 - \hat{p}_j) = \prod_{j=1}^{k}\left(1 - \frac{d_j}{n_j - l_j}\right).$$

Without discretizing the failure times, we can likewise estimate $S(t) = P(\tau > t)$ by

$$\hat{S}(t) = \prod_{s \le t}\left(1 - \frac{\Delta N(s)}{Y(s)}\right). \tag{6.5}$$

Since $N(s)$ has at most $n$ jumps, the product in (6.5) has finitely many factors. The estimator $\hat{S}$ is called the *Kaplan–Meier* estimator of $S$. Note that $\hat{S}(t) = \prod_{s \le t}(1 - \Delta\hat{\Lambda}(s))$ by (6.2) and (6.5). In Sect. 6.7, it is shown that analogous to (6.5),

$$S(t) = \prod_{s \le t}(1 - d\Lambda(s)) \quad \text{for all } t \text{ such that } \Lambda(t) < \infty. \tag{6.6}$$

The product in (6.6) is called the *product integral* of the nondecreasing right-continuous function $\Lambda(\cdot)$ on $[0,t]$; it is defined by the limit of $\prod\{1 - [\Lambda(t_i) - \Lambda(t_{i-1})]\}$, where $0 = t_0 < t_1 < \cdots < t_m = t$ is a partition of $[0,t]$ and the product is in the natural order from left to right, as the mesh size $\max_{1 \le i \le m} |t_i - t_{i-1}|$ approaches 0. Moreover, it is also shown in Sect. 6.6 that

$$(\hat{S}(t) - S(t)) \Big/ \left[ \hat{S}(t) \left\{ \int_0^t \frac{I_{\{Y(s)>0\}}}{Y^2(s)} dN(s) \right\}^{1/2} \right] \xrightarrow{\mathscr{D}} N(0,1) \tag{6.7}$$

by the martingale CLT.

### 6.1.2 Regression Models for Hazard Functions with Covariates

We have focused so far on the estimation of the survival distribution of a failure time $\tau$. In applications, one often wants to use a model for $\tau$ to predict future failures from a vector $\mathbf{x}(t)$ of predictors based on current and past observations; $\mathbf{x}(t)$ is called a time-varying (or time-dependent) covariate. When $\mathbf{x}(t) = \mathbf{x}$ does not depend on $t$, it is called a time-independent (or baseline) covariate. In practice, some predictors may be time independent, while other components of $\mathbf{x}(t)$ may be time-varying. Since prediction of future survival using $\mathbf{x}(t)$ is relevant only if $\tau > t$, one is interested in modeling the conditional distribution of $\tau$ given $\tau > t$, for example, by relating the hazard function $\lambda(t)$ to $\mathbf{x}(t)$. Cox (1972) introduced the *proportional hazards* (or *Cox regression*) model

$$\lambda(t) = \lambda_0(t)\exp(\boldsymbol{\beta}^T\mathbf{x}(t)). \tag{6.8}$$

Putting $\mathbf{x}(t) = 0$ in (6.8) shows that $\lambda_0(\cdot)$ is also a hazard function; it is called the *baseline hazard.*

Instead of assuming a parametric model to estimate the baseline hazard function $\lambda_0(\cdot)$ as done in the previous literature, Cox (1972, 1975) introduced a semiparametric method to estimate the finite-dimensional parameter $\boldsymbol{\beta}$ in the presence of an infinite-dimensional nuisance parameter $\lambda_0(\cdot)$; it is semiparametric in the sense of being nonparametric in $\lambda_0$ but parametric in $\boldsymbol{\beta}$. Cox's partial likelihood method decomposes the likelihood function into two factors, with one involving only $\boldsymbol{\beta}$ and the other involving both $\boldsymbol{\beta}$ and the baseline cumulative hazard function $\Lambda_0$. It estimates $\boldsymbol{\beta}$ by maximizing the *partial likelihood*, which is the first factor that only involves $\boldsymbol{\beta}$ and is described below. Order the observed censored failure times as $\tau_{(1)} < \cdots < \tau_{(m)}$, with $m \le n$. Let $C_j$ denote the set of censored $T_i$'s in the interval $[\tau_{(j-1)}, \tau_{(j)})$, and let $(j)$ denote the individual failing at $\tau_{(j)}$, noting that with probability 1 there is only one failure at $\tau_{(j)}$ because the failure time distributions have density functions. Let $R_{(j)} = \{i : T_i \ge \tau_{(j)}\}$ denote the risk set at $\tau_{(j)}$. Then

$$\begin{aligned} P\left\{(j)\,|\,C_j, (l), C_l, 1 \le l \le j-1\right\} &= P\left\{(j) \text{ fails at } \tau_{(j)}\,|\,R_{(j)}, \text{ one failure at } \tau_{(j)}\right\} \\ &= \exp\left(\boldsymbol{\beta}^T\mathbf{x}_j(\tau_{(j)})\right) \Big/ \sum_{i\in R_{(j)}} \exp\left(\boldsymbol{\beta}^T\mathbf{x}_i(\tau_{(j)})\right). \end{aligned} \tag{6.9}$$

The partial likelihood is $\prod_{j=1}^m P\{(j)\,|\,C_j, (l), C_l, 1 \le l \le j-1\}$. Ignoring the other factors $P\{C_{j+1}|(l), C_l, 1 \le l \le j\}$ in the likelihood function, Cox's regression estimator $\hat{\boldsymbol{\beta}}$ is the maximizer of the partial log-likelihood

$$l(\boldsymbol{\beta}) = \sum_{j=1}^m \left\{\boldsymbol{\beta}^T\mathbf{x}_j(\tau_{(j)}) - \log\left(\sum_{i\in R_{(j)}} \exp\left(\boldsymbol{\beta}^T\mathbf{x}_i(\tau_{(j)})\right)\right)\right\}, \tag{6.10}$$

or equivalently, the solution of $\frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta}) = 0$.

Letting $w_i(\boldsymbol{\beta}) = e^{\boldsymbol{\beta}^T \mathbf{x}_i(\tau_{(j)})} / \sum_{l \in R_{(j)}} e^{\boldsymbol{\beta}^T \mathbf{x}_l(\tau_{(j)})}$ for $i \in R_{(j)}$, note that

$$\frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta}) = \sum_{j=1}^{m} \left\{ \mathbf{x}_j(\tau_{(j)}) - \sum_{i \in R_{(j)}} w_i(\boldsymbol{\beta}) \mathbf{x}_i(\tau_{(j)}) \right\}, \tag{6.11}$$

$$\begin{aligned} -\left( \frac{\partial^2}{\partial \beta_k \partial \beta_h} l(\boldsymbol{\beta}) \right)_{k,h} = \sum_{j=1}^{m} \Bigg\{ & \sum_{i \in R_{(j)}} w_i(\boldsymbol{\beta}) \mathbf{x}_i(\tau_{(j)}) \mathbf{x}_i^T(\tau_{(j)}) \\ & - \left( \sum_{i \in R_{(j)}} w_i(\boldsymbol{\beta}) \mathbf{x}_i(\tau_{(j)}) \right) \left( \sum_{i \in R_{(j)}} w_i(\boldsymbol{\beta}) \mathbf{x}_i(\tau_{(j)}) \right) \Bigg\}^T . \end{aligned} \tag{6.12}$$

Since $\sum_{i \in R_{(j)}} w_i(\boldsymbol{\beta}) = 1$, we can interpret the term $\bar{\mathbf{x}}(\tau_{(j)}) := \sum_{i \in R_{(j)}} w_i(\boldsymbol{\beta}) \mathbf{x}_i(\tau_{(j)})$ in (6.11) and (6.12) as a weighted average of covariates over the risk set. Each summand in (6.11) therefore compares the covariate at an observed failure to its weighted average over the risk set. Moreover, each summand in (6.12) can be expressed as a sample covariance matrix of the form

$$\sum_{i \in R_{(j)}} w_i(\boldsymbol{\beta}) \{\mathbf{x}_i(\tau_{(j)}) - \bar{\mathbf{x}}(\tau_{(j)})\} \{\mathbf{x}_i(\tau_{(j)}) - \bar{\mathbf{x}}(\tau_{(j)})\}^T . \tag{6.13}$$

Making use of martingale theory, Cox's regression estimator $\boldsymbol{\beta}$ can be shown to satisfy the usual asymptotic properties of maximum likelihood estimates even though partial likelihood is used; see Sect. 6.6. In particular, it can be shown that as $n \to \infty$,

$$(-\ddot{l}(\hat{\boldsymbol{\beta}}))^{1/2} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \text{ has a limiting standard normal distribution,} \tag{6.14}$$

where we use $\dot{l}(\boldsymbol{\beta})$ to denote the gradient vector $(\partial / \partial \boldsymbol{\beta}) l(\boldsymbol{\beta})$ and $\ddot{l}(\boldsymbol{\beta})$ to denote the Hessian matrix of second partial derivatives $(\partial^2 / \partial \beta_k \partial \beta_h) l(\boldsymbol{\beta})$, given by (6.12). One can perform usual likelihood inference treating the partial likelihood as a likelihood function and apply likelihood-based selection of covariates. Moreover, even though $\hat{\boldsymbol{\beta}}$ is based on partial likelihood, it has been shown to be asymptotically efficient.

When there are no covariates, $\Lambda = \Lambda_0$ can be estimated by the Nelson–Aalen estimator (6.2). Note that (6.2) has jumps only at uncensored observations and that $Y(s)$ is the sum of 1's over the risk set $\{i : T_i \geq s\}$ at $s$. When $\tau_i$ has hazard function $\exp(\boldsymbol{\beta}^T \mathbf{x}_i(s)) \lambda_0(s)$, we modify $Y(s)$ to

$$Y(s) = \sum_{i \in R_{(j)}} \exp(\boldsymbol{\beta}^T \mathbf{x}_i(s)) \ \text{ at } s = \tau_{(j)}, \tag{6.15}$$

using the same notation as in (6.9)–(6.13). The Breslow estimator of $\Lambda_0$ in the proportional hazards regression model (6.8) is again given by (6.2) but with $Y(s)$ defined by (6.15).

### 6.1.3 Rank Tests Based on Censored Survival Data

In randomized clinical trials with survival endpoints, a primary objective is to compare time to failure between two treatment groups $X$ and $Y$. Suppose that the failure times $X_1, \dots, X_{n'}$ are independent having a common distribution function $F$ and the failure times $Y_1, \dots, Y_{n''}$ are independent having a common distribution function $G$. Let $n = n' + n''$. To test the null hypothesis $H_0 : F = G$ or $H_0' : F \le G$, a commonly used method is to evaluate the ranks $R_i$ of $X_i$ $(i = 1, \dots, n')$ in the combined sample $X_1, \dots, X_{n'}, Y_1, \dots, Y_{n''}$ and to use rank statistics of the form $\ell_n = \sum_{i=1}^{n'} \varphi(R_i/n)$, where $\varphi : (0,1] \to (-\infty, \infty)$. However, because of censoring, one cannot compute $\ell_n$ in these situations. As noted in Gu et al. (1991), a natural extension of $\ell_n$ to censored data is the censored rank statistic of the form

$$S_n = \sum_{k=1}^{K} \psi\big(H_n(Z_{(k)}-)\big)(z_k - m_k/\#_k), \tag{6.16}$$

where $Z_{(1)} \le \cdots Z_{(K)}$ denote the ordered uncensored observations in the combined sample, $z_k = 1$ if $Z_{(k)}$ is an $X$ and $z_k = 0$ if $Z_{(k)}$ is a $Y$, $\#_k$ (resp. $m_k$) denotes the number of observations (resp. $X$'s) in the combined sample that are $\ge Z_{(k)}$, $H_n$ is the Kaplan–Meier estimate (6.5) based on the combined sample, and $\psi$ is related to $\varphi$ by the relation

$$\psi(u) = \varphi(u) - (1-u)^{-1} \int_u^1 \varphi(t)\,dt, \quad 0 < u < 1. \tag{6.17}$$

Taking $\psi(u) = (1-u)^\rho$ $(\rho \ge 0)$ yields the $G^\rho$ statistics proposed by Harrington and Fleming (1982). The case $\rho = 0$ corresponds to Mantel's (1966) logrank statistic, which is a special case of Cox's score statistic (6.11) at $\beta = 0$ since the covariate $z_k$ is binary, and the case $\rho = 1$ corresponds to the generalization of Wilcoxon's statistic by Peto and Peto (1972) and Prentice (1978). Making use of martingale theory, it can be shown that (6.16) is asymptotically normal under the null hypothesis $F = G$; see Sect. 6.6.

## 6.2 The Beta-Blocker Heart Attack Trial (BHAT)

### 6.2.1 Trial Design

The primary objective of BHAT was to determine whether regular, chronic administration of propranolol, a beta-blocker, to patients who had at least one documented myocardial infarction (MI) would result in significant reduction in mortality from all causes during the follow-up period. It was designed as a multicenter, double-blind, randomized placebo-controlled trial with a projected total of 4200 eligible patients recruited within 21 days of the onset of hospitalization for MI. The trial was planned to last for 4 years, beginning in June 1978 and ending in June 1982, with patient accrual completed within the first 2 years so that all patients could be followed for a period of 2–4 years. The sample size calculation was based on a 3-year mortality rate of 18% in the placebo group and a 28% reduction of this rate in the treatment group, with a significance level of 0.05 and 0.9 power using a two-sided logrank test; see Beta-Blocker Heart Attack Trial Research Group (1984, p. 388). In addition, periodic reviews of the data were planned to be conducted by a Data and Safety Monitoring Board (DSMB), roughly once every 6 months beginning at the end of the first year, whose functions were to monitor safety and adverse events and to advise the Steering and Executive Committees on policy issues related to the progress of the trial.

### *6.2.2 Trial Execution and Interim Analysis by DSMB*

The actual recruitment period was 27 months, within which 3837 patients were accrued from 136 coronary care units in 31 clinical centers, with 1916 patients randomized into the propranolol group and 1921 into the placebo group. Although the recruitment goal of 4200 patients had not been met, the projected power was only slightly reduced to 0.89 as accrual was approximately uniform during the recruitment period.

The DSMB arranged meetings at 11, 16, 21, 28, 34, and 40 months to review the data collected so far, before the scheduled end of the trial at 48 months. Besides monitoring safety and averse events, the DSMB also examined the normalized logrank statistics to see whether propranolol was indeed efficacious. The successive values of these statistics are listed below:

| Time (months) | 11 | 16 | 21 | 28 | 34 | 40 |
|---|---|---|---|---|---|---|
| Test statistic | 1.68 | 2.24 | 2.37 | 2.30 | 2.34 | 2.82 |

Instead of continuing the trial to its scheduled end at 48 months, the DSMB recommended terminating it in their last meeting because of conclusive evidence in favor of propranolol. Their recommendation was adopted, and the trial was terminated on October 2, 1981. It drew immediate attention of the biopharmaceutical community to the benefits of sequential methods, not because it reduced the number of patients but because it shortened a 4-year study by 8 months, with positive results for a long-awaited treatment supporting its immediate use.

Note that except for the first interim analysis at 11 months (when there were 16 deaths out of 679 patients receiving propranolol and 25 deaths out of 683 patients receiving placebo), all interim analyses showed normalized logrank statistics exceeding the critical value of 1.96 for a single 5% two-sided logrank test. The lack of significance in the first interim analysis seems to be due to the relatively small number of deaths. In comparison, the last interim analysis at 40 months had 135 deaths in the propranolol group of 1916 patients and 183 deaths in the placebo group of 1921 patients. The final report in Beta-Blocker Heart Attack Trial Research Group (1982) showed more deaths from both groups (138 and 188) due to additional data that were processed after the interim analysis. The Kaplan–Meier estimates of the respective survival functions in this final report show that the mortality distributions were estimable only up to approximately their tenth percentiles, with the cumulative distribution function for propranolol below that of placebo.

The critical value of 1.96 for the standardized logrank statistic only applies to a single analysis. To account for repeated testing, the DSMB used an adjustment (which has a critical value of 5.46 at the first analysis and 2.23 at the sixth analysis) for repeated significance testing with independent, identically distributed normal observations proposed in 1979 by O'Brien and Fleming (1979). Since logrank statistics (rather than normal observations) were actually used, the Beta-Blocker Heart Attack Trial Research Group's (1982) final report of the trial appealed to joint asymptotic normality of time-sequential logrank statistics that was established by Tsiatis (1981) shortly before that.

### *6.2.3 Stochastic Curtailment*

The preceding paragraph shows that time-sequential methodology, which was only at its infancy at that time, was barely adequate to handle the BHAT data. Moreover, the trial had been designed as a fixed-duration (instead of time-sequential) trial. The DSMB used some informal arguments based on *stochastic curtailment* described below, together with the formal group sequential test described in the preceding paragraph, to come to the conclusion that the propranolol therapy was indeed effective, at the time of the sixth interim analysis.

Stochastic curtailment, which was developed during the process of monitoring BHAT and was later described by Lan et al. (1982), is based on the *conditional power*, which is the conditional probability of rejecting the null hypothesis at the scheduled end of the trial given the current data, along with some speculation about the future data. The setting assumed was that of a Wiener process $W(v)$, $0 \le v \le 1$, with drift coefficient $\mu$. Consider the one-sided fixed sample size test of $H_0 : \mu = 0$ versus $H_1 : \mu = \mu_1$ ($> 0$) based on $W(1)$ with type I error probability $\alpha$ and type II error probability $\tilde{\alpha}$. Since the conditional distribution of $W(1)$ given $\{W(v),\ v \le s\}$ is normal with mean $W(s) + \mu(1-s)$ and variance $1-s$, the conditional power at $\mu$ given $\{W(v),\ v \le s\}$ is

$$\beta_s(\mu) = 1 - \Phi\left((1-s)^{-1/2}\left\{\Phi^{-1}(1-\alpha) - W(s) - \mu(1-s)\right\}\right), \qquad (6.18)$$

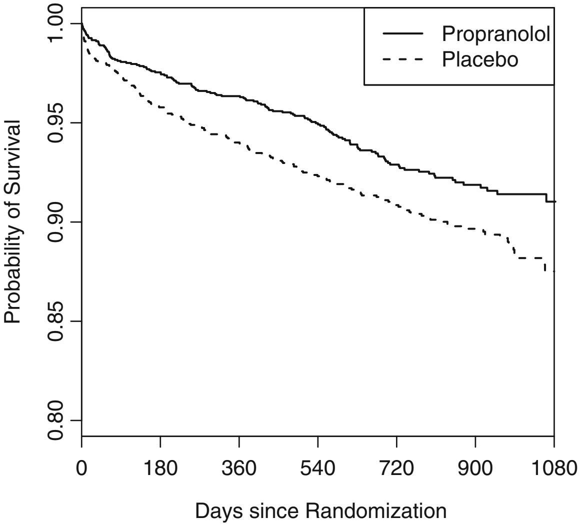
where $\Phi$ is the standard normal distribution function. Lan et al. (1982) proved the following basic result for stochastic curtailment for testing simple hypotheses concerning the drift of a Wiener process. They also appealed to the CLT in extending this argument to asymptotically normal statistics.

**Theorem 6.1.** *Suppose one curtails the preceding fixed sample size test of the drift of a Wiener process by stopping and rejecting $H_0$ when $\beta_s(0) > \rho$ and stopping and rejecting $H_1$ when $\beta_s(\mu_1) < 1 - \tilde{\rho}$, for some $\rho$ and $\tilde{\rho}$ less than but near 1. Then the type I error probability of the test is $\leq \alpha/\rho$, while the type II error probability is $\leq \tilde{\alpha}/\tilde{\rho}$.*

For BHAT, at the sixth interim analysis, the conditional power (6.18) under the null trend was found to range from 0.8 (for 120 additional deaths) to 0.94 (for 60 additional deaths) and to be 0.89 for the projected number of 80 additional deaths. In view of Theorem 6.1, the DSMB concluded that the nominal type I error probability would not be inflated by much if the test should be stochastically curtailed in this manner. The relevance of the Wiener process to the sequentially monitored logrank statistics will be explained in Sect. 6.5.1 that gives the asymptotic joint distribution of sequentially computed logrank statistics under the null hypothesis and under local alternatives.

## 6.3 Terminal Analyses of BHAT Data

The final report of the study was published in Beta-Blocker Heart Attack Trial Research Group (1982), summarizing various results of the final analysis of all the BHAT data that had been collected up to October 2, 1981, when official patient follow-up was stopped. After an average follow-up of 25.1 months, 138 patients in the propranolol group (7.2%) and 188 in the placebo group (9.8%) had died. Figure 6.1 shows the Kaplan–Meier curves by treatment group. The estimated survival curve of the propranolol group was above that of the placebo group. These curves also suggest departures from the proportional hazards model. Using Müller and Wang's (1994) kernel-based estimator with locally optimal bandwidths to estimate the hazard functions, the estimated hazard rates over time are plotted in Fig. 6.2a, which shows largest decrease of hazard rates during the first month after a heart attack for both the propranolol and placebo groups and that the propranolol group has a markedly smaller hazard rate than the placebo group within the first year, after which the hazard rates tend to stabilize. Figure 6.2b plots the hazard ratio of propranolol to placebo over time, and it shows that propranolol has the largest survival benefit over placebo in the first 9 months after a heart attack.



No. at Risk

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Propranolol | 1916 | 1868 | 1844 | 1504 | 1087 | 661 | 203 |
| Placebo | 1921 | 1838 | 1804 | 1464 | 1077 | 658 | 211 |

**Fig. 6.1** The Kaplan–Meier curves by treatment group

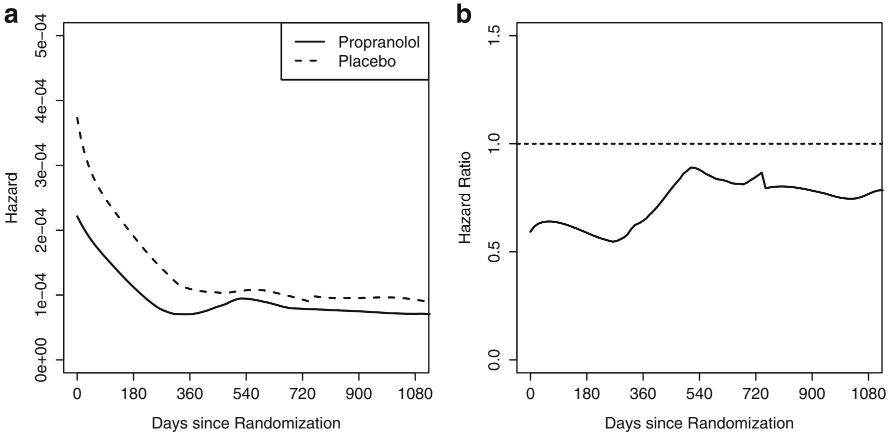


**Fig. 6.2** Estimated (**a**) hazard rates and (**b**) hazard ratio over time

## 6.4 Developments in Group Sequential Methods Motivated by BHAT

### 6.4.1 *The Lan–DeMets Error-Spending Approach*

The unequal group sizes that are proportional to the numbers of deaths during the periods between successive interim analyses in BHAT inspired Lan and DeMets (1983) to introduce the error-spending approach described in Sect. 4.1.3 for group sequential trials. To extend the error-spending approach to more general information-based trials, they let $v$ represent the proportion of information accumulated at time $t$ of interim analysis, so that $\pi(v)$ can be interpreted as the amount of type I error spent up to time $t$, with $\pi(0) = 0$ and $\pi(1) = \alpha$. Analogous to (4.5), they propose to choose $\alpha_j = \pi(v_j) - \pi(v_{j-1})$ to determine the stopping boundary $b_j$ $(j = 1,\dots,K)$ recursively by

$$P_{F=G}\left\{|W_1| \le b_1\sqrt{V_1},\dots,\ |W_{j-1}| \le b_{j-1}\sqrt{V_{j-1}},\ |W_j| > b_j\sqrt{V_j}\right\} = \alpha_j, \qquad (6.19)$$

where $W_i$ denotes the asymptotically normal test statistic at the $i$th interim analysis and $V_i$ denotes the corresponding variance estimate so that $v_i = V_i/V_k$, which will be explained in Sect. 6.5.1.

The error-spending approach has greatly broadened the scope of applicability of group sequential methods. For example, if one wants to use a constant boundary (1.3) for $|S_n|$ as in O'Brien and Fleming (1979), one can consider the corresponding continuous-time problem to obtain $\pi(v)$. The sample sizes at the times of interim analysis need not be specified in advance; what needs to be specified is the maximum sample size $n_K$. Lan and DeMets, who had been involved in the BHAT study, were motivated by BHAT to make group sequential designs more flexible. Although it does not require prespecified "information fractions" at times of interim analysis, the error-spending approach requires specification of the terminal information amount, at least up to a proportionality constant. While this is usually not a problem for immediate responses, for which total information is proportional to the sample size, the error-spending approach is much harder to implement for time-to-event responses, for which the terminal information is not proportional to $n_K$ and cannot be known until one carries the trial to its scheduled end.

### 6.4.2 *Two Time Scales and Haybittle-Type Boundaries*

Lan and DeMets (1983) have noted that there are two time scales in interim analysis of clinical trials with time-to-event endpoints. One is calendar time $t$ and the other is the "information time" $V_n(t)$, which is typically unknown before time $t$ unless restrictive assumptions are made a priori. To apply the error-spending approach to time-to-event responses, one needs an a priori estimate of the null variance of $S_n(t^*)$, where $t^*$ is the pre-scheduled end date of the trial and $S_n(t)$ is the logrank statistic or the more general censored rank statistic (6.16) evaluated at calendar time $t$. Let $v_1$ be such an estimate. Although the null variance of $S_n(t)$ is expected to be nondecreasing in $t$ under the asymptotic independent increments property, its estimate $V_n(t)$ may not be monotone, and we can redefine $V_n(t_j)$ to be $V_n(t_{j-1})$ if $V_n(t_j) < V_n(t_{j-1})$. Let $\pi : [0, v_1] \to [0,1]$ be a nondecreasing function with $\pi(0) = 0$ and $\pi(v_1) = \alpha$, which can be taken as the error-spending function of a stopping rule $\tau$, taking values in $[0, v_1]$, of a Wiener process. The repeated significance test whose boundary is generated by $\pi(\cdot)$ stops at time $t_j$ for $1 \le j < K$ if $V_n(t_j) \ge v_1$ (in which case it rejects $H_0 : F = G$ if $|S_n(t_j)| \ge b_j V_n^{1/2}(t_j)$) or if $V_n(t_j) < v_1$ and $|S_n(t_j)| \ge b_j V_n^{1/2}(t_j)$ (in which case it rejects $H_0$); it also rejects $H_0$ if $|S_n(t^*)| \ge b_K V^{1/2}(t^*)$, and stopping has not occurred prior to $t^* = t_K$. Letting $\alpha_j = \pi(v_1 \wedge V_n(t_j)) - \pi(V_n(t_{j-1}))$ for $j < K$ and $\alpha_K = \alpha - \pi(V_n(t_{K-1}))$, the boundary values $b_1, \ldots, b_K$ are defined recursively by (6.19), in which $\alpha_j = 0$ corresponds to $b_j = \infty$.

This test has type I error probability approximately equal to $\alpha$, irrespective of the choice of $\pi$ and the a priori estimate $v_1$. Its power, however, depends on $\pi$ and $v_1$. At the design stage, one can compute the power under various scenarios to come up with appropriate choice of $\pi$ and $v_1$. The requirement that the trial be stopped once $V_n(t)$ exceeds $v_1$ is a major weakness of the preceding stopping rule. Since one usually does not have sufficient prior information about the underlying survival distributions, the actual accrual rate and the withdrawal pattern, $v_1$ may substantially over- or underestimate the expected value of $V_n(t^*)$. Scharfstein et al. (1997) and Scharfstein and Tsiatis (1998) have proposed re-estimation procedures during interim analyses to address this difficulty, but re-estimation raises concerns about possible inflation of the type I error probability.

Another approach was proposed by Slud and Wei (1982). It requires the user to specify positive numbers $\alpha_1, \ldots, \alpha_K$ such that $\sum_{j=1}^K \alpha_j = \alpha$ so that the boundary $b_j$ for $|S_n(t_j)|/\sqrt{V_n(t_j)}$ is given by (6.19). However, there are no guidelines nor systematic ways to choose the $\alpha_j$. Gu and Lai (1998) proposed to use a Haybittle-type boundary that first chooses $b$ and then determines $c$ by

$$\begin{aligned} P\Big\{ &|W(V_n(t_j))| \ge b V_n^{1/2}(t_j) \text{ for some } j < K \\ &\text{or } |W(V_n(t_K))| \ge c V_n^{1/2}(t_K) \,\Big|\, V_n(t_1), \ldots, V_n(t_K) \Big\} = \alpha, \end{aligned} \tag{6.20}$$

where $\{W(v),\ v \ge 0\}$ is a standard Brownian motion. Lai and Shih (2004) subsequently refined this approach to develop the modified Haybittle–Peto tests that have been discussed in Sect. 4.2.

### 6.4.3 Conditional and Predictive Power

The motivation underlying conditional and predictive power is to forecast the outcome of a given test, called a *reference test*, of a statistical hypothesis $H_0$ from the data $D_t$ up to the time $t$ when such prediction is made. Since the outcome is binary (i.e., whether to reject $H_0$ or not), the forecast can be presented as the probability of rejecting $H_0$ at the end of the study given $D_t$. However, this probability has to be evaluated under some probability measure. In the context of hypothesis testing in a parametric family $\{P_\theta,\ \theta \in \Theta\}$, Lan et al. (1982) proposed to consider the conditional power

$$p_t(\theta) = P_\theta(\text{Reject } H_0 \,|\, D_t). \tag{6.21}$$

Subsequently, Choi et al. (1985) and Spiegelhalter et al. (1986) found it more appealing to put a prior distribution on $\theta$ and consider the posterior probability of rejecting $H_0$ at the end of the trial given $D_t$, and therefore advocated to consider the predictive power

$$P_t = P(\text{Reject } H_0 \,|\, D_t) = \int p_t(\theta)\, d\pi(\theta|D_t), \tag{6.22}$$

where $\pi(\theta|D_t)$ is the posterior distribution of $\theta$. This idea had been proposed earlier by Herson (1979).

While the conditional power approach to futility stopping requires specification of an alternative $\theta_1$, the predictive power approach requires specification of a prior distribution $\pi$. It is often difficult to come up with such specification in practice. On the other hand, one can use $D_t$ to estimate the actual $\theta$ by maximum likelihood or other methods, as suggested by Lan and Wittes (1988). For normal observations $X_i$ with common unknown mean $\theta$ and known variance $\sigma^2$, using Lebesgue measure on the real line as the improper prior for $\theta$ yields the sample mean $\bar{X}_t$, as the posterior mean and also the MLE. In this case, for the fixed sample size test that rejects $H_0 : \theta = 0$ if $\sqrt{n}\bar{X}_n \geq \sigma z_{1-\alpha}$, the predictive power is

$$\Phi\left(\sqrt{\frac{t}{n-t}}\left(\frac{\sqrt{n}}{\sigma}\bar{X}_t - z_{1-\alpha}\right)\right), \tag{6.23}$$

and the conditional power is

$$p_t(\bar{X}_t) = \Phi\left(\sqrt{\frac{n}{n-t}}\left(\frac{\sqrt{n}}{\sigma}\bar{X}_t - z_{1-\alpha}\right)\right), \tag{6.24}$$

in which $\Phi$ denotes the standard normal distribution function and $z_p = \Phi^{-1}(p)$.

Although using the conditional or predictive power to guide early stopping for futility is intuitively appealing, there is no statistical theory for such choice of the stopping criterion. In fact, using the MLE as the alternative already presupposes that the MLE falls outside the null hypothesis, and a widely used default option is to stop when the MLE belongs to $H_0$, which is consistent with (6.24) that falls below the type I error $\alpha$ in this case. However, this ignores the uncertainty in the estimate and can lose substantial power due to premature stopping, as shown in the simulation studies of Bartroff and Lai (2008a,b) on adaptive designs that use this kind of futility stopping; see also Sect. 8.3. Pepe and Anderson (1992) have proposed to adjust for this uncertainty by using $\bar{X}_t + \sigma/\sqrt{t}$ instead of $\bar{X}_t$ to substitute for $\theta_1$ in the conditional power approach.

Instead of estimating the alternative during interim analysis, one can focus on a particular alternative $\theta_1$ and consider the conditional power $p_t(\theta_1)$ or the predictive power with a prior distribution concentrated around $\theta_1$. Although Lan et al. (1982) have shown that adding futility stopping to the reference test of $H_0 : \theta \leq \theta_0$ if $p_t(\theta_1) \leq \gamma$ does not decrease the power of the reference test at $\theta_1$ by more than a factor of $\gamma/(1-\gamma)$, there is no statistical theory justifying why one should use a conditional instead of an unconditional test of $\theta \geq \theta_1$. Furthermore, as noted earlier, this approach leaves open the problem of how $\theta_1$ should be chosen for stopping a study due to futility.

## 6.5 Randomized Clinical Trials with Failure-Time Endpoints and Interim Analyses

### *6.5.1 Time-Sequential Censored Rank Statistics and Their Asymptotic Distributions*

Suppose a clinical trial involves $n = n' + n''$ patients with $n'$ of them assigned to treatment $X$ and $n''$ assigned to treatment $Y$. Let $T_i' \geq 0$ denote the entry time and $X_i > 0$ the survival time (or time to failure) after entry of the $i$th subject in treatment group $X$, and let $T_j''$ and $Y_j$ denote the entry time and survival time after entry of the $j$th subject in treatment group $Y$. The subjects are followed until they fail or withdraw from the study or until the study is terminated. Let $\xi_i'$ ($\xi_j''$) denote the time to withdrawal, possibly infinite, of the $i$th ($j$th) subject in the treatment group $X$ ($Y$). Thus, the data at calendar time $t$ consist of $(X_i(t), \delta_i'(t)),\ i = 1, \ldots, n'$, and $(Y_j(t), \delta_j''(t)),\ j = 1, \ldots, n''$, where

$$\begin{aligned} X_i(t) &= \min\left(X_i, \xi_i', (t - T_i')^+\right), \qquad \delta_i'(t) = I\left(X_i(t) = X_i\right), \\ Y_j(t) &= \min\left(Y_j, \xi_j'', (t - T_j'')^+\right), \quad \delta_j''(t) = I\left(Y_j(t) = Y_j\right), \end{aligned} \tag{6.25}$$

where $a^+$ is the positive part of number $a$. At a given calendar time, on the basis of the observed data (6.25) from the two treatment groups, one can compute the rank statistic (6.16) which can be expressed in the present notation as

$$S_n(t) = \sum_{i=1}^{n'} \delta_i'(t)\psi(H_{n,t}(X_i(t)))\left\{1 - \frac{m_{n,t}'(X_i(t))}{m_{n,t}'(X_i(t)) + m_{n,t}''(X_i(t))}\right\}$$
$$- \sum_{j=1}^{n''} \delta_j''(t)\psi(H_{n,t}(Y_j(t)))\frac{m_{n,t}'(Y_j(t))}{m_{n,t}'(Y_j(t)) + m_{n,t}''(Y_j(t))}, \qquad (6.26)$$

where $\psi$ is a nonrandom function on $[0,1]$ and

$$m_{n,t}'(s) = \sum_{i=1}^{n'} I(X_i(t) \geq s), \qquad m_{n,t}''(s) = \sum_{j=1}^{n''} I(Y_j(t) \geq s), \qquad (6.27)$$

$$N_{n,t}'(s) = \sum_{i=1}^{n'} I\left(X_i \leq \xi_i' \wedge (t - T_i')^+ \wedge s\right),$$
$$N_{n,t}''(s) = \sum_{j=1}^{n''} I\left(Y_j \leq \xi_j'' \wedge (t - T_j'')^+ \wedge s\right), \qquad (6.28)$$

$$1 - H_{n,t}(s) = \prod_{u<s}\left\{1 - \frac{\Delta N_{n,t}'(u) + \Delta N_{n,t}''(u)}{m_{n,t}'(u) + m_{n,t}''(u)}\right\}, \qquad (6.29)$$

where $\wedge$ denotes minimum. Note that unlike the Kaplan–Meier estimator (6.5), we take $\prod_{u<s}$ in (6.29) instead of $\prod_{u\leq s}$. This ensures that $H_{n,t}(s)$ is left continuous in $s$, obviating the need of taking $H_n(s-)$ in (6.16).

Suppose that $\psi$ is continuous and has bounded variation on $[0,1]$ and that the limits

$$b'(t,s) = \lim_{m\to\infty} m^{-1}\sum_{i=1}^{m} P\{\xi_i' \geq s, t - T_i' \geq s\},$$
$$b''(t,s) = \lim_{m\to\infty} m^{-1}\sum_{j=1}^{m} P\{\xi_j'' \geq s, t - T_j'' \geq s\}, \qquad (6.30)$$

exist and are continuous in $0 \leq s \leq t$. Suppose that the distribution functions $G$ and $G$ in Sect. 6.1.3 are are continuous, and let $\Lambda_F = -\log(1-F)$ and $\Lambda_G = -\log(1-G)$ denote their cumulative hazard functions. Let

$$\mu_n(t) = \int_0^t \psi(H_{n,t}(s))\frac{m_{n,t}'(s)m_{n,t}''(s)}{m_{n,t}'(s) + m_{n,t}''(s)}\,(d\Lambda_F(s) - d\Lambda_G(s)).$$

Note that $\mu_n(t) = 0$ if $F = G$. Gu and Lai (1991) have proved the following results on weak convergence of the time-sequential censored rank statistics $S_n(t)$ in $D[0,t^*]$. In Sect. 6.6 we provide some background material on weak convergence in $D[0,t^*]$ and give an outline of the proof of the results.

**Theorem 6.2.** *Assume that for some* $0 < \gamma < 1$,

$$n'/n \to \gamma \quad \text{as } n\,(= n' + n'') \to \infty \ \text{ with } \ 0 < \gamma < 1. \tag{6.31}$$

*(a) For fixed* $F$ *and* $G$, $\{n^{-1/2}(S_n(t) - \mu_n(t)),\ 0 \le t \le t^*\}$ *converges weakly in* $D[0,t^*]$ *to a zero-mean Gaussian process, and* $n^{-1}\mu_n(t)$ *converges in probability as* $n \to \infty$.

*(b) Let* $\{Z(t),\ 0 \le t \le t^*\}$ *denote the zero-mean Gaussian process in (a) when* $F = G$. *This Gaussian process has independent increments and*

$$\operatorname{Var}(Z(t)) = \gamma(1-\gamma) \int_0^t \frac{\psi^2(F(s))b'(t,s)b''(t,s)}{\gamma b'(t,s) + (1-\gamma)b''(t,s)}\, dF(s). \tag{6.32}$$

*(c) For fixed* $F$ *(and therefore* $\Lambda_F$ *also), suppose that as* $n \to \infty$, $G \to F$ *such that* $\int_0^{t^*} |d\Lambda_G/d\Lambda_F - 1| d\Lambda_F = O(n^{-1/2})$ *and* $\sqrt{n}(d\Lambda_G/d\Lambda_F(s) - 1) \to g(s)$ *as* $n \to \infty$, *uniformly in* $s \in I$ *and* $\sup_{s\in I}|g(s)| < \infty$ *for all closed subintervals* $I$ *of* $\{s \in [0,t^*] : F(s) < 1\}$. *Then* $\{n^{-1/2}S_n(t),\ 0 \le t \le t^*\}$ *converges weakly in* $D[0,t^*]$ *to* $\{Z(t) + \mu(t),\ 0 \le t \le t^*\}$, *where* $Z(t)$ *is the same Gaussian process as that in (b) and*

$$\mu(t) = -\gamma(1-\gamma) \int_0^t \frac{\psi(F(s))g(s)b'(t,s)b''(t,s)}{\gamma b'(t,s) + (1-\gamma)b''(t,s)}\, dF(s). \tag{6.33}$$

It follows from Theorem 6.2(b), (c) that the limiting Gaussian process of $\{n^{-1/2}S_n(t),\ t \ge 0\}$ has independent increments under $H_0 : F = G$ and under contiguous alternatives; contiguous alternatives refer to those in (c) that are within $O(n^{-1/2})$ from the null hypothesis $F = G$. Two commonly used estimates $V_n(t)$ of the variance of $S_n(t)$ under $H_0$ are

$$V_n(t) = \int_0^t \frac{\psi^2(H_{n,t}(s))m'_{n,t}(s)m''_{n,t}(s)}{(m'_{n,t}(s) + m''_{n,t}(s))^2}\, d\left(N'_{n,t}(s) + N''_{n,t}(s)\right) \tag{6.34}$$

and

$$V_n(t) = \int_0^t \frac{\psi^2(H_{n,t}(s))}{(m'_{n,t}(s) + m''_{n,t}(s))^2} \left\{ \left(m''_{n,t}(s)\right)^2\, dN'_{n,t}(s) + \left(m'_{n,t}(s)\right)^2\, dN''_{n,t}(s) \right\}. \tag{6.35}$$

As a compromise between these two choices, Gu and Lai (1991, p. 1421) also considered

$$V_n(t) = \{(6.34) + (6.35)\}/2. \tag{6.36}$$

For all three estimates, $n^{-1}V_n(t)$ converges in probability to (6.32) under $H_0$ and under contiguous alternatives. Hence, letting $v = n^{-1}V_n(t)$ and $W(v) = n^{-1/2}S_n(t)$, we can regard $W(v),\ v \geq 0$, as the standard Wiener process under $H_0$. Moreover, if $\psi$ is a scalar multiple of the asymptotically optimal score function, then we can also regard $W(v),\ v \geq 0$, as a Wiener process with some drift coefficient under contiguous alternatives.

When subjects are randomized to $X$ or $Y$ with probability 1/2, $\gamma = 1/2$ and $m'_{n,t} \sim m''_{n,t}$ under $F = G$. Therefore, for the logrank statistic for which $\psi \equiv 1$, (6.34) and (6.35) are asymptotically equivalent to

$$V_n(t) = (\text{total number of deaths up to time } t)/4, \tag{6.37}$$

which is the widely used formula for the null variance estimate of the logrank statistic in randomized clinical trials and was used, in particular, by BHAT.

### 6.5.2 Modified Haybittle–Peto Tests, Power, and Expected Savings

As noted in Sect. 6.4, the assumption of specified group sizes in the Pocock and O'Brien–Fleming boundaries led Lan and DeMets to develop an error-spending counterpart of these and other boundaries, but error spending is difficult to use in the time-sequential setting because the information (in terms of the null variance of the test statistic) at terminal date $t^*$ is not available at an interim analysis. In contrast, the modified Haybittle–Peto test in Sect. 4.2 can be easily applied to time-sequential trials, as shown in (6.20) which considers the two-sided test of $F = G$. For one-sided tests, we can clearly still control the type I error probability by replacing $|W(V_n(t_i))|$ in (6.20) by $W(V_n(t_i)),\ i = 1, \ldots, K$. This is similar to the methodology in Sect. 4.2.2 except that it does not include stopping for futility in choosing $b$ and $c$. He et al. (2012) have noted the difficulties in coming up with a good a priori estimate of $V_n(t_K)$ at the design stage and have developed the following method to handle futility stopping in time-sequential trials.

To begin with, note that the stopping rule and therefore also the test statistic have to be specified clearly in the protocol at the design stage when one does not know the accrual pattern, the withdrawal rate, and the actual survival distributions of the treatment and control groups. The power of the time-sequential test, however, depends on these unknown quantities, and staggered entry of the patients further complicates the power calculations. On the other hand, the time and cost constraints on the trial basically determine the maximum sample size and the maximum study duration at the design stage. In view of these considerations, the power calculations at the design stage for determining the sample size typically assume a working model in which the null hypothesis $F = G$ is embedded in a semiparametric family whose parameters are fully specified for the alternative hypothesis, under which the study duration and sample size of the two-sample semiparametric test are shown to have some prescribed power. The two-sample test statistic $S_n(t)$ is usually chosen to be an efficient score statistic or its asymptotic equivalent in the working model. As shown in Sect. 6.5.1, the asymptotic null variance $nV(t_i)$ of $S_n(t_i)$ depends not only on the survival distribution but also on the accrual rate and the censoring distribution up to the time $t_i$ of the $i$th interim analysis. The observed patterns, however, may differ substantially from those assumed in the working model for the power calculations at the design stage. In addition, the working model under which the test statistic is semiparametrically efficient (e.g., the proportional hazards model when a logrank test is used) may not actually hold. In this case, as the sample size $n$ approaches $\infty$, the limiting distribution of $n^{-1/2}S_n(t)$ is still normal with mean 0 and variance $V(t)$ under $F = G$ and has independent increments, but under local alternatives, the mean $\mu(t)$ of the limiting normal distribution of $n^{-1/2}S_n(t)$ may not be linear in $V(t)$, and may level off or even decrease with increasing $V(t)$, as will be shown at the end of this section.

For the futility stopping decision at interim analysis, He et al. (2012) propose to consider local alternatives, which suggest using the test $H_0 : \mu(t_i) \le 0$ for $1 \le i \le k$ versus $H_\delta : \mu(t_i) \ge \delta V(t_i)$ for some $i$, for the limiting Gaussian process. They choose the same $\delta$ as that used in the design stage to determine the sample size and trial duration, since one does not want to have substantial power loss at or near the alternative assumed at the design stage. Even when the working model does not actually hold, for which $\mu(t)/V(t)$ may vary with $t$, using it to determine the implied alternative for futility stopping only makes it more conservative to stop for futility because $\mu(t)$ tends to level off or even decrease instead of increasing linearly with $V(t)$. It remains to consider how to update, at the $i$th interim analysis, the estimate of the "maximum information" $nV(t^*)$ (and also $nV(t_j)$ for $j > i$ for future interim analyses) after observing accrual, censoring, and survival patterns that differ substantially from those assumed at the design stage. He et al. (2012) propose to replace $V(t)$ by the estimated $\hat{V}(t)$ for $t > t_i$ in the efficient score test of $H_\delta$ that involves these values.

Bayesian modeling provides a natural updating scheme for estimating, at time $t_i$ of interim analysis based on observations up to $t_i$, the null variance $V_n(t)$ of the score statistic $S_n(t)$ for $t > t_i$. Following Susarla and Van Ryzin (1976), He et al. (2012) use Dirichlet process priors for the distribution function $(F+G)/2$ and for the censoring distribution. Note that the null variance $V_n(t)$ is generated by the accrual rate, the censoring distribution, and the survival distributions $F$ and $G$ that are assumed to be equal. The parameter $\alpha$, which is a finite measure on $\mathbb{R}_+ = (0,\infty)$, of the Dirichlet process prior can be chosen to be some constant times the assumed parametric model that is used for power calculation at the design stage, where the constant is $\alpha(\mathbb{R}_+)$ that reflects the strength of this prior measure relative to the sample data. At the $i$th interim analysis, let $n_i$ be the total number of subjects who have been accrued, and let

$$Z_j^{(i)} = \min(Z_j, \xi_j, t_i - T_j), \quad \delta_j^{(i)} = I_{\{Z_j^{(i)} = Z_j\}},$$

$j = 1,\dots,n_i$, where $Z_j$ is the actual survival time of the $j$th patient, $T_j$ is the patient's entry time, and $\xi_j$ is the censoring time. We basically combine the $X$ and $Y$ groups in (6.25) into a combined group of survival times $Z_j$ and use the same idea. By rearranging the observations, we can assume without loss of generality that $Z_1^{(i)},\dots,Z_k^{(i)}$ are the uncensored observation, and let $Z_{[k+1]}^{(i)} < \cdots < Z_{[m]}^{(i)}$ denote the distinct ordered censored observations. Let

$$N_i(u) = \sum_{j=1}^{n_i} I_{\{Z_j^{(i)} \geq u\}}, \quad N_i^+(u) = \sum_{j=1}^{n_i} I_{\{Z_j^{(i)} > u\}},$$

$$\lambda_i(u) = \sum_{j=1}^{n_i} I_{\{Z_j^{(i)} = u, \delta_j = 0\}}, \quad Z_{[k]}^{(i)} = 0, \quad Z_{[m+1]}^{(i)} = \infty.$$

As shown by Susarla and Van Ryzin (1976), for $Z_{[l]}^{(i)} \leq u < Z_{[l+1]}^{(i)}$, the Bayes estimate of $H = 1-(F+G)/2$ at the $i$th interim analysis is given by

$$\hat{H}_i(u) = \frac{\alpha(u,\infty) + N_i^+(u)}{\alpha(\mathbb{R}_+) + n_i} \times \prod_{j=k+1}^{l} \left\{ \frac{\alpha[Z_{[j]}^{(i)},\infty) + N_i(Z_{[j]}^{(i)})}{\alpha[Z_{[j]}^{(i)},\infty) + N_i(Z_{[j]}^{(i)}) - \lambda_i(Z_{[j]}^{(i)})} \right\}. \tag{6.38}$$

Similarly, for updating the estimate $\hat{C}$ of the censoring distribution, He et al. (2012) interchange the roles of $T_j$ and $\xi_j$ above and replace $\alpha$ by $\alpha_c$ that is associated with the specification of the censoring distribution at the design stage. The accrual rates for the period prior to $t_i$ have been observed, and those for the future years can use what is assumed at the design stage. Since $V_n(t) = V_n(t_i) + [V_n(t) - V_n(t_i)]$, they estimate $V_n(t)$ by $V_n(t_i) + \mathrm{E}[V_n^*(t) - V_n^*(t_i)|\hat{H},\hat{C}]$, in which the expectation E assumes the updated accrual rates and can be computed by Monte Carlo simulations to generate the observations $(Z_j^*, \delta_j^*)$ that are independent of the $(Z_j^{(i)}, \delta_j^{(i)})$ observed up to time $t_i$.

We have noted in the second paragraph of this section that the limiting drift (6.33) may not be a monotone function of $t$ even for stochastically ordered alternatives. The following example is given by Gu and Lai (1991) for logrank statistics. Let $F_0$ be the exponential distribution with constant hazard rate $\lambda > 0$, and define for $\theta > 0$,

$$1 - F_\theta(x) = \begin{cases} \exp\{-(1-\theta)\lambda x\} & 0 \leq x \leq 1, \\ \exp\{3\theta\lambda/2 - (1+\theta/2)\lambda x\}, & 1 < x \leq 3, \\ \exp\{-\lambda x\}, & x > 3. \end{cases}$$

Let $F = F_0$ and $G = F_\theta$. Clearly $\{F_\theta, \theta \geq 0\}$ is stochastically ordered; in fact, $F_\theta \geq F_{\theta'}$ for $\theta \leq \theta'$. For $\theta > 0$, the hazard rate of $F_\theta$ is $\lambda_\theta(x) = (1-\theta)\lambda$ $(< \lambda)$ if $0 \leq$

$x \le 1$, $\lambda_\theta(x) = (1+\theta/2)\lambda$ $(> \lambda)$ if $1 < x \le 3$ and $\lambda_\theta(x) = \lambda$ for $x > 3$. Therefore, the function $g$ in Theorem 6.2(c), in which $\theta \to 0$ at rate $n^{-1/2}$, is given by

$$g(x) = \begin{cases} -1, & 0 \le x \le 1, \\ \frac{1}{2}, & 1 < x \le 3, \\ 0, & x > 3. \end{cases}$$

Hence, for the time-sequential logrank statistic, the limiting drift $\mu(t)$, given by (6.33) for contiguous alternatives, is increasing for $0 < t < 1$, decreasing for $1 < t < 3$, and constant for $t \ge 3$, under the assumption that $b'(t,u)b''(t,u) > 0$ for all $0 \le u < t$. This shows that a level-$\alpha$ test of $H_0 : F = G$ based on the logrank statistic $S_n(t_1)$ with $1 < t_1 < 3$ can have higher power at the stochastically ordered alternative $(F,G) = (F_0, F_\theta)$ with $\theta > 0$ than that based on $S_n(t_2)$ evaluated at a later time $t_2 > t_1$, providing therefore both savings in time and increase in power. Gu and Lai (1998, pp. 422–425) confirm this in a simulation study.

Another simulation study in Gu and Lai (1998, pp. 425–426) for time-sequential logrank tests compares the performance of the modified Haybittle–Peto boundaries (6.20) with several other stopping boundaries, including the O'Brien–Fleming boundary that was used in the BHAT report, under the same accrual and withdrawal patterns as those in the BHAT data, and assuming the same times of interim analysis. The fixed duration test that stops at 48 months is also included for comparison. It considers the null hypothesis $H_0 : F = G$, the alternative hypothesis $H_1$ that assumes the proportional hazards model with hazard ratio 0.699 of the treatment to placebo group that roughly corresponds to the planned number of patients, and the alternative hypothesis $H_2$ which has time-varying hazard ratios of 0.599, 0.708, 0.615, 1.560, 0.800, and 0.323 for each of the 6-month periods estimated from the BHAT data. The simulation study shows that all tests have type I error close to the prescribed level 0.05 and that the power of the time-sequential logrank test with the modified Haybittle–Peto or O'Brien–Fleming boundary is very close to that of the fixed-duration logrank test. However, the modified Haybittle–Peto boundary gives the greatest reduction in trial duration under $H_1$ or $H_0$.

## 6.6 Appendix: Martingale Theory and Applications to Sequential/Survival Analysis

### *6.6.1 Optional Stopping Theorem*

A sequence of random variables $S_n$ satisfying $E|S_n| < \infty$ for all $n$ is called a *martingale* if

$$E(S_n|\mathscr{F}_{n-1}) = S_{n-1} \; a.s. \text{ (i.e., almost surely, or with probability 1).}$$

Here $\mathscr{F}_t$ denotes the information set up to time $t$. (To define conditional expectations more generally in terms of Radon–Nikodym derivatives of measures, the $\mathscr{F}_t$ are assumed to be $\sigma$-fields such that $S_t$ is $\mathscr{F}_t$-measurable and $\mathscr{F}_t \subset \mathscr{F}$, where $\mathscr{F}$ is the $\sigma$-field containing all the events under consideration.) Martingale theory can be extended to *submartingales* for which $E(S_n|\mathscr{F}_{n-1}) \geq S_{n-1}$ *a.s.* and (by multiplying the $S_n$ by $-1$) also to *supermartingales* for which $E(S_n|\mathscr{F}_{n-1}) \leq S_{n-1}$ *a.s.* If $S_n$ is a martingale (or submartingale) and $E|\varphi(S_n)| < \infty$ for all $n$, then $\varphi(S_n)$ is a submartingale if $\varphi$ is convex (or convex and nondecreasing). A random variable $N$ taking values in $\{1,2,\ldots\}$ is called a *stopping time* if $\{N = n\} \in \mathscr{F}_n$ for all $n$. The stopped $\sigma$-field $\mathscr{F}_N$ is defined by

$$\mathscr{F}_N = \{A \in \mathscr{F} : A \cap \{N \leq n\} \in \mathscr{F}_n \text{ for all } n\}. \tag{6.39}$$

Martingale theory has provided an important tool to sequential analysis via the *optional stopping theorem*, which roughly says that a martingale (submartingale) up to a stopping time also remains a martingale (submartingale). By "roughly" we mean "under some regularity conditions." Since martingales are defined via conditional expectations, it is not surprising that these regularity conditions involve some sort of integrability. In particular, $\{S_n, n \geq 1\}$ is said to be *uniformly integrable* if $\sup_n E|S_n|I_{\{|S_n|>a\}} \to 0$ as $n \to \infty$. A more precise statement of the optional stopping theorem is the following.

**Theorem 6.3.** *If $N \leq M$ are stopping times and*

$$S_{M\wedge n} \textit{ is a uniformly integrable submartingale}, \tag{6.40}$$

*then $S_N \leq E(S_M|\mathscr{F}_N)$ a.s. and therefore $ES_N \leq ES_M$. In particular, if $S_n = \sum_{i=1}^n X_i$ is a submartingale and $\sup_n E(|X_n| \big| \mathscr{F}_{n-1}) \leq Y$ for some integrable random variable $Y$, then (6.40) holds for all stopping times $M$.*

Since a martingale is both a submartingale and a supermartingale, Theorem 6.3 applied to martingales yields $E(S_M|\mathscr{F}_N) = S_N$ for any stopping times $M$ and $N$. This result is a generalization of Wald's equation (3.7), in which the $X_i$ are i.i.d. so that $E(|X_n| \big| \mathscr{F}_{n-1}) = E|X_n|$ and $\sum_{i=1}^n (X_i - \mu)$ is a martingale, and therefore Theorem 6.3 with $N = 1$ yields (3.7).

Martingale theory can be readily extended to continuous-time stochastic processes if the sample paths are *a.s.* right continuous, which we shall assume in the rest of this section. We replace $n \in \{1,2,\ldots\}$ in the preceding discrete-time setting by $t \in [0,\infty)$. The increasing sequence of $\sigma$-fields $\mathscr{F}_n$ is now replaced by a *filtration* $\{\mathscr{F}_t, t \geq 0\}$, and a stopping time $T$ is a random variable taking values in $[0,\infty)$ and such that $\{T \leq t\} \in \mathscr{F}_t$ for all $t \geq 0$. The optional stopping theorem still holds for right-continuous submartingales. Moreover, with probability 1, a right continuous submartingale has left-hand limits at all $t \in (0,\infty)$; this follows from Doob's upcrossing inequality for right-continuous submartingales.

The stopped field $\mathscr{F}_T$ can again be defined by (6.39) with $n$ replaced by $t$. A filtration $\{\mathscr{F}_t\}$ is said to be *right-continuous* if $\mathscr{F}_t = \mathscr{F}_{t+} := \cap_{\varepsilon>0}\mathscr{F}_{t+\varepsilon}$. It is said to be *complete* if $\mathscr{F}_0$ contains all the $P$-null sets (that have zero probability) in $\mathscr{F}$. In what follows we shall assume that the process $\{S_t, t \geq 0\}$ is right continuous and adapted to a right-continuous and complete filtration $\{\mathscr{F}_t\}$ ("adapted" means that $S_t$ is $\mathscr{F}_t$-measurable.) Letting $\Omega$ denote the sample space, the $\sigma$-field generated on $\Omega \times [0,\infty)$ by the space of adapted processes which are left continuous on $(0,\infty)$ is called the *predictable $\sigma$-field.* A process $\{S_t\}$ is *predictable* if the map $(\omega,t) \mapsto S_t(\omega)$ from $\Omega \times [0,\infty)$ to $\mathbb{R}$ is measurable with respect to the predictable $\sigma$-field.

### 6.6.2 Predictable Variation and Stochastic Integrals

Let $\mathscr{F}_a$ be the class of stopping times such that $P(T \leq a) = 1$ for all $T \in \mathscr{F}_a$. A right-continuous process $\{S_t, t \geq 0\}$ adapted to a filtration $\{\mathscr{F}_t\}$ is said to be of class DL if $\{S_T, T \in \mathscr{F}_a\}$ is uniformly integrable for every $a > 0$. If $\{S_t, \mathscr{F}_t, t \geq 0\}$ is a nonnegative, right continuous submartingale, then it is of class DL. The *Doob–Meyer decomposition* says that if a right-continuous submartingale $\{S_t, \mathscr{F}_t, t \geq 0\}$ is of class DL, then it admits the decomposition

$$S_t = M_t + A_t, \tag{6.41}$$

in which $\{M_t, \mathscr{F}_t, t \geq 0\}$ is a right-continuous martingale with $M_0 = 0$ and $A_t$ is predictable, non-decreasing and right-continuous. Moreover, the decomposition is essentially unique in the sense that if $S_t = M'_t + A'_t$ is another decomposition, then $P\{M_t = M'_t, A_t = A'_t \text{ for all } t\} = 1$. The process $A_t$ in the Doob–Meyer decomposition is called the *compensator* of the submartingale $\{S_t, \mathscr{F}_t, t \geq 0\}$.

Suppose $\{M_t, \mathscr{F}_t, t \geq 0\}$ is a right-continuous martingale that is square integrable, that is, $EM_t^2 < \infty$ for all $t$. Since $M_t^2$ is a right-continuous, nonnegative submartingale (by Jensen's inequality), it has the Doob–Meyer decomposition whose compensator is called the *predictable variation* process and denoted by $\langle M \rangle_t$, that is, $M_t^2 - \langle M \rangle_t$ is a martingale. If $\{N_t, \mathscr{F}_t, t \geq 0\}$ is another right-continuous square-integrable martingale, then $(M_t + N_t)^2 - \langle M + N \rangle_t$ and $(M_t - N_t)^2 - \langle M - N \rangle_t$ are martingales, and the *predictable covariation* process $\langle M, N \rangle_t$ is defined by

$$\langle M,N \rangle_t = \frac{1}{4}\{\langle M+N \rangle_t - \langle M-N \rangle_t\}, \quad t \geq 0. \tag{6.42}$$

Let $\mathscr{M}_2$ denote the linear space of all right-continuous, square-integrable martingales $M$ with $M_0 = 0$. Two processes $X$ and $Y$ on $(\Omega, \mathscr{F}, P)$ are *indistinguishable* if $P(X_t = Y_t \text{ for all } t \geq 0) = 1$. Two martingales $M, N$ belonging to $\mathscr{M}_2$ are said to be *orthogonal* if $\langle M,N \rangle_t = 0$ *a.s.* for all $t \geq 0$ or, equivalently, if $\{M_t N_t, \mathscr{F}_t, t \geq 0\}$ is a martingale. Let $\mathscr{M}_2^c = \{M \in \mathscr{M}_2 : M \text{ has continuous sample paths}\}$ and $\mathscr{M}_2^d = \{N \in \mathscr{M}_2 : N \text{ is orthogonal to } M \text{ for all } M \in \mathscr{M}_2^c\}$. It can be shown that every $M \in \mathscr{M}_2$ has an essentially unique decomposition

$$M = M^c + M^d, \quad \text{with } M^c \in \mathscr{M}_2^c \text{ and } M^d \in \mathscr{M}_2^d. \tag{6.43}$$

While $M^c$ is called the continuous part of $M$, $M^d$ is called its "purely discontinuous" part. Note that $M^c$ and $M^d$ are orthogonal martingales. We can relax the integrability assumptions above by using *localization*. If there exists a sequence of stopping times $T_n$ such that $\{M_{T_n \wedge t}, \mathscr{F}_t, t \geq 0\}$ is a martingale (or a square-integrable martingale, or bounded), then $\{M_t, \mathscr{F}_t, t \geq 0\}$ is called a *local martingale* (or *locally square-integrable martingale*, or *locally bounded*). By a limiting argument, we can define $\langle M \rangle_t$, $\langle M, N \rangle_t$, $M^c$, and $M^d$ for locally square integrable martingales.

We next define the stochastic integral $\int_0^t X_s dY_s$ with *integrand* $X = \{X_s,\ 0 \leq s \leq t\}$ and *integrator* $Y = \{Y_s,\ 0 \leq s \leq t\}$. If $Y$ has bounded variation on $[0,t]$, then the integrand can be taken as an ordinary pathwise Lebesgue–Stieltjes integral over $[0,t]$. If $Y$ is a right-continuous, square-integrable martingale and $X$ is a predictable, locally bounded process such that $\int_0^t X_s^2 d\langle Y \rangle_s < \infty$ *a.s.*, then $\int_0^t X_s dY_s$ can be defined by the limit (in probability) of integrals (which reduce to sums) whose integrands are step functions and converge to $X$ in an $L_2$-sense. In this case, $\int X dY = \{\int_0^s X_u dY_u,\ 0 \leq s \leq t\}$ is a locally square-integrable martingale and

$$\left\langle \int X dY \right\rangle_t = \int_0^t X_s^2 \, d\left\langle Y \right\rangle_s. \tag{6.44}$$

### 6.6.3 Rebolledo's CLT

Rebolledo's CLT for continuous-time martingales (Andersen et al., 1993, Sect. II.5.1) provides a basic tool to derive (6.4), (6.7), and Theorem 6.2. The *Skorohod space* $D[0,\infty)$ (or $D[0,t^*]$) is the metric space (with the Skorohod metric) of all right-continuous functions on $[0,\infty)$ (or $[0,t^*]$) with left-hand limits. A sequence of random variables $X_n$ taking values in a metric space $\mathscr{X}$ is said to converge *weakly* (or *in distribution*) to $Y$ in $\mathscr{X}$ if $\lim_{n\to\infty} Ef(X_n) = Ef(Y)$ for all bounded continuous functions $f: \mathscr{X} \to \mathbb{R}$. Let $\boldsymbol{M}_n(t)$ be a sequence of stochastic processes taking values in $\mathbb{R}^k$ such that each component is a locally square-integrable martingale with right-continuous sample paths. Let $M_{n,i}^{\varepsilon}(t) = M_{n,i}(t) I_{\{|\triangle M_{n,i}(t)| \geq \varepsilon\}}$ be the truncation of the purely discontinuous part of the $i$th component $M_{n,i}$ of $\boldsymbol{M}_n$ that ignores jump sizes less than $\varepsilon$. Rebolledo's CLT gives conditions under which $\boldsymbol{M}_n$ converges weakly to a continuous Gaussian martingale $\boldsymbol{M}$. The martingale property implies that $\boldsymbol{M}$ has uncorrelated increments, which are therefore independent since $\boldsymbol{M}$ is Gaussian. Hence $\boldsymbol{M}$ is a Gaussian process with independent increments: $\boldsymbol{M}(t) - \boldsymbol{M}(s) \sim N(\boldsymbol{0}, \boldsymbol{V}(t) - \boldsymbol{V}(s))$, where $\boldsymbol{V}(t)$ is the covariance matrix of $\boldsymbol{M}(t)$. Let $\mathscr{T} = [0,\infty)$ or $[0,t^*]$.

**Theorem 6.4.** *Let $\mathscr{T}_0 \subset \mathscr{T}$ and assume that as $n \to \infty$,*

$$\langle \boldsymbol{M}_n \rangle_t \xrightarrow{P} \boldsymbol{V}(t) \quad \textit{and} \quad \langle M_{n,i}^{\varepsilon} \rangle_t \xrightarrow{P} 0 \tag{6.45}$$

*for every $t \in \mathscr{T}_0$ and $\varepsilon > 0$. Then*

$$(\boldsymbol{M}_n(t_1),\ldots,\boldsymbol{M}_n(t_\ell)) \xrightarrow{\mathscr{D}} (\boldsymbol{M}(t_1),\ldots,\boldsymbol{M}(t_\ell)) \text{ as } n \to \infty, \tag{6.46}$$

*for all $t_1,\ldots,t_\ell \in \mathscr{T}_0$. If furthermore $\mathscr{T}_0$ is dense in $\mathscr{T}$ and contains $t^*$ in the case $\mathscr{T} = [0,t^*]$, then $\boldsymbol{M}_n$ converges weakly to $\boldsymbol{M}$ in the Skorohod space $(D(\mathscr{T}))^k$.*

If $\boldsymbol{M}_n(t)$ converges in distribution to the normal random vector $\boldsymbol{M}(t)$ that has covariance matrix $\boldsymbol{V}(t)$, then it is natural to expect $\langle \boldsymbol{M}_n \rangle_t$ to converge in probability to $\boldsymbol{V}(t)$, which is the first assumption in (6.45). Although we have only defined the predictable variation process $\langle M \rangle_t$ for a univariate locally square-integrable martingale, we can easily extend to martingale vectors $\boldsymbol{M}$ since we have also defined the predictable covariation process (6.42). If $M_{n,i}$ converges in distribution to a continuous process, then it is natural to expect the jumps of $M_{n,i}$ to be negligible, and this explains the second assumption in (6.45). Thus, the convergence of finite-dimensional distributions (6.46) requires minimal assumptions in (6.45). Weak convergence in $D(\mathscr{T})$ (or the product space $(D(\mathscr{T}))^k$) entails more than convergence in finite-dimensional distributions, but the martingale structure basically satisfies that extra condition called "tightness," yielding Rebolledo's CLT restated in Theorem 6.3.

### 6.6.4 Counting Processes and Applications to Survival Analysis

The stochastic process $N(\cdot)$ defined in (6.1) is called a *counting process*. It is nonnegative, right continuous, and nondecreasing and is therefore a submartingale. The Doob–Meyer decomposition in this case yields $\int_0^t Y(s)d\Lambda(s)$ as the compensator of $N(t)$, where $Y(t) = \sum_{i=1}^n I_{\{T_i \geq t\}}$ and $\Lambda$ is the cumulative hazard functions; see Sect. 6.1.1. Hence,

$$M(t) := N(t) - \int_0^t Y(s)\,d\Lambda(s) \quad \text{is a martingale.} \tag{6.47}$$

Moreover,

$$\langle M \rangle_t = \int_0^t Y(s)(1 - \triangle\Lambda(s))\,d\Lambda(s); \tag{6.48}$$

see Andersen et al. (1993, p. 74). Therefore, if $U = \{U(s),\ 0 \leq s \leq t\}$ is locally bounded and predictable, $\int U\,dM$ is a locally square-integrable martingale and

$$\left\langle \int UdM \right\rangle_t = \int_0^t U^2(s)Y(s)(1 - \triangle\Lambda(s))\,d\Lambda(s). \tag{6.49}$$

If $F$ is continuous, then $\triangle\Lambda = 0$, and (6.4) follows from Theorem 6.4 and the strong law of large numbers. When $F$ has jumps, the standard error estimator of the Nelson–Aalen estimate in the denominator of (6.4) should be replaced by

$$\left\{ \int_0^t \frac{I_{\{Y(s)>0\}}(Y(s) - \triangle N(s))}{Y^3(s)} dN(s) \right\}^{1/2}$$

in view of (6.49); see Andersen et al. (1993, p. 181). This proves the asymptotic normality of the censored rank statistics (6.16) under the null hypothesis $F = G$ and of $\dot{\ell}(\boldsymbol{\beta}_0)$ in the Cox regression model, from which (6.14) follows by the Taylor expansion $0 = \dot{\ell}(\hat{\boldsymbol{\beta}}) \approx \dot{\ell}(\boldsymbol{\beta}_0) + \ddot{\ell}(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$.

To prove (6.7), let $J(t) = I_{\{Y(t)>0\}}$, rewrite (6.5) as $\widehat{S}(t)$ and modify (6.6) as $\widetilde{S}(t)$, where

$$\widehat{S}(t) = \prod_{s\le t}\left(1 - \frac{J(s)}{Y(s)}\triangle N(s)\right), \quad \widetilde{S}(t) = \prod_{s\le t}\left(1 - \widetilde{\Lambda}(s)\right) \text{ with } \widetilde{\Lambda}(t) = \int_0^t J(s)d\Lambda(s).$$

Then the quotient $\widehat{S}(t)/\widetilde{S}(t)$ satisfies Duhamel's equation

$$\frac{\widehat{S}(t)}{\widetilde{S}(t)} - 1 = -\int_0^t \frac{\widehat{S}(s-)}{\widetilde{S}(s)} \frac{J(s)}{Y(s)} \left(dN(s) - Y(s)\,d\Lambda(s)\right); \tag{6.50}$$

see Andersen et al. (1993, pp. 91, 257). From (6.47) and (6.50), it follows that $Z(t) := \widehat{S}(t)/\widetilde{S}(t) - 1$ is a martingale. Since $\widehat{S}/\widetilde{S} \to 1$ on $[0,t]$ with probability 1, (6.7) follows from (6.48) and Theorem 6.4.

### 6.6.5 *Application to Time-Sequential Censored Rank Statistics*

Gu and Lai (1991) proved Theorem 6.2 by making use of Rebolledo's CLT to establish convergence of finite-dimensional distributions of $Z_n(t) := n^{-1/2}(S_n(t) - \mu_n(t))$ in part (a) of the theorem and of $Z_n(t) := n^{-1/2}S_n(t)$ in part (c) of the theorem. To prove tightness, they make use of empirical process theory and exponential inequalities for martingales. Actually one cannot apply martingale theory directly to $Z_n(t)$ since $Z_n(t)$ is not a martingale. The martingales in Sect. 6.6.4 are indexed by the information time $s$ and not by the calendar time $t$; see (6.26)–(6.29). Therefore, instead of $S_n(t)$ and $Z_n(t)$, Gu and Lai (1991) consider more generally $S_n(t,s)$ that replaces $\sum_{i=1}^{n'}$ and $\sum_{j=1}^{n''}$ in (6.26) by $\sum_{1\le i\le n':X_i(t)\le s}$ and $\sum_{1\le j\le n'':Y_j(t)\le s}$, respectively. For fixed calendar time $t$, martingale theory can be applied to the process $S_n(t,s)$. More generally, given calendar times $t_1 < \cdots < t_k \le t^*$, $(S_n(t_1,s),\ldots,S_n(t_k,s))$ is a $k$-dimensional stochastic process to which Rebolledo's CLT for multivariate continuous-time martingales can be applied, similar to what we have done in the preceding section for the case $k = 1$. Hence, Gu and Lai (1991) consider weak convergence of the stochastic processes (random fields) $\widetilde{Z}_n(t,s)$ with two-dimensional time parameters $(t,s)$. Since $Z_n(t) = \widetilde{Z}_n(t,t)$, Theorem 6.2 follows from the weak convergence results of the random field $\widetilde{Z}_n(t,s)$.

## 6.7 Supplements and Problems

1. Suppose $F$ is continuous. Show that the cumulative hazard function $\Lambda$ is simply $-\log S$, where $S = 1-F$ is the survival function. Hence, for given $\Lambda$, $S$ satisfies the Volterra integral equation $S(t) = 1 - \int_0^t S(u)\,d\Lambda(u)$. More generally, given a function $\Lambda$ of bounded variation on $[0,T]$ that is right continuous and has left-hand limits, Volterra's equation

$$S(t) = 1 - \int_0^t S(u-)\,d\Lambda(u) \tag{6.51}$$

has a unique right-continuous solution that has left-hand limits and is given by the product integral

$$S(t) = \prod_{s\le t}\Big(1 - d\Lambda(s)\Big). \tag{6.52}$$

In particular, for the cumulative hazard function $\Lambda(t) = \int_0^t \frac{dF(u)}{S(u-)}$, $S$ clearly satisfies (6.51) and therefore has the representation (6.52).

2. *Extension of classical rank statistics to censored data.*
In Sect. 6.1.3 we have discussed nonparametric group sequential tests using rank statistics of the type $\ell_n = \sum_{i=1}^{n'} \varphi(R_i/n)$, where $R_i$ is the rank of $X_i$ in the combined sample. Since $R_i/n$ can be expressed in terms of $(\hat{F}_{n'}(X_i), \hat{G}_{n''}(X_i))$, where $\hat{F}_{n'}$ and $\hat{G}_{n''}$ are the empirical distribution functions, Sect. 4.3.3 considers somewhat more general functionals $\int J_n(\hat{F}_{n'}, \hat{G}_{n''})\,d\hat{F}_{n'}$ of the empirical distribution functions. Gu et al. (1991) explain why (6.16), with $\psi$ defined by (6.17), provides a natural extension of $\sum_{i=1}^{n'} \varphi(R_i/n)$. Denote $\psi(H_n(Z_{(k)}))$ by $p_n(Z_{(k)})$ and assume $F$ and $G$ to be continuous so that there are no ties among the uncensored observations. Let $N(z)$ denote the number of observations in the combined sample that are $\ge z$. For any pair $(x,y)$ of $X,Y$ values (possibly censored) in the combined sample, define the weights

$$w(x,y) = \begin{cases} -p_n(y)/N(y) & \text{if } y \text{ is uncensored and } y \le x,\\ p_n(x)/N(x) & \text{if } x \text{ is uncensored and } x \le y,\\ 0 & \text{in all other cases.} \end{cases} \tag{6.53}$$

Then (6.16) can be written in the form

$$S_n = \sum_{x,y} w(x,y), \tag{6.54}$$

where $\sum_{x,y}$ denotes summation over all the $n'n''$ pairs of $X,Y$ values in the combined sample. To prove this, note that if $z_i = 1$, then $Z_{(i)} = X_r$ (uncensored) for some $r$ and the corresponding summand in (6.16) reduces to

$$\begin{aligned} p_n(Z_{(i)})(z_i - m_i/\#_i) &= [p_n(Z_{(i)})/\#_i](\#_i - m_i) \\ &= [p_n(X_r)/N(X_r)] \cdot (\text{number of } Y\text{'s} \geq X_r). \end{aligned}$$

Likewise, if $z_i = 0$, then $Z_{(i)} = Y_t$ (uncensored) for some $t$ and

$$p_n(Z_{(i)})(z_i - m_i/\#_i) = -[p_n(Y_t)/N(Y_t)] \cdot (\text{number of } X\text{'s} \geq Y_t).$$

For the special case $p_n(z) = N(z)$, we have

$$w(x,y) = \begin{cases} -1 & \text{if } y \text{ is uncensored and } y \leq x, \\ 1 & \text{if } x \text{ is uncensored and } x \leq y, \\ 0 & \text{in all other cases.} \end{cases} \tag{6.55}$$

In this case, the right-hand side of (6.54) is Gehan's (1965) extension, to censored data, of the Mann–Whitney statistic $\sum_{x,y} w(x,y)$ for complete data (with $w = 1$ or $-1$ according to $x < y$ or $x > y$). Gu et al. (1991) call the function $p_n$ in (6.53) a "payment function," in view of the following two-team game interpretation of (6.16). Consider a contest between two teams: $X$, with $n'$ players, and $Y$, with $n''$ players. All $n = n' + n''$ players simultaneously play, say, a videogame. Once a player makes an error, he is disqualified from further play and pays an amount $p_n(z)$, depending on the length of time $z$ he is in the game, to be distributed equally among all players in the game at that time (including himself). In addition, any player can withdraw from further play before he makes an error (i.e., be "censored"). Thus, the total amount that team $X$ pays team $Y$ is equal to $S_n$ defined by (6.16). Note that $-S_n$ is the amount that team $Y$ pays team $X$ and that a negative value of $S_n$ signifies that $X$ is the better team. The payment function $p_n(z) = N(z)$ used by Gehan is inexorably linked to the censoring pattern. Since censoring is unrelated to the skill of the players, it seems more reasonable to choose a payment function that is relatively unaffected by censoring. One such choice is $p_n(z) = \varphi(\hat{H}_n(z)) - \Phi(\hat{H}_n(z))$, where $\Phi(u) = \int_u^1 \varphi(t)\,dt/(1-u)$ that is used in (6.17). Note that in this two-team contest, $w(x,y)$ defined in (6.53) represents the amount (not necessarily nonnegative) an $X$-player who leaves the game (i.e., either is disqualified or withdraws) at time $x$ pays a $Y$-player who leaves the game at time $y$. This interpretation provides an alternative explanation of (6.54).

3. *Phase II–III cancer trial designs and time-varying hazard ratios of treatment to control.*
Although randomized Phase II studies are commonly conducted in other therapeutic areas, in oncology the majority of Phase II studies leading to Phase III studies are single-arm studies with a binary tumor response endpoint and the most commonly used Phase II designs are Simon's (1989) single-arm two-stage designs for testing $H_0 : p \leq p_0$ versus $H_1 : p \geq p_1$ where $p$ is tumor response rate, as described in Supplement 8 of Sect. 4.5. Whether the new treatment is declared promising in a single-arm Phase II trial, however, depends strongly on the prespecified $p_0$ and $p_1$. As noted by Vickers et al. (2007), uncertainty in the choice of $p_0$ and $p_1$ can increase the likelihood that (a) a treatment with no viable positive treatment effect proceeds to Phase III, for example, if $p_0$ is chosen artificially small to inflate the appearance of a positive treatment effect when one exists, or (b) a treatment with positive treatment effect is prematurely abandoned at Phase II, for example, if $p_1$ is chosen optimistically large. To circumvent the problem of choosing $p_0$, Vickers et al. (2007) and Rubinstein et al. (2009) have advocated randomized Phase II designs. In particular, it is argued that randomized Phase II trials are needed before proceeding to Phase III trials when (a) there is not a good historical control rate, due to either incomparable controls (causing bias), few control patients (resulting in large variance of the control rate estimate), or outcome that is not "antitumor activity", or when (b) the goal of Phase II is to select one from several candidate treatments or several doses for use in Phase III. However, few Phase II cancer studies are randomized with internal controls. The major barriers to randomization include that randomized designs typically require a much larger sample size than single-arm designs and that there are multiple research protocols competing for a limited patient population. Being able to include the Phase II study as an internal pilot for the confirmatory Phase III trial may be the only feasible way for a randomized Phase II cancer trial of such sample size and scope to be conducted.

Although tumor response is an unequivocally important treatment outcome, the clinically definitive endpoint in Phase III cancer trials is usually time to event, such as time to death or time to progression. The go/no-go decision to Phase III is typically based on tumor response because the clinical time-to-failure endpoints in Phase III are often of long latency, such as time to bone metastasis in prostate cancer studies. These failure-time data, which are collected as censored data and analyzed as a secondary endpoint in Phase II trials, can be used for planning the subsequent Phase III trial. Furthermore, because of the long latency of the clinical failure-time endpoints, the patients treated in a randomized Phase II trial carry the most mature definitive outcomes if they are also followed in the Phase III trial. Seamless Phase II–III trials with bivariate endpoints consisting of tumor response and time to event are an attractive idea, but up to now only Bayesian statistical methodologies, introduced by Inoue et al. (2002) and Huang et al. (2009), have been developed for their design and analysis.

The aforementioned Bayesian approach is based on a parametric mixture model that relates survival to response. Let $z_i$ denote the treatment indicator (0

$=$ control, $1 =$ experimental), $\tau_i$ denote survival time, and $y_i$ denote the binary response for patient $i$. Assume that the responses $y_i$ are independent Bernoulli variables and the survival time $\tau_i$ given $y_i$ follows an exponential distribution, denoted $\mathrm{Exp}(\lambda)$ in which $1/\lambda$ is the mean:

$$y_i \mid z_i = z \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\pi_z), \tag{6.56}$$

$$\tau_i \mid \{y_i = y,\ z_i = z\} \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}(\lambda_{z,y}). \tag{6.57}$$

Then the conditional distribution of $\tau_i$ given $z_i$ is a mixture of exponentials:

$$\tau_i \mid z_i = z \overset{\text{i.i.d.}}{\sim} \pi_z \mathrm{Exp}(\lambda_{z,1}) + (1-\pi_z)\mathrm{Exp}(\lambda_{z,0}). \tag{6.58}$$

The parametric relationship of response $y$ on survival $\tau$ assumed by (6.56) and (6.57) enables one to use the Bayesian approach to update the parameters so that various posterior quantities can be used for Bayesian inference. Note that $y$ is a "causal intermediary" because treatment may affect $y$ and then $\tau$ through its effect on $y$ and may also have other effects on $\tau$. The model (6.56)–(6.57) reflects this nicely by considering the conditional distribution of $y$ given $z$ and that of $\tau$ given $(y, z)$.

Let $\mu_z = E(\tau_i \mid z_i = z)$ denote the mean survival time in treatment group $z$. Inoue et al. (2002) proposed the following Bayesian design, assuming independent prior gamma distributions for $\lambda_{z,0}$ and $\lambda_{z,1}$ $(z = 0, 1)$ and beta distributions for $\pi_0$ and $\pi_1$. Each interim analysis involves updating the posterior probability $\hat{p} = P(\mu_1 > \mu_0 \mid \text{data})$ and checking whether $\hat{p}$ exceeds a prescribed upper bound $p_U$ or falls below a prescribed lower bound $p_L$, which is less than $p_U$. If $\hat{p} > p_U$ (or $\hat{p} < p_L$), then the trial is terminated, rejecting (accepting) the null hypothesis that the experimental treatment is not better than the standard treatment; otherwise the study continues until the next interim analysis or until the scheduled end of the study. The posterior probabilities are computed by Markov chain Monte Carlo, and simulation studies of the frequentist operating characteristics under different scenarios are used to determine the maximum sample size, study duration, and the thresholds $p_L$ and $p_U$. Whereas Inoue et al. (2002) considered a more complex scenario in which $y_i$ is observable only if $\tau_i > t_0$, Huang et al. (2009) introduced a more elaborate design that uses the posterior probability $\hat{p}$ after an interim analysis for outcome-adaptive random allocation of patients to treatment arms until the next interim analysis. These Bayesian designs are called Phase II–III because they involve a small number of centers for Phase II after which "the decision of whether to stop early, continue Phase II, or proceed to Phase III with more centers is made repeatedly during a time interval."

While (6.58) provides a parametric approach to modeling the response–survival relationship using mixture of exponential survival times, semiparametric methods such as Cox regression are often preferred for reproducibility considerations and because of the relatively large sample sizes in Phase III studies. Efficient time-sequential methods to carry this out are already available, as shown in this chapter. Moreover, group sequential GLR tests for sample proportions are also available, as shown in Chap. 4. Lai et al. (2012a) combine these methods to develop an alternative seamless Phase II–III design that uses a semiparametric model to relate survival to response and is directly targeted toward frequentist testing with GLR or partial likelihood statistics. Their basic idea is to replace the stringent parametric model involving exponential distributions in (6.57) by a semiparametric counterpart that generalizes the Inoue–Thall–Berry model. Let $y$ denote the response and $z$ denote the treatment indicator, taking the value 0 or 1. Consider the proportional hazards model

$$\lambda(t\,|\,y,z) = \lambda_0(t)\exp(\alpha y + \beta z + \gamma yz). \tag{6.59}$$

The Inoue–Thall–Berry exponential model is a special case of (6.59), with $\lambda_0(\cdot)$ being the constant hazard rate of an exponential distribution. Let $\pi_0 = P(y = 1\,|\,\text{control})$ and $\pi_1 = P(y = 1\,|\,\text{treatment})$. Let $a = e^{\alpha}$, $b = e^{\beta}$, and $c = e^{\gamma}$, and let $S$ be the survival distribution and $f$ be the density function associated with the hazard function $\lambda_0$ so that $\lambda_0 = f/S$. From (6.59), it follows that the survival distribution of $\tau$ is

$$P(\tau > t) = \begin{cases} (1-\pi_0)S(t) + \pi_0(S(t))^a & \text{for the control group } (z = 0), \\ (1-\pi_1)(S(t))^b + \pi_1(S(t))^{abc} & \text{for the treatment group } (z = 1). \end{cases} \tag{6.60}$$

The hazard ratio of the treatment to control survival varies with $t$ because of the mixture form in (6.60). Let $\boldsymbol{\pi} = (\pi_0, \pi_1)$ and $\boldsymbol{\xi} = (a, b, c)$. A commonly adopted premise in the sequenced trials to develop and test targeted therapies of cancer is that the treatment's effectiveness on an early endpoint such as tumor response would translate into long-term clinical benefit associated with a survival endpoint such as progression-free or overall survival, and conversely, that failure to improve the early endpoint would translate into lack of definitive clinical benefit. This explains why the go/no-go decision for Phase III made in a conventional Phase II cancer trial is based on the response endpoint. Under this premise, the complement of the set of parameter values defining an efficacious treatment leads to the null hypothesis $H_0 : \pi_0 \geq \pi_1$, or $\pi_0 < \pi_1$ and $d(\boldsymbol{\pi}, \boldsymbol{\xi}) \leq 0$; see Lai et al. (2012a) for the expression and rationale of $d(\boldsymbol{\pi}, \boldsymbol{\xi})$ and how the modified Haybittle–Peto tests in Sects. 4.2 and 6.5.2 can be extended to test $H_0$.

4. Derive from (6.60) the hazard ratio of treatment to control at every $t$. Show that it is not constant in $t$ even when $S$ is exponential except for certain values of $(\boldsymbol{\pi}, \boldsymbol{\xi})$. On the other hand, show that as $a \to 1$ and $c \to 1$, the limiting hazard ratio is constant in $t$ and express the constant as a function of $\boldsymbol{\pi}$ and $\boldsymbol{\xi}$. In fact, this function is $1 - d(\boldsymbol{\pi}, \boldsymbol{\xi})$, where $d(\boldsymbol{\pi}, \boldsymbol{\xi})$ is the same as that given in Lai et al. (2012a).

# Chapter 7
# Confidence Intervals and *p*-Values

Although group sequential methods allow for early termination of a clinical trial while preserving the overall significance level of the test concerning its primary endpoint, they introduce substantial difficulties in constructing confidence intervals for the parameters of interest following the trial. The naive confidence interval that ignores the data-dependent nature of the sample size needs justification and may be unreliable. Section 7.1 gives an overview of several developments in the literature to address this issue. Most of them focus on the simple case of a normal mean $\mu$ with known variance, for which "exact methods" to construct confidence intervals for $\mu$ following a group sequential test of $\mu \leq 0$ (or of $\mu = 0$) are available.

For samples of fixed size, an important methodology for constructing confidence intervals without distributional assumptions is the bootstrap method. Chuang and Lai (1998) have studied bootstrap confidence intervals for a population mean in a group sequential setting. They have found that the bootstrap method does not give reliable confidence intervals in a group sequential setting and have developed a resampling method, called *hybrid resampling*, to construct confidence intervals whose coverage probabilities are nearly equal to the nominal ones. The term "hybrid" refers to a hybrid of the exact and bootstrap methods. Motivated by applications to time-sequential clinical trials with failure-time endpoints, Lai and Li (2006) have developed a general ordering scheme in the sample space of the observed (possibly censored) failure times up to a stopping time. This ordering scheme, which is described in Sect. 7.3, not only unifies previous exact methods based on *p*-values that are implicitly or explicitly associated with orderings of the sample space, but also resolves a long-standing difficulty due to two time scales in time-sequential trials noted in Chap. 6. Applying the ordering scheme and hybrid resampling to time-sequential survival data, Lai and Li (2006) circumvent this difficulty and construct confidence intervals with accurate coverage probabilities for hazard ratios in time-sequential clinical trials. Their results are summarized in Sect. 7.4.

In a sequential clinical trial whose stopping rule depends on the primary endpoint, inference on secondary endpoints is an important and long-standing

J. Bartroff et al., *Sequential Experimentation in Clinical Trials,*
Springer Series in Statistics 298, DOI 10.1007/978-1-4614-6114-2_7,

problem. Ignoring the possibility of early stopping based on the primary endpoint may result in substantial bias. To address this problem, a commonly used approach, described in Sect. 7.1.4, is to develop bias correction by estimating the bias in the case of bivariate normal outcomes and appealing to joint asymptotic normality of the statistics associated with the primary and secondary endpoints. Lai et al. (2009) have developed a new approach, described in Sect. 7.5, that uses hybrid resampling. This approach is shown in Sect. 7.6 to provide accurate inference in complex clinical trials, including time-sequential trials with survival endpoints and covariates.

## 7.1 Inference Following a Sequential Trial: Some Developments

### *7.1.1 Naive Confidence Intervals, Anscombe's Theorem, and Bayesian Methods*

For fully sequential tests such as the SPRT based on i.i.d. observations, Anscombe's (1952) theorem was used to prove the asymptotic validity of the naive confidence interval $(\hat{\theta}_T - \theta)/\widehat{\mathrm{se}}(\hat{\theta}_T)$, where $T$ is the stopping time and $\widehat{\mathrm{se}}(\hat{\theta}_T)$ denotes the estimated standard error. This is tantamount to replacing the fixed sample size $n$ by the random variable $T$ in the traditional confidence interval $(\hat{\theta}_n - \theta)/\widehat{\mathrm{se}}(\hat{\theta}_n)$ for an unknown parameter $\theta$. Anscombe's theorem states that if $Z_n := (\hat{\theta}_n - \theta)/\mathrm{se}(\hat{\theta}_n)$ has a limiting standard normal distribution and as $n \to \infty$ and $\varepsilon \uparrow 1$,

$$T_n/n \xrightarrow{P} c, \quad \widehat{\mathrm{se}}(\hat{\theta}_n)/\mathrm{se}(\hat{\theta}_n) \xrightarrow{P} 1, \quad \max_{\varepsilon n \le j \le n} |Z_n - Z_j| \xrightarrow{P} 0 \tag{7.1}$$

for some nonrandom constant $c$, then $(\hat{\theta}_{T_n} - \theta)/\widehat{\mathrm{se}}(\hat{\theta}_{T_n})$ also has a limiting standard normal distribution as $n \to \infty$. We embed the stopping time $T$ in a sequence of stopping times $T_n$ in (7.1) not only to indicate that the stopping time is large for the central limit theorem to hold but also to show that it is asymptotically nonrandom, behaving like $cn$. Siegmund (1985, p.22) has pointed out the inadequacy of this normal approximation because early stopping in the fully sequential test may result in a sample size not large enough to satisfy (7.1) unless an "artificially large" minimum sample size is enforced for the purpose of estimation, and because (7.1) cannot hold at certain parameter values, for example, $\theta$ where "$E_\theta\{\log[f_1(X)/f_0(X)]\} \approx 0$" for the SPRT.

Imposing a minimum sample size is not an issue for group sequential tests, but the condition $T_n/n \xrightarrow{P} c$ in (7.1) is violated at more parameter values. Chuang and Lai (1998) have shown that even though $\sqrt{n}(\bar{X}_n - \mu)$ is a pivot in the case of $X_i \sim N(\mu, 1)$, $\sqrt{T}(\bar{X}_T - \mu)$ is highly non-pivotal for group sequential stopping times. In the fully sequential case, Siegmund (1985, p.24) has suggested using Edgeworth-type asymptotic expansions to improve the coverage accuracy of the intervals, but notes that derivation of these expansions "seems remarkably difficult, even for fairly simple stopping rules." Lai and Wang (1994) have developed these expansions which are, however, too complicated for practical use because they involve fluctuation-theoretic quantities for random walks. Woodroofe (1986, 1992) has developed simpler expansions which will be described in Sect. 7.1.4 and the second supplement in Sect. 7.7, but these expansions are related to coverage probabilities integrated with respect to some prior distribution on the parameter space.

Woodroofe's use of these prior distributions is for technical reasons to simplify the expansions of the frequentist coverage probabilities, which involve adjustments for the randomness of the stopping time $T$. It is different from the Bayesian approach that considers credible sets instead of confidence sets. The credible sets do not need adjustments for the randomness of $T$ as they are defined by a prescribed level for the posterior probability, given the observations up to $T$, that $\theta$ belongs to the set. In particular, for the case $X_i \sim N(\mu, 1)$ and $\mu \sim N(0, \sigma^2)$, a $(1-\alpha)$-level credible interval for $\mu$ is

$$\frac{\sigma^2 T \bar{X}_T}{\sigma^2 T + 1} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{\sigma^2 T + 1}},$$

where $z_p = \Phi^{-1}(p)$ is the $p$th quantile of the standard normal distribution. Letting $\sigma \to \infty$ (corresponding to the flat prior) yields the naive confidence interval $\bar{X}_T \pm z_{1-\alpha/2}/\sqrt{T}$ for $\mu$. Since both the naive confidence interval and the Bayesian credible interval make no adjustments for the sampling fluctuations of $T$, it is not surprising that their frequentist coverage probabilities may differ substantially from the prescribed level.

### 7.1.2 *Exact Confidence Intervals in the Normal Case*

For the prototypical problem of $X_i \sim N(\mu, 1)$ in the literature, there is only a single unknown parameter $\mu$, so one can construct confidence intervals by the *exact method* that defines a $(1-\alpha)$-level lower confidence bound as the set of parameters $\mu$ for which a level-$\alpha$ test of $H_\mu : \mu' \geq \mu$ with the given stopping time $T$ accepts $H_\mu$; the $(1-\alpha)$-level upper confidence bound is defined similarly. Thus, a $(1-2\alpha)$-level confidence set can be defined by

$$\left\{\mu : u_\alpha(\mu) \leq \sqrt{T}(\bar{X}_T - \mu) \leq u_{1-\alpha}(\mu)\right\}, \tag{7.2}$$

where $u_\alpha(\mu)$ and $u_{1-\alpha}(\mu)$ are the quantiles for every fixed $\mu$:

$$P_\mu\left\{\sqrt{T}(\bar{X}_T - \mu) < u_\alpha(\mu)\right\} = \alpha = P_\mu\left\{\sqrt{T}(\bar{X}_T - \mu) > u_{1-\alpha}(\mu)\right\}. \tag{7.3}$$

Note that the confidence set (7.2) reduces to an interval whose endpoints are found by intersecting the line $\sqrt{T}(\bar{X}_T - \mu)$ with the curves $u_\alpha(\mu)$ and $u_{1-\alpha}(\mu)$ if there is only one intersection with each curve, which is the case commonly encountered in practice. Rosner and Tsiatis (1988) propose this exact method to construct confidence intervals for $\mu$ and to use recursive numerical integration as in Sect. 4.3.1 to determine the quantiles $u_\alpha(\mu)$ and $u_{1-\alpha}(\mu)$.

### 7.1.3 Siegmund's Ordering and Group Sequential Trials with Random Group Sizes

As noted in Sect. 4.1.3, the number of subjects available for analysis at the $i$th interim analysis is often unknown in advance, and the stopping time $T$ that takes values in $J = \{n_1, n_2, \ldots, n_k\}$ is therefore not completely specified to implement (7.3). This difficulty can be circumvented by using an appropriate ordering scheme of the sample space of $(T, S_T)$. Under a total ordering $\leq$ of the sample space, an exact $(1-2\alpha)$-level confidence interval for $\mu$ is $\mu_\alpha < \mu < \mu_{1-\alpha}$, where $\mu_c$ is the value of $\mu$ such that

$$P_\mu\{(T, S_T) > (t_0, s_0)\} = c, \tag{7.4}$$

in which $(t_0, s_0)$ denotes the observed value of $(T, S_T)$ and $>$ in (7.4) denotes $\not\leq$. Such confidence intervals were first introduced by Siegmund (1978) for stopping rules of the form

$$T = \min\{n \in J : S_n \geq b_n \text{ or } S_n \leq a_n\} \qquad (a_n < b_n). \tag{7.5}$$

He used the following ordering of the sample space of $(T, S_T)$ : $(t, s) > (\tilde{t}, \tilde{s})$ whenever (1) $t = \tilde{t}$ and $s > \tilde{s}$, or (2) $t < \tilde{t}$ and $s \geq b_t$, or (3) $t > \tilde{t}$ and $\tilde{s} \leq a_{\tilde{t}}$. Emerson and Fleming (1990) proposed to order $(T, S_T)$ according to $S_T/T$. Under their "sample mean ordering," $(t, s) > (\tilde{t}, \tilde{s})$ whenever $s/t > \tilde{s}/\tilde{t}$.

For Siegmund's ordering, the event $\{(T, S_T) > (t_0, s_0)\}$ in (7.4) only involves sample points that stop before or at $t_0$ unless $s_0 \leq a_{t_0}$, in which case the event reduces to $\{T = t_0, S_T > s_0\} \cup \{T > t_0\} \cup \{T < t_0, S_T > b_T\}$. Since $\{T > t_0\}$ is the complement of $\{T \leq t_0\}$, the sample sizes $n_j$ under Siegmund's ordering need only be specified for $j \leq j(t_0)$, that is, up to the stopping time $t_0$ which is equal to $n_{j(t_0)}$. We can therefore condition on $n_1, \ldots, n_{j(t_0)}$ in evaluating the probability in (7.4) when Siegmund's ordering is used and thereby still obtain an exact $1-2\alpha$ confidence interval for $\mu$ even when it is not known how the $n_j$ are generated for $j > j(t_0)$. This important property of Siegmund's ordering is not shared by the sample mean ordering, under which the event $\{(T, S_T) > (t_0, s_0)\}$ contains sample points with $T > t_0$ when $t_0$ is smaller than the largest allowable sample size $N = n_k$. Therefore, unless one imposes assumptions on the probability mechanism, which is typically unknown, generating the group sizes after $t_0$, one cannot evaluate the probability in (7.4). For the sample mean ordering, Emerson and Fleming (1990) proposed to assign the remaining $N - n_{j(t_0)}$ observations to a single group in evaluating the probability in (7.4) after conditioning on the observed $n_1, \ldots, n_{j(t_0)}$. This is tantamount to changing the number of groups from $k$ to $j(t_0) + 1$.

Instead of using recursive numerical integration, Siegmund (1978) uses analytic approximations to evaluate (7.4). For the case $\mu = 0$, note the similarity of (7.4) and (4.19) that considers the particular null hypothesis $\mu = 0$. Rosner and Tsiatis (1988) call the ordering of $(T, S_T)$ according to $\sqrt{T}(\bar{X}_T - \mu)$, which they used in (7.2), "the likelihood ratio ordering". This exact method for constructing $(1-2\alpha)$-level confidence intervals in the present normal setting amounts to running a family of sequential tests, one for each $\mu$, with stopping rule $T$.

### 7.1.4 Bias Correction for A Modified Pivot

Siegmund (1978), Emerson and Fleming (1990), and Whitehead (1986; 1992, Chap. 5) have introduced bias-corrected or unbiased estimators following sequential trials. Instead of using bias correction for point estimation, Woodroofe (1992) uses bias correction to correct the pivot for optional stopping in the case of a normal population with unknown mean $\mu$ and known variance 1. Suppose the stopping rule is of the form $T = \min\{n_0(a), \max(t_a, n_1(a))\}$, where $n_0(a) \sim a/\varepsilon_0$ and $n_1(a) \sim a/\varepsilon_1$, with $0 < \varepsilon_0 < \varepsilon_1$, and

$$t_a = \inf\{n \geq 1 : n g(S_n/n) \geq a\},$$

in which $g$ is continuously differentiable. Let $R_0(\mu) = T^{1/2}(\bar{X}_T - \mu)$. Noting that $a/T \xrightarrow{P} \kappa(\mu) := \max\{\varepsilon_0, \min(g(\mu), \varepsilon_1)\}$, Woodroofe (1986) has shown that

$$ER_0(\mu) \doteq a^{-1/2}[(d/d\mu)\kappa^{1/2}(\mu)] = (\kappa(\mu)/a)^{1/2} b(\mu),$$

where $b(\mu) = [(d/du)\kappa^{1/2}(\mu)]\kappa^{-1/2}(\mu) = \dot{\kappa}(\mu)/\{2\kappa(\mu)\}$. This suggests the bias-corrected pivot

$$R_1(\mu) = T^{1/2}(\bar{X}_T - \mu) - T^{-1/2} b(\bar{X}_T). \tag{7.6}$$

Section 7.7 gives a (nonparametric) multivariate extension of the modified pivot (7.6) and Woodroofe's theory for confidence intervals based on (7.6).

### 7.1.5 Bivariate Normal Outcomes

A long-standing problem in the terminal analysis of sequential clinical trials is testing secondary hypotheses. Usually such hypotheses are concerned with parameters associated with secondary endpoints, whereas the stopping rule of the trial depends on the primary endpoint. When the primary and secondary endpoints are correlated, conventional nonsequential inference on a secondary endpoint that ignores the sequential design for the primary endpoint is invalid. To address this problem, Liu et al. (2000) and Whitehead et al. (2000) propose to use corrections for

**Table 7.1** Quantiles $q_\alpha$ of $\sqrt{T}(\bar{Y}_T - \theta)/\sigma$

| $\alpha$ (in %) | 2.5 | 5 | 10 | 20 | 50 | 80 | 90 | 95 | 97.5 |
|---|---|---|---|---|---|---|---|---|---|
| $z_\alpha$ | −1.96 | −1.65 | −1.28 | −0.84 | 0.00 | 0.84 | 1.28 | 1.65 | 1.96 |
| $q_\alpha(\mu = 0.0)$ | −2.12 | −1.72 | −1.31 | −0.84 | 0.01 | 0.86 | 1.31 | 1.68 | 2.04 |
| $q_\alpha(\mu = 0.5)$ | −2.08 | −1.69 | −1.29 | −0.85 | 0.03 | 0.97 | 1.46 | 1.85 | 2.17 |
| $q_\alpha(\mu = 1.0)$ | −1.97 | −1.64 | −1.28 | −0.82 | 0.15 | 1.06 | 1.50 | 1.84 | 2.14 |
| $q_\alpha(\mu = 1.5)$ | −1.89 | −1.56 | −1.15 | −0.62 | 0.24 | 1.04 | 1.43 | 1.78 | 2.05 |
| $q_\alpha(\mu = 2.0)$ | −1.74 | −1.41 | −1.04 | −0.61 | 0.20 | 0.99 | 1.40 | 1.75 | 2.06 |

the bias of the randomly stopped test statistic, based on estimation of the bias in the prototypical case of bivariate normal endpoints and on joint asymptotic normality of their test statistics for the primary and secondary hypotheses.

Let $(X_i, Y_i)$ be independent bivariate normal with $E(Y_i) = \theta$, $\text{Var}(Y_i) = \sigma^2$, $E(X_i) = \mu$, $\text{Var}(X_i) = 1$, and $\text{Corr}(X_i, Y_i) = \rho$, where $\rho$ and $\sigma$ are known and nonzero, and $\mu$ and $\theta$ are unknown. Consider the problem of testing $H_0 : \theta \leq 0$ based on a randomly stopped sample $\{(X_i, Y_i), 1 \leq i \leq T\}$, where $T$ is a stopping rule based on $X_1, X_2, \ldots$. If $T$ is replaced by a fixed sample size $n$, then $\sqrt{n}(\bar{Y}_n - \theta)/\sigma$ is standard normal. However, $\sqrt{T}(\bar{Y}_T - \theta)/\sigma$ has a non-normal distribution, which depends on $\mu$ because the sampling distribution of $T$ depends on $\mu$, as illustrated by Lai et al. (2009) in which $T = \inf\{m \leq 75 : X_1 + \cdots + X_m > 2.413\sqrt{m}$, $m$ is divisible by $15\}$, and $\rho = 0.8$; see Table 7.1. Therefore, whereas the likelihood ratio test based on a sample of fixed size $n$ rejects $H_0$ if the sample mean $\bar{Y}_n$ exceeds $\sigma z_{1-\alpha}/\sqrt{n}$, the test with stopping rule $T$ should replace $z_{1-\alpha}$ by $d_{1-\alpha}$ that is defined by

$$\sup_\mu P_{\mu,0}\left\{\sqrt{T}\bar{Y}_T/\sigma \geq d_{1-\alpha}\right\} = \alpha, \tag{7.7}$$

because $H_0$ is composite and the type I error probability constraint is $\sup_{\mu,\theta\leq 0} P_{\mu,\theta}$ $\{$the test rejects $H_0\} \leq \alpha$. Since the test rejects $H_0$ if $\sqrt{T}\bar{Y}_T \geq d_{1-\alpha}\sigma$, the $p$-value of the test is

$$\sup_\mu P_{\mu,0}\left\{\sqrt{T}\bar{Y}_T \geq \sqrt{t}\bar{y}_t\right\}, \tag{7.8}$$

in which $(t, \bar{y}_t)$ is the observed value of $(T, \bar{Y}_T)$ in the sample. This is the *exact method*, which is a generalization of that in Sect. 7.1.2 for the univariate case in which the $Y_i$ are absent. Instead of solving for $d_{1-\alpha}$, it is more convenient to compute (7.8) and to reject $H_0$ if (7.8) does not exceed $\alpha$. The probability in (7.8) for a given $\mu$ can be computed by using the recursive numerical integration method of Armitage et al. (1969). A standard numerical optimization algorithm can then be used to maximize the computed probability over $\mu$.

Instead of replacing the incorrect quantile $z_{1-\alpha}$ by its correct version in (7.7), Liu et al. (2000) propose to adhere to $z_{1-\alpha}$ but to modify $\sqrt{T}\bar{Y}_T/\sigma$ by a bias-corrected modification $\sqrt{T}(\bar{Y}_T - \rho\sigma\hat{b}_T)/\sigma$, where $\hat{b}_T = \bar{X}_T - \hat{\mu}_T$ and $\hat{\mu}_T$ is an unbiased estimate of $\mu$ given by Emerson and Fleming (1990) and Liu and Hall (1999). Their basic idea is that the normal approximation would be applicable to $\sqrt{T}(\bar{Y}_T - \rho\sigma\hat{b}_T - \theta)/\sigma$ so that $z_{1-\alpha}$ can still be used. Whitehead et al. (2000)

use a different approach that involves estimating the mean $m_U$ and the standard deviation $s_U$ of $U = \sqrt{T}(\bar{X}_T - \mu)$ and using the independence between $\{X_1, X_2, \dots\}$ and $\{Y_i - \rho\sigma X_i : i \geq 1\}$ to form an approximate pivot

$$\frac{\{\sqrt{T}(\bar{Y}_T - \theta)/\sigma - \rho m_U\}}{\{1 + \rho^2(s_U^2 - 1)\}^{1/2}}.$$

As shown by Todd and Whitehead (1996), $m_U$ and $s_U$ can be estimated by using $\bar{X}_T$ to replace $\mu$ in a numerical integration procedure that assumes $\mu$ to be known. Although Whitehead et al. (2000) only use this approximate pivot to construct confidence intervals for $\theta$, it can also be used to test $H_0 : \theta \leq 0$.

### *7.1.6 Extensions Beyond the Normal Cases*

The traditional literature on group sequential designs focuses on the prototypical problem of testing for the mean $\mu$ of a normal distribution with known variance, usually assumed to be 1 after normalization, or more generally the mean vector $\boldsymbol{\mu}$ of a multivariate normal distribution with known covariance matrix. For more general test statistics and parameters, the group sequential nature is used to justify normal approximation of the increments of the test statistics within successive groups, thereby reducing the problem to that of normal increments; see Jennison and Turnbull (2000, Chap. 8), Liu et al. (2000), and Whitehead et al. (2000) that focus primarily on ordering schemes and bias-corrected pivots for group sequential tests, and confidence intervals following the tests, in the normal and bivariate normal cases.

## 7.2 A Hybrid Resampling Approach

### *7.2.1 A General Formulation of Exact, Bootstrap, and Hybrid Resampling Methods*

Chuang and Lai (2000) have provided the following general framework for the statistical problem of constructing confidence intervals. Let $\boldsymbol{X}$ be a vector of observations from distribution $F$ in some family $\mathscr{F}$ of distributions. For nonparametric problems, $\mathscr{F}$ is the family of distributions satisfying certain prespecified regularity conditions. For parametric models with parameter $\eta \in \Gamma$, we can denote $\mathscr{F}$ by $\{F_\eta : \eta \in \Gamma\}$. The problem of interest is to find a confidence interval for the real-valued parameter $\theta = \theta(F)$. Let $\Theta$ denote the set of all possible values of $\theta$.

*Exact method:* If $\mathscr{F} = \{F_\theta : \theta \in \Theta\}$ is indexed by a real-valued parameter $\theta$, an exact equal-tailed confidence region can always be found by using the well-known duality between hypothesis tests and confidence regions, which we have described in Sect. 7.1.2 for the special case of $F_\theta = N(\theta, 1)$. Suppose one would like to test the null hypothesis that $\theta$ is equal to $\theta_0$. Let $R(\boldsymbol{X}, \theta_0)$ be some real-valued test statistic. Let $u_\alpha(\theta_0)$ be the $\alpha$-quantile of the distribution of $R(\boldsymbol{X}, \theta_0)$ under the distribution $F_{\theta_0}$. The null hypothesis is accepted if $u_\alpha(\theta_0) < R(\boldsymbol{X}, \theta_0) < u_{1-\alpha}(\theta_0)$. An exact equal-tailed confidence region with coverage probability $1-2\alpha$ consists of all $\theta_0$ not rejected by the test and is therefore given by

$$\{\theta : u_\alpha(\theta) < R(\boldsymbol{X}, \theta) < u_{1-\alpha}(\theta)\}. \tag{7.9}$$

*Bootstrap method:* The exact method applies only when there are no nuisance parameters and this assumption is rarely satisfied in practice. The bootstrap method replaces the quantiles $u_\alpha(\theta)$ and $u_{1-\alpha}(\theta)$ by the approximate quantiles $u^*_\alpha$ and $u^*_{1-\alpha}$ obtained in the following manner. Based on $\boldsymbol{X}$, construct an estimate $\hat{F}$ of $F \in \mathscr{F}$. The quantile $u^*_\alpha$ is defined to be $\alpha$-quantile of the distribution of $R(\boldsymbol{X}^*, \hat{\theta})$ with $\boldsymbol{X}^*$ generated from $\hat{F}$ and $\hat{\theta} = \theta(\hat{F})$. Thus, the bootstrap method yields the following confidence region for $\theta$ with approximate coverage probability $1-2\alpha$:

$$\left\{\theta : u^*_\alpha < R(\boldsymbol{X}, \theta) < u^*_{1-\alpha}\right\}. \tag{7.10}$$

In particular, when $\hat{F}$ is the empirical distribution of i.i.d. $X_1, \ldots, X_n$ and the root $R(\boldsymbol{X}, \theta)$ is equal to $(\hat{\theta} - \theta)/\hat{\sigma}$ for some estimate $\hat{\sigma}$ of the standard error of $\hat{\theta}$, the bootstrap confidence interval (7.10) is called the bootstrap-$t$ interval.

*Hybrid resampling method:* The hybrid confidence region is based on reducing the family of distributions $\mathscr{F}$ to another family of distributions $\{\hat{F}_\theta : \theta \in \Theta\}$, which is used as the "resampling family" and in which $\theta$ is the unknown parameter of interest. This reduction depends on $\boldsymbol{X}$, and some ways for carrying it out are given in the rest of this chapter. Let $\hat{u}_\alpha(\theta)$ be the $\alpha$-quantile of the sampling distribution of $R(\boldsymbol{X}, \theta)$ under the assumption that $\boldsymbol{X}$ has distribution $\hat{F}_\theta$. The hybrid confidence region results from applying the exact method to $\{\hat{F}_\theta : \theta \in \Theta\}$ and is given by

$$\{\theta : \hat{u}_\alpha(\theta) < R(\boldsymbol{X}, \theta) < \hat{u}_{1-\alpha}(\theta)\}. \tag{7.11}$$

The construction of (7.11) typically involves simulations to compute the quantiles as in the bootstrap method. Chuang and Lai (1998, 2000) call this the *hybrid resampling* method because it "hybridizes" the exact method (that uses test inversion) with the bootstrap method (that uses the observed data to determine the resampling distribution). Note that hybrid resampling is a generalization of the bootstrap method, which uses the singleton $\{\hat{F}\}$ as the resampling family $\{\hat{F}_\theta\}$.

In practice, it is often desirable to express a confidence set for $\theta$ as an interval. Although (7.9), (7.10), and (7.11) may not be intervals, it often suffices to give only the upper and lower limits of the confidence set. Chuang and Lai (2000) describe an algorithm, based on method of successive secant approximations, to find the upper limit of (7.11). Let $f(\theta) = R(\boldsymbol{X}, \theta) - \hat{u}_\alpha(\theta)$ and consider solving the equation $f(\theta) = 0$. First, we find $a_1 < b_1$ such that $f(a_1) > 0$ and $f(b_1) < 0$. Let $f_1(\theta)$ be linear in $\theta \in [a_1, b_1]$ with $f_1(a_1) = f(a_1)$ and $f_1(b_1) = f(b_1)$, and $\theta_1$ be the root of $f_1(\theta) = 0$. If $f(\theta_1) > 0$, set $a_2 = \theta_1$ and $b_2 = b_1$. If $f(\theta_1) < 0$, set $b_2 = \theta_1$ and $a_2 = a_1$. Proceeding inductively in this manner, we let $f_k(\theta)$ linearly interpolate $f(a_k)$ and $f(b_k)$ for $a_k \le \theta \le b_k$, and let $\theta_k \in (a_k, b_k)$ be the root of $f_k(\theta) = 0$. This procedure terminates if $\theta_k$ differs little from $\theta_{k-1}$ and the terminal value $\theta_k$ is taken to be the upper limit of (7.11). Typically $f(\hat{\theta}) > 0$, so $\hat{\theta}$ can be chosen as $a_1$. To find $b_1$, one can start with $b_1' = \hat{\theta} + 2\hat{\sigma}$, where $\hat{\sigma}$ is an estimate of the standard error of $\hat{\theta}$. If $f(b_1') < 0$, let $b_1 = b_1'$; otherwise let $b_2' = b_1' + \hat{\sigma}/2$ and check whether $f(b_2') < 0$. This procedure is repeated until one arrives at $f(b_h') < 0$ and sets $b_1 = b_h'$. The quantiles $\hat{u}_\alpha(\theta_j)$ can be computed from independent samples from $\hat{F}_{\theta_j}$, as was done in these examples. It is often possible to try to reuse the same random sample for all $\theta$ values.

### 7.2.2 *Hybrid Resampling Confidence Intervals for Population Means Following Group Sequential Tests Based on Sample Means*

The hybrid resampling method provides a way to relax the assumption of normally distributed $X_i$ in the exact method of Sect. 7.1.2 for constructing confidence intervals for the mean $\mu$ of $X_i$ that has known variance 1. Let $G$ denote the common distribution of $X_i - \mu$, which has mean 0 and variance 1. An obvious estimate of $G$ is the empirical distribution $\hat{G}_T$ of $(X_i - \bar{X}_T)/\hat{\sigma}_T$ $(1 \le i \le T)$, where $\hat{\sigma}^2 = T^{-1}\sum(X_i - \bar{X}_T)^2$. Let $\varepsilon_1, \varepsilon_2, \dots$ be independent with common distribution $\hat{G}_T$, and let $X_i(\mu) = \mu + \varepsilon_i$. Let $T_\mu$ be the stopping rule $T$ applied to $X_1(\mu), X_2(\mu), \dots$, instead of to $X_1, X_2, \dots$. By analogy with (7.3), define the quantiles $\hat{u}_\alpha(\mu)$ and $\hat{u}_{1-\alpha}(\mu)$ of the distribution of $(\varepsilon_1 + \dots + \varepsilon_{T_\mu})/\sqrt{T_\mu}$ given $\hat{G}_T$. An approximate $1 - 2\alpha$ confidence set is

$$\left\{\mu : \hat{u}_\alpha(\mu) \le \sqrt{T}\,(\bar{X}_T - \mu) \le \hat{u}_{1-\alpha}(\mu)\right\}. \tag{7.12}$$

For every fixed $\mu$, the quantiles $\hat{u}_\alpha(\mu)$ and $\hat{u}_{1-\alpha}(\mu)$ in (7.12) can be computed by simulation.

A simpler alternative to this resampling method is the bootstrap method. Instead of using the empirical distribution $\hat{G}_T$ of the $(X_i - \bar{X}_T)/\hat{\sigma}_T$ to generate $\varepsilon_i$ and thereby to form $X_i(\mu) = \varepsilon_i + \mu$, the bootstrap method uses the empirical distribution $\hat{F}_T$ of $X_i\,(1 \le i \le T)$ and generates $X_1^*, X_2^*, \dots, X_{T^*}^*$ directly from $\hat{F}_T$, where $T^*$ is the stopping rule $T$ applied to $X_1^*, X_2^*, \dots$, instead of to $X_1, X_2, \dots$. Let $u_\alpha^*$ and $u_{1-\alpha}^*$ denote the $\alpha$- and $(1-\alpha)$-quantiles of the distribution of $\sqrt{T^*}(\bar{X}_{T^*}^* - \bar{X}_T)/\hat{\sigma}_T$, which can be computed from $\hat{F}_T$ by simulation. The bootstrap confidence interval is given by

$$\bar{X}_T - u^*_{1-\alpha}/\sqrt{T} \le \mu \le \bar{X}_T - u^*_\alpha/\sqrt{T}. \tag{7.13}$$

The bootstrap method is known to give second-order accurate confidence intervals when the stopping rule $T$ is replaced by a fixed sample size $n$. In fact, under $\hat{F}_n$, the distribution $\sqrt{n}(\bar{X}^*_n - \bar{X}_n)/\hat{\sigma}_n$ differs from the standard normal distribution, which is the distribution of $\sqrt{n}(\bar{X}_n - \mu)$ in the present example of normal $X_i$, by an $O_p(n^{-1})$ term (Hall, 1992). However, this asymptotic theory is no longer valid when $n$ is replaced by $T$ in the group sequential setting, and $\sqrt{T}(\bar{X}_T - \mu)$ is no longer an approximate pivot since its distribution changes substantially with $\mu$, as will be shown in a more general context in Sect. 7.4.1.

The preceding hybrid resampling scheme assumes that the stopping rule $T$ is completely specified. For group sequential stopping rules, this implies prespecified sample sizes $n_1, \dots, n_k$ at interim monitoring times. As noted in Sect. 7.1.3, Siegmund's ordering scheme can be used for normal $X_i$ even when the $n_j$ are not prespecified. Without assuming the $X_i$ to be normal, we approximate $P_\mu\{(T, S_T) \ge (t_0, s_0)\}$ by $P\{(T_\mu, S_{T_\mu}(\mu)) \ge (t_0, s_0)\}$, where $S_n(\mu) = (\mu + \varepsilon_1) + \cdots + (\mu + \varepsilon_n)$ and $\varepsilon_i$ and $T_\mu$ are defined above. Thus an approximate $(1-2\alpha)$-level confidence interval is $\hat{\mu}_\alpha \le \mu \le \hat{\mu}_{1-\alpha}$, where $\hat{\mu}_c$ is the value of $\mu$ for which

$$P\left\{\left(T_\mu, S_{T_\mu}(\mu)\right) \ge (t_0, s_0)\right\} = c. \tag{7.14}$$

It can be shown that the probability in (7.14) is an increasing function of $\mu$. This probability can be computed for any fixed $\mu$ using simulation by generating the $\varepsilon_i$ from the empirical distribution $\hat{G}_T$ of $(X_j - \bar{X}_T)/\hat{\sigma}_T$ $(1 \le j \le T)$.

### 7.2.3 A Comparative Study

Consider the stopping rule (7.5) with $b_n = a_n = 2.413\sqrt{n}$ and $J = \{n_1, \dots, n_5\}$, in which $n_1, n_2 - n_1, n_3 - n_2, n_4 - n_3$ are independent and uniformly distributed on $\{16, 17, \dots, 24\}$ and that $n_5 = 100$. Table 7.2a considers the case of normally distributed $X_i$ with unknown mean $\mu$ and known variance 1. The table gives the coverage errors of upper and lower confidence limits for $\mu$ obtained by different methods with a common nominal coverage error probability $\alpha = 0.05$. It shows that the hybrid resampling method, which uses 2000 simulations to compute the probability in (7.14) by Monte Carlo, has coverage errors that are close to those of Siegmund's exact method which is based on the assumption of normal $X_i$. Also given for comparison are the Emerson–Fleming method, which also uses the Gaussian assumption on the $X_i$, and the naive normal confidence limits that treat $\sqrt{T}(\bar{X}_T - \mu)$ as if it were normal. The coverage errors of the naive normal confidence limits in Table 7.2a differ substantially from the nominal value of 5%, and the Emerson–Fleming upper confidence limit shows a relatively large coverage error of 7.5% at $\mu = \frac{3}{4}$. Whereas Siegmund's method is exact for normal $X_i$ and the Emerson–Fleming method is also based on the assumption of normally distributed $X_i$, the

**Table 7.2** Coverage errors in percentages for four types of lower, L, and upper, U, confidence limits. (S, Siegmund's method; H, hybrid resampling method; EF, Emerson–Fleming methods; N, naive normal method)

| | $\mu=0$ | | $\mu=1/8$ | | $\mu=1/4$ | | $\mu=1/2$ | | $\mu=3/4$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | L | U | L | U | L | U | L | U | L | U |
| (a) Normal mean | | | | | | | | | | |
| S | 5.48 | 4.84 | 5.24 | 4.86 | 4.78 | 4.92 | 5.22 | 4.88 | 5.26 | 4.28 |
| H | 5.62 | 4.96 | 5.14 | 4.76 | 4.82 | 5.00 | 5.34 | 4.94 | 5.32 | 4.30 |
| EF | 5.48 | 4.84 | 5.36 | 4.86 | 4.98 | 4.92 | 5.24 | 5.68 | 5.26 | 7.46 |
| N | 6.48 | 5.86 | 10.54 | 5.04 | 7.10 | 4.86 | 5.38 | 1.56 | 5.28 | 3.04 |
| (b) Non-normal mean | | | | | | | | | | |
| S | 6.36 | 3.50 | 6.70 | 4.34 | 5.94 | 4.38 | 6.52 | 4.38 | 6.36 | 4.00 |
| H | 5.44 | 4.60 | 5.66 | 5.10 | 5.62 | 4.96 | 5.96 | 4.76 | 5.88 | 4.70 |
| EF | 6.34 | 3.50 | 6.92 | 4.34 | 6.46 | 4.38 | 6.66 | 5.40 | 6.36 | 7.84 |
| N | 7.46 | 4.50 | 12.28 | 4.38 | 9.14 | 4.32 | 7.22 | 2.26 | 6.40 | 2.66 |

hybrid resampling method is nonparametric in nature. Table 7.2b considers the case in which $X_i - \mu + 1$ is exponential with mean 1, so that $E_\mu(X_i) = \mu$ and $\mathrm{Var}_\mu(X_i) = 1$. It shows that the hybrid resampling method still yields coverage errors close to the nominal value of 5% and that the other methods perform markedly worse than their counterparts in Table 7.2a. Each result in Table 7.2 is based on 5000 simulations.

## 7.3 A General Ordering Scheme and $p$-Values

### *7.3.1 Total Ordering of Sample Space for p-Values*

Section 7.1.3 has described different ordering schemes of the sample space of $(T, S_T)$. Since an exact method for constructing confidence regions is based on inverting a test, such a method is implicitly or explicitly linked to an ordering of the sample space of the test statistic used. The ordering defines the $p$-value of the test as the probability (under the null hypothesis) of more extreme values (under the ordering) of the test statistic than that observed in the sample. Equivalently, the test rejects the null hypothesis, one for each given $\mu$, if the test statistic exceeds or falls below a specified quantile of its null distribution. Thus, the ordering scheme for $(T, S_T)$ in the exact methods of Siegmund (1978) and Rosner and Tsiatis (1988) can be associated with corresponding bivariate quantiles of $(T, S_T)$. Under a total ordering $\leq$ of the sample space of $(T, S_T)$, Lai and Li (2006) call $(t, s)$ a $q$th quantile if

$$P\{(T, S_T) \leq (t, s)\} = q, \tag{7.15}$$

under the assumption that the $X_i$ have a strictly increasing continuous distribution function, as in the normal case. This is a natural generalization of the $q$th quantile of a univariate random variable. For randomly stopped sums of independent normal random variables with unknown mean $\mu$, the bivariate vector $(T, S_T)$ is sufficient for $\mu$. For the general setting where a stochastic process $\boldsymbol{X}_u$, in which $u$ denotes either discrete or continuous time, is observed up to a stopping time $T$, Lai and Lai (2006) define $\boldsymbol{x} = \{\boldsymbol{x}_u : u \leq t\}$ to be a $q$th quantile if

$$P\{\boldsymbol{X} \leq \boldsymbol{x}\} \geq q, \qquad P\{\boldsymbol{X} \geq \boldsymbol{x}\} \geq 1-q, \tag{7.16}$$

under a total ordering $\leq$ for the sample space of $\boldsymbol{X} = \{\boldsymbol{X}_u : u \leq T\}$.

For applications to confidence intervals of a real parameter $\theta$, the choice of the total ordering should be targeted toward the objective of interval estimation. Let $\{U_r : r \leq T\}$ be real-valued statistics based on the observed process $\{\boldsymbol{X}_s : s \leq T\}$. For example, let $U_r$ be an estimate of $\theta$ based on $\{\boldsymbol{X}_s : s \leq r\}$. A total ordering on the sample space of $\boldsymbol{X}$ can be defined via $\{U_r : r \leq T\}$ as follows:

$$\boldsymbol{X} \geq \boldsymbol{x} \quad \text{if and only if} \quad U_{T\wedge t} \geq u_{T\wedge t}, \tag{7.17}$$

where $T \wedge t = \min(T,t)$ and $\{u_r : r \leq t\}$ is defined from $\boldsymbol{x} = \{\boldsymbol{x}_r : r \leq t\}$ in the same way as $\{U_r : r \leq T\}$ is defined from $\boldsymbol{X}$.

In particular, consider the case of independent normal $X_i$, and let $U_n$ be the sample mean $\bar{X}_n$ of $X_1, \ldots, X_n$. Whereas the sample mean ordering of Emerson and Fleming (1990) defines $(T, S_T) \geq (t, s_t)$ by $S_T/T \geq s_t/t$, (7.17) yields the somewhat different ordering

$$(T, S_T) \geq (t, s_t) \quad \text{if and only if} \quad \bar{X}_{T\wedge t} \geq s_{T\wedge t}/(T \wedge t). \tag{7.18}$$

With $(t, s_t)$ being the observed sample values, note that (7.18) is equivalent to $S_{T\wedge t} \geq s_{T\wedge t}$, which is the same as Siegmund's ordering for stopping rules $T$ of the type (7.5). Thus (7.17) can be considered as a generalization of Siegmund's ordering; it also relates Siegmund's ordering to the intuitively appealing ordering via sample means advocated by Emerson and Fleming (1990). Moreover, if $U_r = \sqrt{r}(\bar{X}_r - \mu_0)$, then (7.18) again holds, and the modified form (7.17) of the likelihood ratio ordering of Rosner and Tsiatis (1988) is again equivalent to Siegmund's ordering. The original Rosner–Tsiatis ordering requires $n_1, \ldots, n_k$ (or the stochastic mechanism generating them) to be completely specified; see Fig. 7.1. It has the same difficulties as the Emerson–Fleming ordering described in the last paragraph of Sect. 7.1.3 if this is not the case.

Like Siegmund's ordering, (7.17) has the attractive feature that the probability mechanism generating $X_t$ needs only to be specified up to the stopping time $T$ in order to define the quantile. For example, consider the case of a Wiener process $\{W_t : t \geq 0\}$ with drift coefficient $\theta$, and let $\tau(1) < \tau(2) < \ldots$ be a sequence of positive random variables that are independent of $\{W_t : t \geq 0\}$. Let $X_n = W_{\tau(n)}$ and let $T$ be a stopping time for $\{X_n : n \geq 1\}$. Given the values of $\tau(1), \ldots, \tau(T)$, we do not need to know the stochastic mechanism generating the $\tau(n)$ for $n > T$ in order to

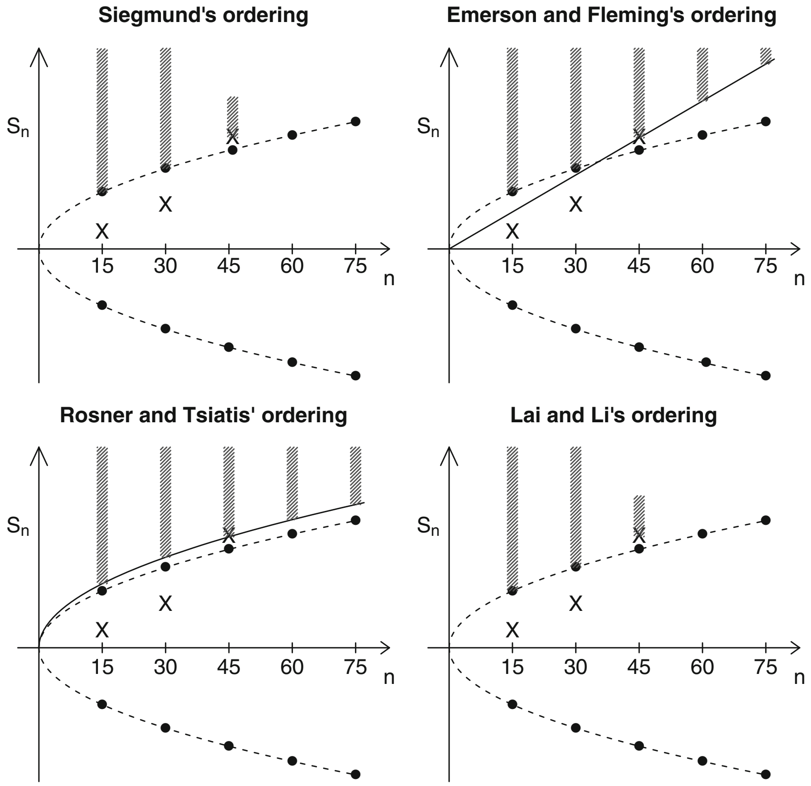


**Fig. 7.1** More extreme values (*shaded*) of $(T,S_T)$ than the observed (marked by X) under different orderings of the sample space when the group sizes are prespecified

compute the $q$th quantile of $\{X_n : n \le T\}$ under the ordering (7.17) with for example, $U_n = W_{\tau(n)}/\tau(n)$. This is an important advantage of (7.17) that will be used in the next section.

### 7.3.2 Bootstrap Methods for p-Values and Hybrid Resampling

In the second paragraph of Sect. 4.3.3, we have described what is tantamount to a $p$-value implementation of the modified Haybittle–Peto test of $\theta = \theta_0$ in (4.19), or more generally, of $u(\boldsymbol{\theta}) = u_0$, by the bootstrap method in Sect. 4.3.2. We have pointed out in Sect. 7.2.1 that hybrid resampling is a hybrid of exact and bootstrap methods by noting that it captures the essential feature of both methods. Here we elaborate further on this point. The exact method is related to inverting a family of tests, one for each $\theta$, to construct confidence intervals. A test, implemented through $p$-values, involves an ordering of the sample space. Hybrid resampling basically uses the bootstrap method to compute the $p$-values of these tests, one for each fixed $u_0$ to determine if $u_0$ should be included in the confidence set; see (7.12) for the case of the mean $\mu$ $(= u_0)$ of a normal distribution.

## 7.4 Hybrid Resampling Approach for Secondary Endpoints

### *7.4.1 Bivariate Mean Vectors*

To extend their approach beyond the bivariate normal case, Liu et al. (2000) apply their bias-correction method to other statistics than sample means that converge weakly, after normalization, to a bivariate Wiener process with correlation coefficient $\rho$. Lai et al. (2009) have developed an alternative approach which is based on a generalization of (7.8) for the $p$-value of an exact test in parametric models and which uses the hybrid resampling method in Sect. 7.2 to extend the approach to nonparametric settings. To begin with, suppose the primary and secondary endpoints $X$ and $Y$ have a joint density function $f_{\mu,\theta}$ such that the marginal distribution of $Y$ depends only on $\theta$ while that of $X$ depends on $\mu$. Let $U_n = U_n(Y_1,\ldots,Y_n)$ be a test statistic of $H_0 : \theta = \theta_0$ based on a sample $\{(X_i,Y_i), 1 \le i \le n\}$. Let $T$ be a stopping time whose distribution depends on $\mu$, as is the case in which stopping is determined by $X_1, X_2, \ldots$. The $p$-value of a one-sided test of $H_0$ is $\sup_\mu P_{\mu,\theta_0}(U_T \ge u_t)$, in which $(t,u_t)$ is the observed value of $(T,U_T)$. If a fixed sample size $n$ was used instead of $T$, then the $p$-value would be $P_{\theta_0}(U_n \ge u_n)$ since the distribution of $U_n$ does not depend on $\mu$. We can therefore regard the supremum over $\mu$ as an adjustment for using a random $T$ whose distribution depends on $\mu$. This adjustment can be carried out by using numerical integration or Monte Carlo simulations to evaluate the function $P_{\mu,\theta_0}(U_T \ge u_t)$.

*Example 7.1.* To compare the power functions of the tests of Liu et al. (2000) and Whitehead et al. (2000) with those of the exact test whose $p$-value is given by (7.8), Lai et al. (2009) simulate group sequential trials with up to five looks at sample sizes 15, 30, 45, 60, and 75, from the bivariate normal distribution with mean $(\mu,\theta)$, $\mathrm{Var}(X_i) = \mathrm{Var}(Y_i) = 1$, and $\mathrm{Corr}(X_i,Y_i) = \rho = 0.8$. Letting $S_n = X_1 + \cdots + X_n$, the trial is stopped at sample size $T \in \{15,30,45,60,75\}$ if $S_T \ge 2.413\sqrt{T}$ or if $T = 75$, as in Table 7.1. Table 7.3a gives the type I error probability and power of five tests of $H_0 : \theta \le 0$, including the tests of Liu et al. (2000) and Whitehead et al. (2000), the exact test defined by (7.8), the bootstrap, and hybrid resampling tests. Each entry in Table 7.3 is based on 2000 simulations, with the same simulated datasets for each method. It shows that the exact test and the hybrid resampling test maintain the type I error probability, while the other tests inflate it. The type I error probability is not constant across all values of the primary parameter; instead, the nominal type I error probability 0.05 is achieved at the least favorable values. Moreover, the power of the hybrid resampling test is comparable to that of the exact test. The naive normal test that ignores early stopping has also been considered in the simulation study. It inflates the type I error probability even more, and its results are not included in the table.

### *7.4.2 Bivariate Nonparametric Functionals*

Without assuming a parametric model under which the distribution of $U_T$ can be evaluated, for example, by Monte Carlo, when the parameter values are given, we can proceed nonparametrically by using hybrid resampling instead of parametric Monte Carlo. In the nonparametric setting, $\theta$ is some functional $\theta(G)$ of the distribution function $G$ of $Y$, and $\mu = \mu(F)$ is some functional of the distribution function $F$ of $X$ that determines the distribution of $T$. The joint distribution of $(X,Y)$ under the constraints $\theta(G) = \theta_0$ and $\mu(F) = \mu$ can be estimated from $\{(X_i,Y_i), 1 \leq i \leq T\}$ by empirical likelihood (see see Sect. 3.2 of Chuang and Lai, 2000) or by simpler methods that use the particular structure of $\mu(F)$ and $\theta(G)$ for the problem at hand. Whereas the exact test in the second paragraph of Sect. 7.1.5 uses the probability measure $P_{\mu,\theta_0}$ and the bootstrap approach samples from $P_{\hat{F},\theta_0}$, the hybrid resampling approach uses the empirical measure $\hat{P}_{\mu,\theta_0}$ from which $(X_i^*,Y_i^*)$, $1 \leq i \leq T^*$, are drawn to form the hybrid resample, yielding a hybrid of the exact and bootstrap tests, as in Sect. 7.2.

The hybrid resampling approach can be extended to handle more general situations in which the observations are independent random vectors $Z_1, Z_2, \ldots$ having a common distribution $\Psi$ and the stopping rule $T$ depends on the primary parameter $\mu(\Psi)$, where $\mu$ is a functional of $\Psi$. This includes the preceding bivariate example as a special case with $Z_i = (X_i,Y_i)$ and $\mu(\Psi) = \mu(F)$, and it also allows the primary and secondary parameters to be functions of both $X_i$ and $Y_i$. The underlying distribution $\Psi$ can be nonparametric, parametric, or semiparametric. In the nonparametric case, the hybrid resampling approach can be implemented by using empirical likelihood or simpler variants thereof. The parametric approach uses parametric instead of empirical likelihood. Section 7.5 applies the semiparametric approach to Cox's regression models for censored survival data. The preceding discussion has focused on testing the hypothesis $\theta = \theta_0$ for a secondary endpoint following a group sequential trial. Confidence intervals for $\theta$ can be constructed by inverting these tests. We have assumed so far that the group sizes in the group sequential trial are prespecified constants. When the group sizes $n_j$ are random variables such that $n_j$ is unobservable if $n_j$ exceeds the stopping time of the sequential experiment, we can implement the hybrid resampling approach by using the ordering scheme in Sect. 7.3; details are given at the end of this section. Lai et al. (2009) have shown that for the hybrid resampling test of $H_0 : \theta \leq 0$,

$$\sup_{\mu,\theta \leq 0} P_{\mu,\theta}\{\text{Test rejects } H_0\} = \alpha + O(n^{-1}). \tag{7.19}$$

**Table 7.3** Type I error and power of tests of $H_0 : \theta = 0$ when (a) the primary endpoint $X$ and the secondary endpoint $Y$ are bivariate normal with mean $(\mu, \theta)$, $\mathrm{Var}(X) = \mathrm{Var}(Y) = 1$, and $\mathrm{Corr}(X,Y) = 0.8$; (b) the secondary endpoint $Y' = G^{-1}_{1+\theta}(\Phi(Y)) - 1$, with $(X,Y)$ being the same as in (a) except that $Y$ has zero mean. (BT, bootstrap test; BC, bias-corrected test of Liu et al.; P, pivot-based test of Whitehead et al.; EX, exact test; H, hybrid resampling test)

| | $\theta = 0$ | | | | | $\theta = 1.0/\sqrt{15}$ | | | | | $\theta = 1.5/\sqrt{15}$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\sqrt{15}\mu$ | BT | BC | P | EX | H | BT | BC | P | EX | H | BT | BC | P | EX | H |
| (a) Bivariate normal | | | | | | | | | | | | | | | |
| −2.0 | 4.2 | 10.1 | 6.1 | 2.7 | 3.0 | 44.3 | 50.1 | 48.6 | 35.3 | 35.5 | 62.2 | 65.2 | 61.7 | 54.2 | 52.9 |
| −1.5 | 5.2 | 8.8 | 5.5 | 2.9 | 3.2 | 48.8 | 58.2 | 57.5 | 42.2 | 45.9 | 69.2 | 74.6 | 72.5 | 65.4 | 65.2 |
| −1.0 | 4.9 | 7.1 | 4.0 | 2.8 | 3.5 | 58.9 | 69.1 | 70.3 | 53.0 | 57.5 | 79.9 | 82.1 | 80.6 | 75.2 | 76.1 |
| −0.5 | 5.3 | 5.7 | 4.2 | 3.4 | 3.1 | 70.8 | 75.1 | 73.7 | 62.3 | 61.4 | 88.3 | 90.9 | 89.8 | 86.7 | 84.6 |
| 0.0 | 4.8 | 4.9 | 4.6 | 3.5 | 3.7 | 69.5 | 73.7 | 74.1 | 63.6 | 63.4 | 94.8 | 95.3 | 96.0 | 93.7 | 92.1 |
| 0.5 | 7.4 | 4.5 | 6.6 | 4.9 | 4.8 | 66.5 | 70.6 | 68.4 | 64.6 | 64.5 | 95.2 | 96.0 | 97.1 | 92.4 | 93.7 |
| 1.0 | 6.9 | 5.1 | 5.7 | 5.1 | 4.9 | 66.3 | 58.6 | 62.2 | 61.1 | 60.7 | 93.2 | 93.4 | 92.2 | 92.7 | 91.9 |
| 1.5 | 6.5 | 5.9 | 5.3 | 4.6 | 5.2 | 57.1 | 38.1 | 56.5 | 49.4 | 48.0 | 90.6 | 82.7 | 88.6 | 88.3 | 89.5 |
| 2.0 | 5.8 | 6.8 | 4.5 | 4.3 | 4.8 | 44.8 | 30.8 | 44.2 | 34.5 | 35.1 | 78.0 | 62.9 | 75.9 | 73.7 | 73.5 |
| (b) Normal $X$ and exponential $Y'$ | | | | | | | | | | | | | | | |
| −2.0 | 4.4 | 8.5 | 6.3 | 1.3 | 4.0 | 39.1 | 41.8 | 40.5 | 20.0 | 32.7 | 55.3 | 53.7 | 56.8 | 38.6 | 48.9 |
| −1.5 | 4.4 | 7.0 | 5.5 | 2.0 | 3.4 | 45.4 | 49.8 | 46.3 | 31.1 | 42.3 | 61.4 | 63.6 | 60.5 | 45.6 | 60.2 |
| −1.0 | 4.5 | 4.2 | 4.1 | 1.7 | 3.2 | 54.6 | 59.1 | 57.2 | 42.9 | 49.5 | 74.5 | 74.1 | 72.9 | 66.4 | 72.1 |
| −0.5 | 4.7 | 3.4 | 3.9 | 1.7 | 3.7 | 63.4 | 61.3 | 62.5 | 42.7 | 56.6 | 84.1 | 83.8 | 83.6 | 66.7 | 82.3 |
| 0.0 | 4.0 | 2.3 | 4.8 | 2.2 | 3.9 | 63.5 | 56.5 | 60.8 | 42.8 | 60.0 | 87.2 | 82.0 | 82.5 | 75.5 | 83.2 |
| 0.5 | 6.0 | 2.8 | 5.7 | 2.3 | 4.4 | 59.9 | 50.4 | 55.3 | 41.1 | 54.2 | 86.9 | 80.9 | 84.2 | 75.6 | 84.0 |
| 1.0 | 6.4 | 2.7 | 6.0 | 2.1 | 4.6 | 58.4 | 30.8 | 50.2 | 37.0 | 54.8 | 86.2 | 70.1 | 83.8 | 73.8 | 82.1 |
| 1.5 | 7.5 | 3.9 | 5.5 | 2.0 | 5.0 | 49.8 | 16.8 | 45.1 | 22.7 | 43.9 | 76.3 | 41.3 | 71.7 | 54.4 | 72.4 |
| 2.0 | 5.6 | 3.8 | 5.1 | 1.9 | 5.1 | 41.3 | 14.0 | 37.9 | 14.3 | 31.7 | 68.5 | 26.4 | 62.3 | 33.9 | 58.2 |

*Example 7.2.* Let $(X_i, Y_i)$ be the same as in Example 7.1, and let

$$Y_i' = G^{-1}_{1+\theta}(\Phi(Y_i)) - 1,$$

where $\Phi$ is the standard normal distribution function and $G_\lambda(u) = 1 - e^{-u/\lambda}$ is the distribution function of the exponential distribution with mean $\lambda$. Suppose the observations are actually $(X_i, Y_i')$. Note that $E(Y_i') = \theta$ but $\mathrm{Corr}(X_i, Y_i') \neq \mathrm{Corr}(X_i, Y_i)$. As pointed out in Sect. 7.1.6, to handle the possibility of non-normality of the bivariate outcomes, Liu et al. (2000) and Whitehead et al. (2000) appeal to the group sequential nature of the stopping rule so that the sample sums within each group can be regarded as approximately bivariate normal, thereby justifying the normal approximation to their test statistics but with the sample correlation coefficient of $(X_i, Y_i')$ and the sample variance $\hat{\sigma}_T^2$ of the $Y_i'$ replacing the corresponding population quantities. This approximate (bivariate) normality within each group can also be used to justify the use of the exact test that assumes bivariate normality. Without making such an assumption or approximation, the hybrid resampling approach draws $B$ resamples from the empirical distribution of $(X_i - \bar{X}_T + \mu, (Y_i' - \bar{Y}_T')/\hat{\sigma}_T)$, $1 \leq i \leq T$, whereas the bootstrap test draws $B$ bootstrap samples from the empirical distribution of $(X_i, (Y_i' - \bar{Y}_T')/\hat{\sigma}_T)$, with $B = 2000$ for the results in Table 7.3. Table 7.3b gives the power functions of the tests for this non-normal setting. It shows that the hybrid resampling approach maintains the type I error probability while the exact test, which is exact only under the normal distribution assumption for $Y'$, is conservative and other tests have inflated type I error probability.

The ordering scheme (7.17) can be applied to the case in which the interim sample sizes $n_j$ of the group sequential test concerning the primary parameter $\mu = \mu(\Psi)$ are random variables that are observable only up to the stopping time of the trial. As in the second paragraph of Sect. 7.4.2, the observations are $Z_1, \dots, Z_T$ and $\hat{\Psi}_T$ is the empirical distribution of $Z_1, \dots, Z_T$. Let $W_T = \theta(\hat{\Psi}_T)$. To test $H_0 : \theta(\Psi) = \theta_0$ for a secondary parameter at the end of the group sequential trial, we can use the ordering (7.17) to compute the $p$-value

$$\sup_{\mu} \hat{P}_{\mu,\theta_0} \{(T, W_T) \geq (\tau, w_\tau)\},$$

where $(\tau, w_\tau)$ denotes the observed value of $(T, W_T)$ and $\hat{P}_{\mu,\theta_0}$ refers to the probability measure under which $Z_i$ are generated from the nonparametric maximum likelihood estimator of $\Psi$ subject to the constraints $\mu(\Psi) = \mu$ and $\theta(\Psi) = \theta_0$. In particular, we can apply this ordering, with $W_T = \sqrt{T}\bar{Y}_T$, in evaluating the probability (7.8) when the interim sample sizes $n_j$ are random variables instead of being fixed in advance. In this case, $W_{T\wedge t} \geq w_{T\wedge t}$ is equivalent to $\bar{Y}_{T\wedge t} \geq \bar{y}_{T\wedge t}$. In Example 7.2, the $n_j$ are fixed in advance and therefore the distribution of $\sqrt{T}\bar{Y}_T$ under $E(X) = \mu$ and $E(Y) = 0$ can be evaluated by simulation or numerical integration. On the other hand, if $n_j$ are random and one does not know the probability mechanism generating them except that they are due to the accrual pattern which is independent of $(X_i, Y_i)$, then one can only condition on the observed $n_j$ in evaluating the probability (7.8).

## 7.5 Applications to Time-Sequential Survival Trials

Lai and Li (2006) apply the ordering scheme (7.17) to construct confidence intervals following time-sequential tests in Cox regression (proportional hazards model). The time-sequential trial described in Sect. 6.5 has time to failure as the primary endpoint, and involves interim analyses of the trial at calendar times $t_j$ $(1 \leq j \leq k)$, with $0 < t_1 < \cdots < t_k = t^*$, where $t^*$ is the prescribed duration of the trial. Suppose $n$ patients enter the trial serially. The data at calendar time $t$ consist of $(Y_i(t), \delta_i(t), z_i I_{\{T_i \leq t\}})$, for $i = 1, \dots, n$, where $Y_i(t) = \min\{Y_i, \xi_i, (t - T_i)^+\}$, $\delta_i(t) = I_{\{Y_i(t) = Y_i\}}$, $T_i \geq 0$ denotes the entry time and $Y_i > 0$ the time to failure after entry of the $i$th subject, and $z_i$ is the subject's covariate while $\xi_i$ is the withdrawal time, possibly infinite. (In Sect. 6.5 we have not included the covariates $z_i$.)

Assume that $T_i$ is independent of $(Y_i, \xi_i, z_i)$, that $Y_i$ and $\xi_i$ are conditionally independent given $z_i$, and that the hazard function of $Y_i$ is given by Cox's (1972) proportional hazards model, for which

$$P\{y \le Y_i \le y + dy \,|\, Y_i \ge y, z_i\} = e^{\beta z_i}\, d\Lambda(y), \tag{7.20}$$

where $\beta$ is an unknown parameter and $\Lambda$ is the baseline cumulative hazard function that is assumed to be continuous. To test the null hypothesis $H_0 : \beta = 0$, which corresponds to no covariate effect on survival, differentiation of the log partial likelihood for $\beta$ at $\beta = 0$ and calendar time $t$ yields Cox's score statistic:

$$S_n(t) = \sum_{i=1}^{n} \delta_i(t) \left\{ z_i - \left( \sum_{j \in R_i(t)} z_j \right) \Big/ |R_i(t)| \right\}, \tag{7.21}$$

where $R_i(t) = \{j : Y_j(t) \ge Y_i(t)\}$ and $|R_i(t)|$ denotes the size of the "risk set" $R_i(t)$. The observed Fisher information at calendar time $t$ is

$$V_n(t) = \sum_{i=1}^{n} \delta_i(t) \left[ \sum_{j \in R_i(t)} z_j^2 \Big/ |R_i(t)| - \left\{ \sum_{j \in R_i(t)} z_j \Big/ |R_i(t)| \right\}^2 \right], \tag{7.22}$$

which provides an estimate of the null variance of $S_n(t)$; see Sect. 6.1.2 where it is shown that in analogy with (7.5) for the case of normal random walks, one can use a repeated significance test that rejects $H_0$ at the $j$th interim analysis $(1 \le j \le k)$ if

$$S_n(t_j)/V_n^{1/2}(t_j) \ge b_j \quad \text{or} \quad S_n(t_j)/V_n^{1/2}(t_j) \le a_j, \tag{7.23}$$

and stops the trial as soon as (7.23) occurs, where $a_j < 0 < b_j$.

### 7.5.1 Implementation of the Hybrid Resampling Approach

Lai and Li (2006) consider interval estimation of $\beta$ following the time-sequential test (7.23). Let $\tau$ denote the calendar time of stopping, that is, $\tau = \min\{t_j : (7.23)$ holds at time $t_j\}$. For notational simplicity, denote $S_n(t)$ by $S(t)$ and $V_n(t)$ by $V(t)$. Letting $\Psi_t = S(t)/V(t)$, they use the ordering scheme (7.17), which orders the sample space of $(\tau, \Psi_\tau)$ by

$$(\tau_1, \Psi_{\tau_1}^{(1)}) \le (\tau_2, \Psi_{\tau_2}^{(2)}) \quad \text{if and only if} \quad \Psi_{\tau_1 \wedge \tau_2}^{(1)} \le \Psi_{\tau_1 \wedge \tau_2}^{(2)}. \tag{7.24}$$

Similarly to the normal-mean case, let $p(\beta) = P_\beta\{(\tau, \Psi_\tau) > (\tau, \Psi_\tau)_{\text{obs}}\}$, where $(\tau, \Psi_\tau)_{\text{obs}}$ denotes the observed value of $(\tau, \Psi_\tau)$. Then $\{\beta : \alpha < p(\beta) < 1 - \alpha\}$ is a confidence set for $\beta$ with coverage probability $1 - 2\alpha$, and an important ingredient of the hybrid resampling approach is to replace the unknown nuisance parameters in $p(\beta)$ by suitably chosen estimators. To begin with, suppose that $\xi_i$ is independent of $z_i$ and has distribution function $C$. The baseline distribution $G = 1 - e^{-\Lambda}$ in $p(\beta)$

can be estimated by $\hat{G} = 1 - e^{-\hat{\Lambda}}$, where $\hat{\Lambda}$ is Breslow's estimator of the cumulative hazard function from all the data at the end of the trial:

$$\hat{\Lambda}(s) = \sum_{i:Y_i(\tau)\le s} \left\{ \delta_i(\tau) \Big/ \left( \sum_{j\in R_i(\tau)} e^{\hat{\beta} z_j} \right) \right\}, \tag{7.25}$$

in which $\hat{\beta}$ is Cox's (1972) estimate of $\beta$ that maximizes the partial likelihood at time $\tau$; see Sect. 6.1.2. Since the $\xi_i$ are censored by $\min\{Y_i, (\tau - T_i)^+\}$, $C$ can be estimated by the Kaplan–Meier estimator $\hat{C}$. This suggests replacing $p(\beta)$ by

$$\hat{p}(\beta) = P\left\{ \left( \tau^{(\beta)}, \Psi^{(\beta)}_{\tau^{(\beta)}} \right) > (\tau, \Psi_\tau)_{\text{obs}} \right\}, \tag{7.26}$$

where the superscript $(\beta)$ means that the observations are generated by the proportional hazards model with baseline distribution $\hat{G}$ and regression parameter $\beta$. Usually $\hat{p}(\beta)$ is monotone in $\beta$, so the confidence set $\{\beta : \alpha < \hat{p}(\beta) < 1-\alpha\}$ with approximate coverage probability $1-2\alpha$ can be expressed as an interval, whose endpoints $\beta_L < \beta_U$ are defined by $\hat{p}(\beta_L) = \alpha$, $\hat{p}(\beta_U) = 1-\alpha$, and can be computed by using the procedure in the last paragraph of Sect. 7.2.1.

Concerning the Monte Carlo evaluation of $\hat{p}(\beta)$, note that the observed entry times $T_i$ and covariates $z_i$ are taken as fixed constants in $\hat{p}(\beta)$, for which we need only generate the survival times $Y_i^*$ and censoring times $\xi_i^*$. Since $\hat{G}$ (or $\hat{C}$) can only be estimated up to the longest observed survival, or censoring, time, denoted by $t'$, or $t''$, we can only generate $Y_i^* \wedge t'$ and $\xi_i^* \wedge t''$. This suffices, however, for the time-sequential score statistic (7.21) and its estimated null variance (7.22) for $t \le \tau$. To generate $Y_i^* \wedge t'$, note that it has the same distribution as $(1-\hat{G})^{-1}[U^{\exp(-\beta z_i)} \vee \{1 - \hat{G}(t')\}]$, where $U \sim$ Uniform $[0,1]$.

We have assumed in the preceding that $\xi_i$ is independent of $(z_i, Y_i)$. For dichotomous covariates, we can easily extend the methodology to the case where the control group $(z_i = 0)$ and the treatment group $(z_i = 1)$ have different censoring distributions $C_0$ and $C_1$, by using separate Kaplan–Meier estimators $\hat{C}_0$ and $\hat{C}_1$. Clearly the same idea can be used for discrete covariates that have a finite number of possible values. For continuous covariates, usually the rate of loss to follow-up, that is, censoring by the $\xi_i$, is small relative to that of administrative censoring by $(\tau - T_i)^+$. In this case, one can simply use the same Kaplan–Meier estimate $\hat{C}$ as that used under the additional assumption of independence between $\xi_i$ and $z_i$. Alternatively we can treat $\xi_i$ as ancillary and impute the censored $\xi_i$ by using the methods of Lai and Li (2006, p. 644).

### 7.5.2 Tests on Secondary Parameters in the Cox Model

Consider the time-sequential clinical trial with failure-time primary endpoint. Assuming Cox's proportional hazards model

$$P\{y \le Y_i \le y + dy \,|\, Y_i \ge y, x_i, \boldsymbol{u}_i\} = e^{\beta x_i + \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{u}_i} \, d\Lambda(y), \tag{7.27}$$

in which $\Lambda$ is the baseline cumulative hazard function that is assumed to be continuous, $x_i$ is the primary covariate, for example, treatment, and $\boldsymbol{u}_i$ is a vector of concomitant covariates. In addition, $T_i$ is assumed to be independent of $(Y_i, \xi_i, x_i, \boldsymbol{u}_i^{\mathrm{T}})$, and $Y_i$ and $\xi_i$ are assumed to be conditionally independent of $(x_i, \boldsymbol{u}_i^{\mathrm{T}})$. The log partial likelihood at calendar time $t$ is

$$l_t(\beta, \boldsymbol{\theta}) = \sum_{i=1}^{n} \delta_i(t) \left[ (\beta x_i + \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{u}_i) - \log \left( \sum_{j \in R_i(t)} e^{\beta x_j + \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{u}_j} \right) \right], \tag{7.28}$$

where $R_i(t) = \{j : Y_j(t) \ge Y_i(t)\}$. To test the primary hypothesis $\beta = \beta_0$, one can proceed as follows. Let $(\hat{\beta}_t, \hat{\boldsymbol{\theta}}_t)$ be the maximizer of $l_t(\beta, \boldsymbol{\theta})$, and let $V_t$ be the first diagonal element of $(-\ddot{l}_t(\hat{\beta}_t, \hat{\boldsymbol{\theta}}_t))^{-1}$, which is an estimate of the asymptotic variance of $\hat{\beta}_t$. At the $j$th interim analysis $(j = 1, \ldots, k)$, the trial is terminated if

$$\left( \hat{\beta}_{t_j} - \beta_0 \right) \Big/ V_{t_j}^{1/2} \ge b_j \quad \text{or} \quad \left( \hat{\beta}_{t_j} - \beta_0 \right) \Big/ V_{t_j}^{1/2} \le a_j, \tag{7.29}$$

rejecting the primary hypothesis $\beta = \beta_0$ upon stopping, where $a_j < 0 < b_j$ are the stopping boundaries of the repeated significance test.

A commonly used alternative to the preceding Wald statistic is Cox's score statistic:

$$S_t(\beta, \boldsymbol{\theta}) = \frac{\partial l_t}{\partial \beta} = \sum_{i=1}^{n} \delta_i(t) \left[ x_i - \frac{X_{it}^{(1)}(\beta, \boldsymbol{\theta})}{r_{it}(\beta, \boldsymbol{\theta})} \right], \tag{7.30}$$

as in Sect. 6.1.2, where

$$r_{it}(\beta, \boldsymbol{\theta}) = \sum_{j \in R_i(t)} e^{\beta x_j + \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{u}_j}, \qquad X_{it}^{(1)}(\beta, \boldsymbol{\theta}) = \sum_{j \in R_i(t)} x_j e^{\beta x_j + \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{u}_j}. \tag{7.31}$$

At the $j$th interim analysis $(j = 1, \ldots, k)$, the trial is terminated if

$$S_{t_j}\left(\beta_0, \tilde{\boldsymbol{\theta}}_{t_j}\right) / v_{t_j}^{1/2} \ge b_j \quad \text{or} \quad S_{t_j}\left(\beta_0, \tilde{\boldsymbol{\theta}}_{t_j}\right) / v_{t_j}^{1/2} \le a_j, \tag{7.32}$$

rejecting the primary hypothesis $\beta = \beta_0$ upon stopping, where $a_j < 0 < b_j$ are the stopping boundaries of the repeated significance test and $\tilde{\boldsymbol{\theta}}_t$ is the maximizer of $l_t(\beta_0, \boldsymbol{\theta})$. The estimate $v_t$ of the null variance of $S_t(\beta_0, \tilde{\boldsymbol{\theta}}_t)$ is $I_{11} - I_{12} I_{22}^{-1} I_{21}$, where $I_{ij}$ is the $(i, j)$th component of $-\ddot{l}_t(\beta_0, \tilde{\boldsymbol{\theta}}_t)$, as in usual likelihood inference, treating partial likelihood like usual likelihood. Let

$$\boldsymbol{U}_{it}^{(1)}(\beta, \boldsymbol{\theta}) = \sum_{j \in R_i(t)} \boldsymbol{u}_j e^{\beta x_j + \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{u}_j}$$

$$
\begin{aligned}
U_{it}^{(2)}(\beta,\boldsymbol{\theta}) &= \sum_{j\in R_i(t)} e^{\beta x_j+\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{u}_j}\,\boldsymbol{u}_j\boldsymbol{u}_j^{\mathrm{T}}\\
X_{it}^{(2)}(\beta,\boldsymbol{\theta}) &= \sum_{j\in R_i(t)} x_j^2 e^{\beta x_j+\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{u}_j}. \qquad (7.33)
\end{aligned}
$$

The components of the Hessian matrix $\ddot{l}_t(\beta,\boldsymbol{\theta})$ are given by

$$
\begin{aligned}
\frac{\partial^2 l_t}{\partial\beta^2} &= -\sum_{i=1}^{n}\delta_i(t)\left[\frac{X_{it}^{(2)}(\beta,\boldsymbol{\theta})}{r_{it}(\beta,\boldsymbol{\theta})}-\left(\frac{X_{it}^{(1)}(\beta,\boldsymbol{\theta})}{r_{it}(\beta,\boldsymbol{\theta})}\right)^2\right],\\
\frac{\partial^2 l_t}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}} &= -\sum_{i=1}^{n}\delta_i(t)\left[\frac{U_{it}^{(2)}(\beta,\boldsymbol{\theta})}{r_{it}(\beta,\boldsymbol{\theta})}-\left(\frac{U_{it}^{(1)}(\beta,\boldsymbol{\theta})}{r_{it}(\beta,\boldsymbol{\theta})}\right)\left(\frac{U_{it}^{(1)}(\beta,\boldsymbol{\theta})}{r_{it}(\beta,\boldsymbol{\theta})}\right)^{\mathrm{T}}\right],\\
\frac{\partial^2 l_t}{\partial\beta\partial\boldsymbol{\theta}^{\mathrm{T}}} &= -\sum_{i=1}^{n}\delta_i(t)\left[\frac{\sum_{j\in R_i(t)}x_j\boldsymbol{u}_j e^{\beta x_j+\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{u}_j}}{r_{it}(\beta,\boldsymbol{\theta})}-\left(\frac{X_{it}^{(1)}(\beta,\boldsymbol{\theta})}{r_{it}(\beta,\boldsymbol{\theta})}\right)\left(\frac{U_{it}^{(1)}(\beta,\boldsymbol{\theta})}{r_{it}(\beta,\boldsymbol{\theta})}\right)\right],
\end{aligned}
$$
(7.34)

and $\partial^2 l/\partial\beta\partial\boldsymbol{\theta}^{\mathrm{T}} = (\partial^2 l/\partial\beta\partial\boldsymbol{\theta})^{\mathrm{T}}$. Making use of (7.29)–(7.34) and the ordering scheme (7.17), we now proceed to give hybrid resampling tests and confidence intervals for the secondary parameters in the proportional hazards model (7.27) following a time-sequential trial with stopping rule (7.29) or (7.32).

In particular, consider the bivariate Cox regression model (7.27) in which $\theta$ is univariate. The Wald test statistic of $H_0:\theta=\theta_0$ is $W_T=(\hat{\theta}_T-\theta_0)/\tilde{V}_T^{1/2}$, where $T$ is the time at which the trial is terminated and the estimate $\tilde{V}_T$ of the null variance of $\hat{\theta}_T$ is the second diagonal element of $(-\ddot{l}_T(\hat{\beta}_T,\hat{\theta}_T))^{-1}$. In analogy with (7.8), the $p$-value, evaluated by hybrid resampling, of the one-sided test of $H_0$ is

$$
\sup_{\beta}\hat{P}_{\beta,\theta_0}\{(T,W_T)\geq(\tau,w_\tau)\}, \qquad (7.35)
$$

in which $(\tau,w_\tau)$ is the observed value of $(T,W_T)$. The ordering scheme (7.17) is used in the inequality in (7.35), in which $\hat{P}_{\beta,\theta_0}$ replaces the baseline distribution $G=1-e^{-\Lambda}$ in $P_{\beta,\theta_0}$ by $\hat{G}_T=1-e^{-\hat{\Lambda}_T}$, where $\hat{\Lambda}_t$ is Breslow's estimator of the cumulative hazard function from all the data up to time $t$ and uses the Kaplan–Meier estimator or variants thereof to estimate the censoring mechanism in $P_{\beta,\theta_0}$.

## 7.6 Simulation Studies

### 7.6.1 Confidence Intervals for Primary Endpoint of Hazard Rate

A standard approach in the literature on time-sequential survival analysis is to use the space–time Brownian motion approximation of $(S(t), V(t))$, to which Siegmund's ordering can be applied since the stopping rule (7.23) has the form (7.5) under this approximation; see Whitehead (1992, Chap. 5). However, as will be shown in this section, using the Brownian motion approximation to treat the problem as that of a normal mean following a group sequential test may not provide an adequate approximation to the coverage probability of the confidence interval based on Siegmund's ordering unless $\beta$ is very close to 0. Instead of applying this Brownian motion approximation directly, an alternative is to apply hybrid resampling to an extension of Siegmund's ordering to the present setting. Recalling the form (7.23) of the stopping rule with $a_j < 0 < b_j$, Lai and Li (2006) discuss the following extensions of Siegmund's ordering to the sample space of $(\tau, S(\tau), V(\tau))$:

*Extension 1.* We have $(\tau, s, v) > (\tilde{\tau}, \tilde{s}, \tilde{v})$ whenever

1. $\tau = \tilde{\tau}$ and $s/\sqrt{v} > \tilde{s}/\sqrt{\tilde{v}}$, or
2. $\tau < \tilde{\tau}$ and $s > 0$, or
3. $\tau > \tilde{\tau}$ and $\tilde{s} < 0$,

in which $\tau$ and $\tilde{\tau}$ take values in $\{t_1, \ldots, t_k\}$. We have used $s/\sqrt{v}$ in (i) because the stopping rule (7.23) involves $S(t_j)/V^{1/2}(t_j)$.

*Extension 2.* Alternatively, by analogy with the Emerson–Fleming mean ordering, we can use $s/v$ in lieu of $s/\sqrt{v}$, so that $(\tau, s, v) > (\tilde{\tau}, \tilde{s}, \tilde{v})$ whenever

1. $\tau = \tilde{\tau}$ and $s/v > \tilde{s}/\tilde{v}$, or
2. $\tau < \tilde{\tau}$ and $s > 0$, or
3. $\tau > \tilde{\tau}$ and $\tilde{s} < 0$.

These extensions of Siegmund's ordering can be used in conjunction with the hybrid resampling method to construct confidence intervals for $\beta$. However, the results below show that the confidence intervals thus constructed do not improve the coverage probabilities of those constructed via Brownian motion approximations. This illustrates the importance of using hybrid resampling in conjunction with a suitably chosen ordering scheme.

Lai and Li (2006) have carried out a simulation study comparing different methods for constructing confidence intervals for $\beta$ in the proportional hazards model (7.20) with dichotomous covariates $z_j$, for which $\theta = e^{\beta}$ corresponds to the hazard ratio of a new treatment ($z_j = 1$) relative to a control ($z_j = 0$). Since the patients are randomized to either treatment with probability $\frac{1}{2}$, the score statistic (7.21) is the log-rank statistic for testing $\beta = 0$, and the null variance estimate (7.22) can be replaced by the more convenient approximation (6.37). The simulation study considers a time-sequential trial in which $n = 350$ subjects enter the trial uniformly during a 3-year recruitment period. The trial is designed to last for a maximum of $t^* = 5.5$ years, with interim analyses after 1 year and every 6 months thereafter. The log-rank statistic is used to test $H_0 : \beta = 0$ at each data monitoring time $t_j$ $(j = 1, \ldots, 10)$, and the test is stopped at the smallest $t_j$ such that

$$V_n(t_j) \geq 55, \tag{7.36a}$$

or

$$V_n(t_j) \geq 11, \qquad |S_n(t_j)| \Big/ V_n^{1/2}(t_j) \geq 2.85, \tag{7.36b}$$

or at $t_{10}(= t^*)$ when (7.36) does not occur, where $V_n(t)$ is defined by (6.37). If the test stops with $V_n(t_j) \geq 55$ or at $t^*$, reject $H_0$ if $|S_n(t^*)|/V_n^{1/2}(t^*) \geq 2.05$. Also reject $H_0$ if (7.36b) occurs for some $j < 10$. The threshold 2.05 for the final analysis at $t^*$ is chosen so that the type I error probability of the test is approximately 5% using the Brownian motion approximation. It is assumed that there is no loss to follow-up, that the lifetimes of the control group have an exponential distribution with mean 3 years, and that those of the treatment group have an exponential distribution with mean $3e^{-\beta}$ years, with $e^{\beta} = 1, 2/3, 1/2$. Table 7.4 gives the coverage errors, with nominal value $\alpha = 0.05$, of the upper and lower confidence limits for $\beta$, and the coverage probabilities, with nominal value 90%, of two-sided confidence intervals. Besides the hybrid resampling method with the ordering scheme (7.24), Table 7.4 also considers the following six methods for constructing confidence limits for $\beta$: Siegmund's ordering and the Emerson–Fleming ordering, both applied to the space–time Brownian motion approximation of $(S_n(t), V_n(t))$ and thereby yielding score-based confidence intervals; the Emerson–Fleming ordering applied to $\hat{\beta}_t$ in place of the normal mean, with $V_n(t_j)$ playing the role of $n_j$, yielding a Wald-type confidence interval; Extension 1 of the Siegmund's ordering and its variant Extension 2 used instead of (7.24) for hybrid resampling; and the naive confidence interval $\hat{\beta}_\tau \pm 1.645/V_n^{1/2}(\tau)$, noting that $V_n^{-1/2}(t)$ is the asymptotic standard error of $\hat{\beta}_t$; see (6.14). For the Emerson–Fleming method, if the trial stops at time $t_j < t_{10}$ and $V_n(t_j) < 55$, then the ordering applied to $\hat{\beta}_t$, or to $S_n(t)/V_n(t)$, entails consideration of the event that $\hat{\beta}_{\tau^*}$ exceeds the observed value of $\hat{\beta}_\tau$, or that $S_n(\tau^*)/V_n(\tau^*)$ exceeds the observed $S_n(\tau)/V_n(\tau)$, where $\tau^*$ is the calendar time at which $V_n(\tau^*) = 55$. Recall that $4V_n(t)$ is the total number of deaths up to time $t$; see (6.37).

Concerning the coverage error of the lower limit $\beta_L$ of the hybrid resampling confidence interval in Table 7.4, monotonicity of $\hat{p}(\beta)$ defined in (7.26) implies that

$$P_\beta\{\beta < \beta_L\} = P_\beta\left\{\hat{p}(\beta) < \hat{p}(\beta_L)\right\} = P_\beta\left\{\hat{p}(\beta) < \alpha\right\}, \tag{7.37}$$

since $\hat{p}(\beta_L) = \alpha$. To compute (7.37) by Monte Carlo, this suggests that we need only evaluate $\hat{p}(\beta)$ for each simulated dataset and check if it is less than $\alpha$, without solving for $\beta_L$ and thereby greatly reducing the computation time in repeated

**Table 7.4** Coverage errors in percentages for lower (L) and upper (U) confidence limits and coverage probabilities ($P$) of confidence intervals for $\beta$ when interim analyses are performed at fixed calendar times. (H, hybrid resampling method based on the ordering (7.24); S, Siegmund's method; $\mathrm{EF_S}$, Emerson–Fleming score-based method; $\mathrm{EF_W}$, Emerson–Fleming Wald-type method; $\mathrm{H_S}$, hybrid resampling method based on Extension 1 of the Siegmund ordering; $\mathrm{H_{S'}}$, hybrid resampling method based on Extension 2 of the Siegmund ordering; N, naive normal method)

| | $\beta = 0$ | | | $\beta = \log(2/3)$ | | | $\beta = \log(1/2)$ | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | L | U | P | L | U | P | L | U | P |
| H | 4.45 | 4.55 | 91.00 | 5.25 | 5.35 | 89.40 | 5.05 | 4.05 | 90.90 |
| S | 4.45 | 5.05 | 90.50 | 4.65 | 0.35 | 95.00 | 5.75 | 0.00 | 94.25 |
| $\mathrm{EF_S}$ | 4.45 | 5.05 | 90.50 | 4.65 | 0.65 | 94.70 | 7.40 | 0.00 | 92.60 |
| $\mathrm{EF_W}$ | 4.35 | 4.80 | 90.85 | 4.70 | 0.60 | 94.70 | 5.35 | 0.00 | 94.65 |
| $\mathrm{H_S}$ | 1.35 | 6.35 | 92.30 | 3.15 | 6.30 | 90.55 | 2.15 | 3.70 | 94.15 |
| $\mathrm{H_{S'}}$ | 1.10 | 7.20 | 91.70 | 5.50 | 4.10 | 90.40 | 3.75 | 4.15 | 92.10 |
| N | 4.15 | 5.05 | 90.80 | 5.80 | 7.75 | 86.45 | 3.75 | 3.15 | 93.10 |

simulation runs. The same idea is also used in Table 7.4 for the upper confidence limit $\beta_U$ of the hybrid resampling interval and for Siegmund's and Emerson and Fleming's confidence limits that are based on test inversion. Each result in Table 7.4 is based on 2000 simulations.

Table 7.4 shows that the hybrid resampling method using the ordering (7.24) yields quite accurate confidence intervals, with all probabilities within 1% of their nominal values. In contrast, using Extension 1 or 2 of the Siegmund ordering in conjunction with the hybrid resampling method produces upper and lower confidence bounds whose coverage errors differ substantially from 0.05 in Table 7.4. Although the other methods also perform well at $\beta = 0$, they have obvious difficulties with the upper confidence bound when $\beta < 0$. Therefore, using the Brownian motion approximation to treat the problem as that of a normal mean following a group sequential test does not seem to provide an adequate approximation unless $\beta$ is very close to 0. The Brownian motion approximation induces an ordering scheme that depends on $(S(\tau), V(\tau))$ but ignores the sampling fluctuations in $\tau$. In this connection, Lai and Li (2006) also point out the following difference between the ordering (7.24) and Extension 2 of Siegmund's ordering. Let $\tilde{\tau}$ denote the observed value of $\tau$, let $(\tilde{s}, \tilde{v})$ denote $(S(\tau), V(\tau))_{\text{obs}}$, and let $(\tilde{s}_t, \tilde{v}_t)$ denote $(S(t), V(t))_{\text{obs}}$ at $t = t_j \le \tilde{\tau}$. Under the ordering (7.24), the event $\{(\tau, \Psi_\tau) > (\tau, \Psi_\tau)_{\text{obs}}\}$ that is associated with $p(\beta)$ can be expressed as

$$\begin{aligned}\{(\tau, \Psi_\tau) > (\tau, \Psi_\tau)_{\text{obs}}\} = &\left\{\tau > \tilde{\tau} \text{ and } S(\tilde{\tau})\big/V(\tilde{\tau}) > \tilde{s}\big/\tilde{v}\right\} \\ &\cup \left\{\tau \le \tilde{\tau} \text{ and } S(\tau)/V(\tau) > \tilde{s}_\tau\big/\tilde{v}_\tau\right\}. \qquad (7.38)\end{aligned}$$

On the other hand, under the ordering of Extension 2, $p(\beta)$ is associated with the event

$$\{\tau > \tilde{\tau} \text{ and } \tilde{s} < 0\} \cup \{\tau < \tilde{\tau} \text{ and } S(\tau) > 0\} \cup \left\{\tau = \tilde{\tau} \text{ and } S(\tau)/V(\tau) > \tilde{s}\big/\tilde{v}\right\}.$$

The inherent incompatibility between calendar and information time scales is addressed in the ordering (7.24) by designating values $(\tau, V(\tau), S(\tau))$ in the sample space to be more extreme than those in the observed sample if stopping occurs after $\tilde{\tau}$ and $S(\tilde{\tau})/V(\tilde{\tau})$ exceeds $\tilde{s}/\tilde{v}$ or if stopping occurs at or before $\tilde{\tau}$ and $S(\tau)/V(\tau)$ exceeds $\tilde{s}_\tau/\tilde{v}_\tau$; see (7.38). Note that the observed sample consists of not only $\tilde{\tau}$, $\tilde{s}$, and $\tilde{v}$ but also $(\tilde{s}_t, \tilde{v}_t)$ for $t = t_j < \tilde{\tau}$. In contrast, Extension 1 or 2 of Siegmund's ordering only involves the information time $V(\tau)$ when $\tau = \tilde{\tau}$, and Table 7.4 shows that it falls short of incorporating the full extent of the randomness of the information time upon stopping.

### 7.6.2 *Test for Secondary Endpoint in Cox Model*

Lai et al. (2009) consider a time-sequential trial in which $n = 350$ subjects enter the trial uniformly during a 3-year recruitment period and are randomized to treatment, $x = 1$, or control, $x = 0$, with probability $1/2$. A baseline covariate $u$ is also measured for each subject upon entry. Assume the survival time follows the proportional hazards model (7.27). Like that in Sect. 7.6.1, the trial is designed to last for a maximum of $t^* = 5.5$ years, with interim analyses after 1 year and every 6 months thereafter. The Wald statistic is used to test $\beta = 0$ at each interim monitoring time $t_j (j = 1, \ldots, 10)$. Using the notation of Sect. 7.5.2, the trial is stopped at the smallest $t_j$ such that

$$V_{t_j}^{-1} \geq 55 \tag{7.39}$$

or

$$V_{t_j}^{-1} \geq 11 \quad \text{and} \quad |\hat{\beta}_{t_j}|/V_{t_j}^{1/2} \geq 2.85 \tag{7.40}$$

or at $t_{10} = t^*$ when (7.39) and (7.40) do not occur. If the trial stops with (7.39) or at $t_j = t^*$, reject $\beta = 0$ if $|\hat{\beta}_{t_j}|/V_{t_j}^{1/2} \geq 2.05$. Also reject $\beta = 0$ if (7.40) occurs for some $j < 10$, similar to Sect. 7.6.1. Lai et al. (2009) compare the type I error probability and power of the proposed hybrid resampling method for testing $H_0 : \theta = 0$ with those of the naive Wald test that ignores early stopping, an extension of the bias-correction method of Liu et al. (2000), and the bootstrap test; $B = 2000$ resamples are used in the hybrid resampling and the bootstrap approaches. The results are given in Table 7.5, in which each entry is based on 2000 simulations, with the same 2000 simulated datasets for each method. The table shows that the hybrid resampling test has type I error probabilities estimated to be at most 0.052 at all values of the primary parameter $\beta$, whereas the type I error probabilities of the other methods exceed 0.06 at some $\beta$ values.

For large $n$, the maximum partial likelihood estimates $(\hat{\beta}_t, \hat{\theta}_t)$, after multiplication by $-\ddot{l}_t(\hat{\beta}_t, \hat{\theta}_t)$, behave like a bivariate zero-mean Gaussian process with independent increments under $(\beta, \theta) = (0, 0)$. For approximate bias-corrected inference on a secondary endpoint, Liu et al. (2000) assume weak convergence of the test statistics, after suitable normalization, to a bivariate Wiener process,

**Table 7.5** Type I error probability and power of different tests of $\theta = 0$. (N, naive normal method; B, bias-correction method; BOOT, bootstrap method; H, hybrid resampling method)

| Method | $e^{\beta}$=0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 | 1.2 | 1.5 | 2.0 |
|---|---|---|---|---|---|---|---|---|---|
| (a) $e^{\theta} = 1.0$ $(\theta = 0)$ | | | | | | | | | |
| N | 4.6 | 4.4 | 4.6 | 6.7 | 7.0 | 5.3 | 5.3 | 5.8 | 4.4 |
| B | 5.4 | 5.2 | 4.3 | 5.2 | 5.5 | 4.5 | 6.0 | 6.6 | 7.2 |
| BOOT | 5.7 | 4.8 | 4.5 | 6.4 | 6.1 | 5.9 | 5.2 | 5.7 | 5.3 |
| H | 4.1 | 3.7 | 3.5 | 5.2 | 4.7 | 3.8 | 4.0 | 3.8 | 4.0 |
| (b) $e^{\theta} = 1.5$ | | | | | | | | | |
| N | 32.1 | 42.3 | 57.5 | 60.1 | 61.5 | 62.2 | 65.6 | 43.4 | 30.8 |
| B | 30.6 | 41.9 | 59.0 | 57.8 | 59.8 | 60.1 | 68.1 | 47.9 | 33.3 |
| BOOT | 33.1 | 44.8 | 59.7 | 58.2 | 62.3 | 61.8 | 63.4 | 42.0 | 34.9 |
| H | 27.0 | 37.5 | 52.3 | 53.1 | 54.1 | 54.0 | 61.6 | 37.2 | 27.1 |
| (c) $e^{\theta} = 2.0$ | | | | | | | | | |
| N | 60.6 | 70.0 | 86.5 | 91.6 | 94.7 | 96.2 | 95.7 | 86.2 | 61.4 |
| B | 58.9 | 68.7 | 88.1 | 90.2 | 92.9 | 97.6 | 95.6 | 88.8 | 65.3 |
| BOOT | 58.8 | 68.5 | 84.9 | 90.5 | 94.8 | 95.8 | 93.2 | 85.1 | 65.2 |
| H | 52.7 | 65.2 | 79.3 | 87.3 | 90.6 | 90.9 | 90.8 | 80.4 | 58.6 |

with covariance matrix $t\boldsymbol{\Sigma}$ at time $t$. A simple extension of their method to the present problem of testing $H_0 : \theta = \theta_0$ is to replace $t\boldsymbol{\Sigma}$ by $-\ddot{l}_t(\hat{\beta}_t, \hat{\theta}_t)$, yielding a bias-corrected test statistic of the form $W_T - \hat{\rho}_T(\hat{\beta}_T - \bar{\beta}_T)/V_T^{1/2}$, where $\bar{\beta}_T$ is the unbiased estimator of $\beta$ as in Liu and Hall (1999) and the correlation coefficient $\hat{\rho}_T$ can be determined from $(-\ddot{l}_T(\hat{\beta}_T, \hat{\theta}_T))^{-1}$. This is the bias-correction method used in Table 7.5. However, its asymptotic justification, as $n \to \infty$, requires the limiting correlation matrix to be equal for all $t$, as noted by Hall and Yakir (2003, p. 599).

## 7.7 Supplements and Problems

1. *Anscombe's theorem and fixed-width confidence intervals*

   The condition $\max_{\varepsilon n \le j \le n} |Z_n - Z_j| \xrightarrow{P} 0$ as $n \to \infty$ and $\varepsilon \uparrow 1$ in (7.1) is called "uniform continuity in probability." Anscombe's theorem is applicable to asymptotically normal sequences $Z_n$ and random times $T_n$ satisfying (7.1). It basically says that $T_n$ can be treated as nonrandom and no adjustments for stopping are needed for constructing confidence intervals for $\theta$ based on $Z_{T_n}$. Although this does not work well for stopping times associated with sequential tests of $\theta$, it has been found to provide adequate approximations when the stopping time is related to $\widehat{\text{se}}(\hat{\theta}_i)$, as in the case of fixed-width confidence intervals, for which $T_n$ is of the form $\inf\{i : z_{1-\alpha}\, \widehat{\text{se}}(\hat{\theta}_i) \le d_n\}$ so that $\hat{\theta}_{T_n} \pm d_n$ is an approximately $(1-2\alpha)$-level confidence interval, with fixed width $2d_n$, for $\theta_i$. Unlike stopping rules associated with sequential tests of $\theta$, stopping when $\widehat{\text{se}}(\hat{\theta}_i)$ is sufficiently small typically introduces negligible bias in $\hat{\theta}_{T_n}$. Section 4.1 of Lai (2001) gives a review of the literature on fixed-width confidence intervals. Note that the normal quantile $z_{1-\alpha}$ in the stopping time $T_n$ already presupposes that $(\hat{\theta}_{T_n} - \theta)/\widehat{\text{se}}(\hat{\theta}_{T_n})$ is approximately standard normal.

2. *Woodroofe's modified pivot following a sequential test*
Woodroofe (1986) has shown that the bias-corrected pivot (7.6) satisfies

$$P(R_1(\mu) \le x) \doteq \Phi(x) - (2a)^{-1} x\phi(x)[(d/d\mu)\kappa^{1/2}(\mu)]^2 \tag{7.41}$$

in a very weak sense, that is, the integral of the left-hand side of (7.41) with respect to $\xi(\mu)d\mu$ has an asymptotic expansion given by that of the right-hand side for a large class of prior densities $\xi$, where $\phi$ and $\Phi$ denote the standard normal density and distribution function, respectively. His underlying idea is that since posterior distributions are unaffected by optional stopping, the posterior probability can be expanded about a normal limit, for example, by using Johnson's (1970) expansions that we have referred to in Sect. 1.5. Note that the derivation of the bias-corrected pivot is based on frequentist calculations; in particular, the expectation of $R_0(\mu)$ in Sect. 7.1.4 is taken under the true probability measure. Woodroofe's very weak expansion is for the coverage error, only in an average sense with $\xi$ centered around the true parameter $\mu_0$, of the confidence interval using the corrected pivot $R_1(\mu)$. The expansion itself does not lead to a new pivot. Moreover, since it takes a Bayesian approach to circumvent difficulties in analyzing the randomly stopped pivot, it does not suggest corrections for the pivot due to optional stopping.

3. *Multivariate and nonparamteric extensions of the pivot $R_1(\mu)$*
Lai et al. (2006a) have extended Woodroofe's (1992) arguments to derive the corrected pivot (7.6) as follows. Let $\boldsymbol{X}, \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ be i.i.d. $d \times 1$ random vectors with $E(\boldsymbol{X}) = \boldsymbol{\mu}$, $\text{Cov}(\boldsymbol{X}) = \boldsymbol{V}$, and $E\|\boldsymbol{X}\|^r < \infty$ for some $r > 3$. Let $h : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable in some neighborhood of $\boldsymbol{\mu}$. Consider the stopping rule $T = \min\{n_0(a), \max(t_a, n_1(a))\}$, where $t_a = \inf\{n \ge 1 : n g(\boldsymbol{S}_n/n) \ge a\}$ and $g : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable in some neighborhood of $\boldsymbol{\mu}$. Suppose $\varepsilon_0 < g(\boldsymbol{\mu}) < \varepsilon_1$. Then application of the strong law of large numbers in conjunction with Taylor's theorem yields

$$\begin{aligned}
&\sqrt{T}\{h(\bar{\boldsymbol{X}}_T) - h(\boldsymbol{\mu})\} \\
&\doteq \sqrt{T}(\nabla h(\boldsymbol{\mu}))^{\mathrm{T}}(\bar{\boldsymbol{X}}_T - \boldsymbol{\mu}) + \sqrt{T}(\bar{\boldsymbol{X}}_T - \boldsymbol{\mu})^{\mathrm{T}}\nabla^2 h(\boldsymbol{\mu})(\bar{\boldsymbol{X}}_T - \boldsymbol{\mu})/2 \\
&\doteq \frac{1}{\sqrt{a}} g^{1/2}(\boldsymbol{S}_T/T)(\boldsymbol{S}_T - \mu T)^{\mathrm{T}}\nabla h(\boldsymbol{\mu}) + \frac{1}{2\sqrt{T}}\{T(\bar{\boldsymbol{X}}_T - \boldsymbol{\mu})^{\mathrm{T}}\nabla^2 h(\boldsymbol{\mu})(\bar{\boldsymbol{X}}_T - \boldsymbol{\mu})\},
\end{aligned} \tag{7.42}$$

in which the last approximate equality follows from $Tg(\boldsymbol{S}_T/T) \doteq a$ (ignoring overshoot) so that $\sqrt{T} \doteq \sqrt{a}/g^{1/2}(\boldsymbol{S}_T/T) \doteq \{a/g(\boldsymbol{\mu})\}^{1/2}$. By Wald's equation (3.7), $E\{g^{1/2}(\boldsymbol{\mu})(\boldsymbol{S}_T - \boldsymbol{\mu}T)^{\mathrm{T}}\nabla h(\boldsymbol{\mu})\} = 0$. Moreover,

$$g^{1/2}(\boldsymbol{S}_T/T) - g^{1/2}(\boldsymbol{\mu}) \doteq \{(\nabla g(\boldsymbol{\mu}))^{\mathrm{T}}(\boldsymbol{S}_T - T\boldsymbol{\mu})\}/\{2g^{1/2}(\boldsymbol{\mu})T\}. \tag{7.43}$$

By Anscombe's theorem, $\sqrt{T}(\bar{\boldsymbol{X}}_T - \boldsymbol{\mu}) = (\boldsymbol{S}_T - \boldsymbol{\mu}T)/\sqrt{T}$ has a limiting $N(\boldsymbol{0}, \boldsymbol{V})$ distribution. Combining (7.42) with (7.43) and taking expectation, it can be shown that

$$
\begin{aligned}
&E[\sqrt{T}\{h(\bar{\boldsymbol{X}}_T) - h(\boldsymbol{\mu})\}] \\
&\quad = \frac{(\nabla g(\boldsymbol{\mu}))^{\mathrm{T}} \boldsymbol{V} \nabla h(\boldsymbol{\mu})}{2(ag(\boldsymbol{\mu}))^{1/2}} + \frac{1}{2}\left(\frac{g(\boldsymbol{\mu})}{a}\right)^{1/2} \operatorname{tr}(\nabla^2 h(\boldsymbol{\mu})\boldsymbol{V}) + o(a^{-1/2}). \qquad (7.44)
\end{aligned}
$$

The second term on the right-hand side of (7.44) follows from $E(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{Z}) = \operatorname{tr}(\boldsymbol{A}\boldsymbol{V})$ if $\boldsymbol{A}$ is a nonrandom matrix and $\boldsymbol{Z}$ is a random vector with $E(\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}) = \boldsymbol{V}$.

For $g(\boldsymbol{\mu}) < \varepsilon_0$ (or $g(\boldsymbol{\mu}) > \varepsilon_1$), stopping occurs at $n_0(a)$ (or $n_1(a)$) with probability approaching 1, and it can be shown that

$$
E[\sqrt{T}\{h(\bar{\boldsymbol{X}}_T) - h(\boldsymbol{\mu})\}] = \frac{1}{2}(n_i(a))^{-1/2} \operatorname{tr}(\nabla^2 h(\boldsymbol{\mu})\boldsymbol{V}) + o(a^{1/2}), \qquad (7.45)
$$

with $i = 0$ or $1$ according as $g(\boldsymbol{\mu}) < \varepsilon_0$ or $g(\boldsymbol{\mu}) > \varepsilon_1$. Let $\kappa(\boldsymbol{\mu}) = \max\{\varepsilon_0, \min(g(\boldsymbol{\mu}), \varepsilon_1)\}$, and note that $\nabla \kappa^{1/2}(\boldsymbol{\mu}) = \frac{1}{2}\nabla g(\boldsymbol{\mu})/(g(\boldsymbol{\mu}))^{1/2}$ if $\varepsilon_0 < g(\boldsymbol{\mu}) < \varepsilon_1$, and $\nabla \kappa^{1/2}(\boldsymbol{\mu}) = 0$ if $g(\boldsymbol{\mu}) < \varepsilon_0$ or $g(\boldsymbol{\mu}) > \varepsilon_1$. Recalling that $1/n_0(\boldsymbol{\mu}) \sim \varepsilon_0/a$ and $1/n_1(\boldsymbol{\mu}) \sim \varepsilon_1/a$, Lai et al. (2006a) combine (7.44) and (7.45) into

$$
E[\sqrt{T}\{h(\bar{\boldsymbol{X}}_T) - h(\boldsymbol{\mu})\}] = b(\boldsymbol{\mu}, \boldsymbol{V})(\kappa(\boldsymbol{\mu})/a)^{1/2} + o(a^{1/2}), \qquad (7.46)
$$

where

$$
b(\boldsymbol{\mu}, \boldsymbol{V}) = (\nabla \kappa^{1/2}(\boldsymbol{\mu}))^{\mathrm{T}} \boldsymbol{V} \nabla h(\boldsymbol{\mu})/\kappa^{1/2}(\boldsymbol{\mu}) + \operatorname{tr}(\nabla^2 h(\boldsymbol{\mu})\boldsymbol{V})/2. \qquad (7.47)
$$

For the special case $d{=}1$, $h(\mu){=}\mu$, and $V{=}1$, $b(\mu, V){=}[(d/d\mu)\kappa^{1/2}(\mu)]/\kappa^{1/2}(\mu)$, which agrees with that in (7.6).

4. *Unbiased or bias-corrected estimators following a sequential test*

The preceding two supplements consider bias-corrected pivots for confidence intervals. Bias-corrected maximum likelihood estimators following a sequential test have been introduced by Siegmund (1978, Sect. 3) and Whitehead (1986). For group sequential tests involving normal observations with unknown mean $\mu$ and common known variance, Emerson and Fleming (1990) propose to use the first-stage sample mean as a preliminary unbiased estimator and to condition it on $(T, S_T)$ when $T$ is associated with the O'Brien–Fleming or Pocock test, since $(T, S_T)$ is a sufficient statistic for $\mu$; see Supplement 5. Their simulation study shows, however, that this has markedly larger mean squared error than Whitehead's bias-corrected estimator, even though it is unbiased while Whitehead's is not. Liu and Hall (1999) prove that the sufficient statistic $(T, S_T)$ is not complete for $\mu$ and that there exist infinitely many unbiased estimators of $\mu$ but none has uniformly minimum variance. They show, however, that there is an unbiased estimator that minimizes the variance of unbiased estimators in a restricted class of so-called “truncation-adaptive” estimators.

5. *Sufficiency of* $(T, S_T)$

   (a) One way to prove sufficiency is to write down the conditional density of $(X_1, \dots, X_{m-1})$ given $T = m$ and $S_m = s$ and show that it does not depend on $\mu$. Carry this out for $T = \inf\{n \geq 1 : S_n \geq b_n \text{ or } S_n \leq a_n\} \wedge M$ with $M = 3$.

   (b) Another way is to argue probabilistically using the fact that conditional on $S_m = s$, $X_{m-1}, X_{m-2}, \dots$ are normal with mean $s/m$ and therefore do not depend on $\mu$. Although this does not involve the stopping rule $T$, the stopping rule only stipulates that the $X_i$ $(1 \leq i \leq m-1)$ must fall in a certain region when we evaluate their joint conditional density given $T = m$ and $S_m = s$, and this region does not depend on $\mu$. In fact, Siegmund (1985, Sect. IV.2) uses this idea and a time-reversal argument to compute the type I error of repeated significance tests.

   (c) We can in fact go far beyond normal random walks and stopping rules of sequential tests. A sequence of statistics $\boldsymbol{S}_j = \boldsymbol{S}_j(\boldsymbol{X}_1, \dots, \boldsymbol{X}_j)$ is called a *sufficient sequence* for a parameter vector $\boldsymbol{\theta}$ if for every $j$, $\boldsymbol{S}_j$ is sufficient for $\boldsymbol{\theta}$ based on $\boldsymbol{X}_1, \dots, \boldsymbol{X}_j$. Show that if $\{\boldsymbol{S}_j\}$ is a sufficient sequence for $\boldsymbol{\theta}$, then for any stopping time $T$, $(T, \boldsymbol{S}_T)$ is sufficient for $\boldsymbol{\theta}$ in the sense that the conditional distribution of $(\boldsymbol{X}_1, \dots, \boldsymbol{X}_T)$ given $(T, \boldsymbol{S}_T)$ does not depend on $\boldsymbol{\theta}$.

6. *Coverage errors for bootstrap and hybrid resampling confidence sets*
   We begin with a review of the theory on second-order accuracy of bootstrap confidence intervals based on a fixed number of i.i.d. observations $X_1, \dots, X_n$. Hall (1992) shows that the essence of this theory lies in comparing the Edgeworth expansions of (a) the sampling distribution of the approximate pivot used to construct the confidence interval and (b) the bootstrap distribution of that based on the empirical distribution (for nonparametric bootstrap) or the estimated parametric distribution (usually by maximum likelihood for parametric bootstrap). Thus, (a) refers to the actual probability measure, which is unknown, and (b) refers to the estimated measure that is used to generate bootstrap samples. Hall (1992) focuses on smooth functions of means as the parameter of interest, for which Edgeworth expansions are available. These expansions in turn yield Cornish–Fisher expansions for bootstrap quantiles that are used to construct the bootstrap confidence intervals, and the coverage error of a bootstrap confidence interval can be evaluated by using the Edgeworth expansion. This approach can be extended to much more complex settings than smooth functions of means, and Gross and Lai (1996) have shown how it works in the setting of censored survival data considered in Sect. 6.1. Carrying this idea out for stopped random walks is much harder because the stopping time is integer valued and approximating its lattice distribution by a continuous normal distribution function plus higher-order correction terms results in certain sawtooth functions that the bootstrap distribution cannot match to the actual sampling distribution, as shown by Lai and Wang (1994).

Lai et al. (2006a) have reported several simulation studies on the coverage errors of confidence intervals for $\mu$ using the pivot $R_0(\mu)$, the bias-corrected pivot $R_1(\mu)$, another modified pivot proposed by Woodroofe (1992), and bootstrap confidence intervals using $R_0$ and $R_1$, together with hybrid resampling intervals, in the setting of fully sequential tests for which Woodroofe's (1986) very weak expansions and Lai and Wang's (1994) Edgeworth expansions have been developed. The finding is that the hybrid resampling method produces coverage errors close to the nominal values over the range of parameter values of $\mu$ studied, but the other methods may be quite inaccurate at some parameter values.

We next describe how Chuang and Lai (1998) use Edgeworth expansions to prove the second-order accuracy of hybrid resampling confidence sets (7.12) and (7.14). Let

$$T = \min\{n_j : S_{n_j} \geq \gamma_j \text{ or } S_{n_j} \leq \lambda_j\} \quad (\min \emptyset = n_k), \tag{7.48}$$

in which $n_1 < \cdots < n_k = n$ are positive integers and $\lambda_j < \gamma_j$ are real numbers. Here $X_1, X_2, \ldots, X_{n_k}$ are i.i.d. random variables with common characteristic function $\psi$ such that for some constant $C$,

$$\liminf_{n\to\infty} (n_j - n_{j-1})/n > 0 \quad (1 \leq j \leq k), \tag{7.49}$$

$$E(X_1 - \mu)^4 \leq C, \quad \limsup_{|t|\to\infty} |\psi(t)| < 1. \tag{7.50}$$

The second part of (7.50) is commonly called "Cramér's condition" for Edgeworth expansions. Conditions (7.49) and (7.50) ensure that

$$P_\mu \Big\{ \tilde{\lambda}_1 < (S_{n_1} - \mu n_1)/n^{1/2} < \tilde{\gamma}_1, \ldots, \tilde{\lambda}_{j-1} < (S_{n_{j-1}} - \mu n_{j-1})/n^{1/2} < \tilde{\gamma}_{j-1}, \\ (S_{n_j} - \mu n_j)/n^{1/2} \in (-\infty, z] \cap ((-\infty, \tilde{\lambda}_j] \cup [\tilde{\gamma}_j, \infty)) \Big\} \tag{7.51}$$

has an Edgeworth expansion of the form

$$\int \cdots \int_{C_j} \prod_{i=1}^{j} \phi(x_i) \Big\{ 1 + (n_i - n_{i-1})^{-1/2} Q_1(x_i) + (n_i - n_{i-1})^{-1} Q_2(x_i) \Big\} dx_i + o(n^{-1}) \tag{7.52}$$

uniformly in $\mu$, $z$, $\tilde{\lambda}_1$, $\tilde{\gamma}_1, \ldots \tilde{\lambda}_j$, and $\tilde{\gamma}_j$, where $\phi$ is the standard normal density function, $Q_1$ and $Q_2$ are polynomials, and

$$C_j = \Big\{ (x_1, \ldots, x_j) : \tilde{\lambda}_l < (n_1/n)^{1/2} x_1 + \cdots + \{(n_l - n_{l-1})/n\}^{1/2} x_l < \tilde{\gamma}_l \ (l < j) \\ \text{and } (n_1/n)^{1/2} x_1 + \cdots + \{(n_j - n_{j-1})/n\}^{1/2} x_j \in (-\infty, z] \cap ((-\infty, \tilde{\lambda}_j] \cup [\tilde{\gamma}_j, \infty)) \Big\},$$

since the $X_i$ have common variance 1; see Lemma 5.4 of Hall (1992). If $\tilde{\lambda}_l = (\lambda_l - \mu n_l)/n^{1/2}$ and $\tilde{\gamma}_l = (\lambda_l - \mu n_l)/n^{1/2}$ for $1 \le l \le j$, then (7.51) can be expressed as $P_\mu\{T = n_j, (S_T - T\mu)/T^{1/2} \le z\}$. Hence $P_\mu\{(S_T - T\mu)/T^{1/2} \le z\}$ is a sum of probabilities of the form (7.51) which can be approximated by Edgeworth expansions within an $o(n^{-1})$ error, uniformly in $\mu$ and $z$. Define $\hat{G}_T$ and $\varepsilon_i$ as in the first paragraph of Sect. 7.2.2. An argument similar to that in Sect. 5.2 of Hall (1992) shows that

$$P\left\{\tilde{\lambda}_1 < \left(\sum_{i=1}^{n_1} \varepsilon_i\right) \Big/ n^{1/2} < \tilde{\gamma}_1, \ldots, \tilde{\lambda}_{j-1} < \left(\sum_{i=1}^{n_{j-1}} \varepsilon_i\right) \Big/ n^{1/2} < \tilde{\gamma}_{j-1},\right.$$
$$\left.\left(\sum_{i=1}^{n_j} \varepsilon_i\right) \Big/ n^{1/2} \in (-\infty, z] \cap ((-\infty, \tilde{\lambda}_j] \cup [\tilde{\gamma}_j, \infty)) \,\middle|\, \hat{G}_T \right\}$$

also has the Edgeworth expansion (7.52) with $o(n^{-1})$ replaced with $o_p(n^{-1})$ and with $Q_1$, $Q_2$ replaced by $\hat{Q}_1$, $\hat{Q}_2$ such that $\sup_x |\hat{Q}_l(x) - Q_l(x)| = O_p(n^{-1/2})$ for $l = 1, 2$. Using this and an argument similar to that in Sects. 3.5 and 5.2 of Hall (1992), Chuang and Lai (1998) obtain the coverage probability $1 - 2\alpha + O(n^{-1})$ for the hybrid resampling confidence interval (7.12).

Similarly, for the hybrid resampling confidence set (7.14) using Siegmund's ordering, note that for $s \ge \gamma_j$, $P_\mu\{(T, S_T) \ge (n_j, s)\}$ is equal to

$$P_\mu\Big\{\tilde{\lambda}_1 < (S_{n_1} - \mu n_1)/n^{1/2} < \tilde{\gamma}_1, \ldots, \tilde{\lambda}_{j-1} < (S_{n_{j-1}} - \mu n_{j-1})/n^{1/2} < \tilde{\gamma}_{j-1},$$
$$(S_{n_j} - \mu n_j)/n^{1/2} \ge s\Big\}$$
$$+ \sum_{i=1}^{j-1} P_\mu\Big\{\tilde{\lambda}_t < (S_{n_t} - \mu n_t)/n^{1/2} < \tilde{\gamma}_t \text{ for } t < i, \;\; (S_{n_i} - \mu n_i)/n^{1/2} \ge \tilde{\gamma}_i\Big\},$$

and that for $s \le \lambda_j$, $P_\mu\{(T, S_T) \le (n_j, s)\}$ is equal to

$$P_\mu\Big\{\tilde{\lambda}_1 < (S_{n_1} - \mu n_1)/n^{1/2} < \tilde{\gamma}_1, \ldots, \tilde{\lambda}_{j-1} < (S_{n_{j-1}} - \mu n_{j-1})/n^{1/2} < \tilde{\gamma}_{j-1},$$
$$(S_{n_j} - \mu n_j)/n^{1/2} \le s\Big\}$$
$$+ \sum_{i=1}^{j-1} P_\mu\Big\{\tilde{\lambda}_t < (S_{n_t} - \mu n_t)/n^{1/2} < \tilde{\gamma}_t \text{ for } t < i, \;\; (S_{n_i} - \mu n_i)/n^{1/2} \le \tilde{\lambda}_i\Big\}.$$

This shows that the preceding argument involving Edgeworth expansions for (7.12) can again be used to show that the hybrid resampling confidence set (7.14) also has coverage probability $1 - 2\alpha + O(n^{-1})$.

7. *Confidence intervals for median survival in the Cox regression model*
Median survival times and their associated confidence intervals are often used to summarize the survival outcome of patients in clinical trials with failure-time endpoints. Interval estimation for the median $m$ has been recognized to be a difficult problem, even when there are no covariates and the sample consists of fully observable data without censoring. Lai and Su (2006) have reviewed bootstrap methods to address this estimation problem in the absence of censoring and covariates and an alternative test-based approach that inverts a generalized sign test similar to the approach proposed by Brookmeyer and Crowley (1982). The difficulty is compounded by the presence of covariates and censoring in survival studies. Lai and Su (2006) have generalized the Brookmeyer–Crowley confidence interval to that for the median $m(\boldsymbol{x})$ in the Cox model (6.8) with time-invariant covariate vector $\boldsymbol{x}$. Specifically, $m(\boldsymbol{x})$ is the median of the survival distribution $S(t|\boldsymbol{x}) = (S_0(t))^{\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x})}$, where $S_0$ is the survival distribution with hazard function $\lambda_0$ given in (6.8). Instead of working directly with $S(\cdot|\boldsymbol{x})$, they find it more convenient to work with the cumulative hazard function $\Lambda(\cdot|\boldsymbol{x})$. The key idea underlying their approach is that it is much easier to find an approximate pivot

$$\{\hat{\Lambda}(m|\boldsymbol{x}) - \Lambda(m|\boldsymbol{x})\}/\hat{\sigma}(m|\boldsymbol{x}) \tag{7.53}$$

at given $m$ than finding an approximate pivot that involves the estimated median $\hat{m}(\boldsymbol{x})$. This explains why they use a test-based confidence set:

$$\left\{m : \hat{c}_\alpha(m) \le (\hat{\Lambda}(m|\boldsymbol{x}) - \log 2)/\hat{\sigma}(m|\boldsymbol{x}) \le \hat{c}_{1-\alpha}(m)\right\}, \tag{7.54}$$

in which $\hat{c}_\alpha(m)$ and $\hat{c}_{1-\alpha}(m)$ are the quantiles of the bootstrap distribution of $(\hat{\Lambda}(m|\boldsymbol{x}) - \Lambda(m|\boldsymbol{x}))/\hat{\sigma}(m|\boldsymbol{x})$ for given $m$, and include $m$ in the confidence set if the null hypothesis $\Lambda(m|\boldsymbol{x}) = \log 2$ (which is equivalent to $S(m|\boldsymbol{x}) = 1/2$ corresponding to median) is accepted.

Lai et al. (2009) subsequently extended this approach to the time-sequential setting in which a secondary analysis involves estimation of $m(\boldsymbol{x})$. To be specific, let $\boldsymbol{\gamma} = (\beta, \boldsymbol{\theta}^T)^T$ and $\boldsymbol{x}_i = (x_i, \boldsymbol{u}_i^T)^T$ in the proportional hazards model (7.27). Given a covariate vector $\boldsymbol{x}$, let $\hat{\Lambda}_t(s|\boldsymbol{x}) = \hat{\Lambda}_t(s)e^{\hat{\boldsymbol{\gamma}}_t^{\mathrm{T}}\boldsymbol{x}}$ be the estimate, at time $t$, of the cumulative hazard function $\Lambda(s|\boldsymbol{x})$, where $\hat{\boldsymbol{\gamma}}_t = (\hat{\beta}_t, \hat{\boldsymbol{\theta}}_t^{\mathrm{T}})^{\mathrm{T}}$ and $\hat{\Lambda}_t, \hat{\beta}_t, \hat{\boldsymbol{\theta}}_t$ are the same as in Sect. 7.5.2. To construct an upper $(1-\alpha)$-level confidence bound for $m(\boldsymbol{x})$ of the survival distribution $e^{-\Lambda(\cdot|\boldsymbol{x})}$, Lai et al. (2009) use a test-based approach that considers testing $H_0 : \Lambda(m|\boldsymbol{x}) = \log 2$ with the test statistic $W_t(m)$, where

$$W_t(m) = \left\{\hat{\Lambda}_t(m|\boldsymbol{x}) - \Lambda_t(m|\boldsymbol{x})\right\}/\hat{v}_t^{1/2}(m|\boldsymbol{x}), \tag{7.55}$$

as in (7.53). For a time-sequential trial with stopping rule $T$ of the form (7.32), we can use the ordering scheme (7.17) to evaluate the $p$-value, denoted by $\hat{p}(m)$, of the test based on $W_T(m)$:

**Table 7.6** Coverage errors in percentages for lower (L) and upper (U) confidence limits and coverage probabilities (P) of confidence intervals for the median survival of subjects with covariates $x = 1$ and $u = 0.5$

| | Naive Normal | | | Hybrid | | | Bootstrap | | |
|---|---|---|---|---|---|---|---|---|---|
| $e^{\beta}$ | L | U | P | L | U | P | L | U | P |
| (a) $e^{\theta} = 1.0$ | | | | | | | | | |
| 1.00 | 8.3 | 3.8 | 87.9 | 3.7 | 4.1 | 92.2 | 5.7 | 5.4 | 88.9 |
| 0.90 | 9.0 | 3.3 | 87.7 | 3.6 | 3.9 | 92.5 | 4.4 | 4.2 | 91.4 |
| 0.80 | 9.1 | 3.1 | 87.8 | 4.0 | 3.9 | 92.1 | 5.1 | 5.6 | 89.3 |
| 0.70 | 8.1 | 2.6 | 89.3 | 4.1 | 4.4 | 91.5 | 5.7 | 6.0 | 88.3 |
| 0.60 | 6.8 | 3.5 | 89.7 | 3.2 | 3.9 | 92.9 | 4.2 | 5.1 | 90.7 |
| 0.50 | 6.9 | 3.0 | 90.1 | 4.5 | 4.7 | 90.8 | 6.2 | 5.1 | 88.7 |
| (b) $e^{\theta} = 1.5$ | | | | | | | | | |
| 1.00 | 7.2 | 3.3 | 89.5 | 3.7 | 4.3 | 92.0 | 5.8 | 5.5 | 88.7 |
| 0.90 | 7.7 | 3.6 | 88.7 | 4.1 | 4.1 | 91.8 | 4.3 | 5.6 | 90.1 |
| 0.80 | 7.5 | 3.0 | 89.5 | 3.9 | 3.6 | 92.5 | 6.1 | 5.5 | 88.4 |
| 0.70 | 8.1 | 3.4 | 88.5 | 3.5 | 4.5 | 92.0 | 4.8 | 5.9 | 89.3 |
| 0.60 | 6.6 | 3.2 | 90.2 | 3.7 | 3.6 | 92.7 | 4.3 | 4.1 | 91.6 |
| 0.50 | 6.8 | 2.8 | 90.4 | 4.2 | 4.7 | 91.1 | 5.1 | 5.5 | 89.4 |
| (c) $e^{\theta} = 2.0$ | | | | | | | | | |
| 1.00 | 8.2 | 2.9 | 88.9 | 4.0 | 4.2 | 91.8 | 5.7 | 6.0 | 88.3 |
| 0.90 | 8.1 | 3.3 | 88.6 | 3.6 | 4.4 | 92.0 | 4.3 | 4.3 | 91.4 |
| 0.80 | 7.7 | 2.8 | 89.5 | 3.9 | 4.2 | 91.9 | 5.3 | 5.8 | 88.9 |
| 0.70 | 7.4 | 3.3 | 89.3 | 4.3 | 3.9 | 91.8 | 4.4 | 4.9 | 90.7 |
| 0.60 | 6.5 | 3.4 | 90.1 | 4.9 | 3.3 | 91.8 | 5.2 | 4.9 | 89.9 |
| 0.50 | 7.2 | 3.2 | 89.6 | 3.6 | 4.3 | 92.1 | 6.0 | 5.6 | 88.4 |

$$\hat{p}(m) = \sup_{\beta} \hat{P}_{\beta}\left\{(T, W_T(m)) \geq (\tau, w_{\tau}(m))\right\}, \tag{7.56}$$

in which $(\tau, w_{\tau}(m))$ is the observed value of $(T, W_T(m))$ and $\hat{P}_{\beta}$ uses $\tilde{\boldsymbol{\theta}}_T(\beta)$ and $\hat{G}_T = 1 - e^{-\hat{\Lambda}_T}$ to replace $\boldsymbol{\theta}$ and $G$ in $P_{\beta,\boldsymbol{\theta}}$, where $\tilde{\boldsymbol{\theta}}_t(\beta)$ is the maximizer of $l_t(\beta, \boldsymbol{\theta})$ for a given $\beta$. The probability in (7.56) can be evaluated by Monte Carlo simulations. The hybrid resampling confidence set for $m(\boldsymbol{x})$, with nominal confidence level $1 - 2\alpha$, is $\{m : \alpha < \hat{p}(m) < 1 - \alpha\}$. When $\hat{p}(m)$ is monotone in $m$, the confidence set is an interval, which is typically the case.

For the time-sequential trial in Sect. 7.6.2, Lai et al. (2009) also consider the problem of constructing confidence intervals for median survival in patients receiving the new treatment, $x = 1$, and having risk factor $u = 0.5$. Table 7.6 gives the coverage errors of the nominal 90% confidence intervals for the median survival, obtained by (a) the hybrid resampling method, (b) the naive normal approximation that assumes $W_T(m)$ to be approximately normal, with $W_t(m)$ given by (7.55), and (c) the bootstrap method which resamples from $P_{\hat{\boldsymbol{\gamma}}_T, \hat{\Lambda}_T}$ to evaluate the *p*-value. Each result is based on 2000 simulations; the hybrid and bootstrap methods use $B = 2000$ resamples. Table 7.6 shows that the hybrid resampling method maintains the nominal coverage errors of both the lower and upper confidence limits, and the naive method has substantially larger coverage errors for the lower confidence limits while the bootstrap method has inflated coverage errors for some lower and upper confidence limits.

# Chapter 8
# Adaptive Design of Confirmatory Trials

Because of the ethical and economic considerations in the design of clinical trials to test the efficacy of new treatments and because of lack of information on the magnitude and sampling variability of the treatment effect at the design stage, there has been increasing interest from the biopharmaceutical industry in sequential methods that can adapt to information acquired during the course of the trial. Beginning with Bauer (1989), who introduced sequential adaptive test strategies over a planned series of separate trials, and Wittes and Brittain (1990), who discussed internal pilot studies, a large literature has grown on adaptive design of clinical trials. Depending on the topics covered, the term "adaptive design" in this literature is sometimes replaced by "sample size re-estimation," "trial extension," or "internal pilot studies." In standard clinical trial designs, the sample size is determined by the power at a given alternative, but in practice, it is often difficult for investigators to specify a realistic alternative at which sample size determination can be based. Although a standard method to address this difficulty is to carry out a preliminary pilot study, the results from a small pilot study may be difficult to interpret and apply, as pointed out by Wittes and Brittain (1990), who proposed to treat the first stage of a two-stage clinical trial as an internal pilot from which the overall sample size can be re-estimated. The problem of sample size re-estimation based on observed treatment difference at some time before the prescheduled end of a clinical trial has attracted considerable attention since the 1990s. Much of the literature has focused on finding ways to adjust the test statistics after midcourse sample size modification so that the type I error probability is maintained at the prescribed level. Section 8.1.1 gives a summary of the major developments and the methods proposed. Section 8.1.2 describes their extensions to the group sequential setting.

By making use of a generalization of the Neyman–Pearson lemma, Tsiatis and Mehta (2003) showed that these adaptive tests of a simple null versus a simple alternative hypothesis are inefficient because they are not based on likelihood ratio statistics. Jennison and Turnbull (2003) gave a general weighted form of these test statistics and demonstrated in simulation studies that the adaptive tests performed

J. Bartroff et al., *Sequential Experimentation in Clinical Trials,*
Springer Series in Statistics 298, DOI 10.1007/978-1-4614-6114-2_8,

considerably worse than group sequential tests. Assuming normally distributed outcomes with known variances, Jennison and Turnbull (2006a) introduced adaptive group sequential tests that choose the $j$th group size and stopping boundary on the basis of the cumulative sample size $n_{j-1}$ and the sample sum $S_{n_{j-1}}$ over the first $j-1$ groups and that are optimal, in the sense of minimizing a weighted average of the expected sample sizes over a collection of parameter values, subject to prescribed error probabilities at the null and a given alternative hypothesis. They showed how the corresponding optimization problem can be solved numerically by using backward induction algorithms (see Sect. 3.6.1). Jennison and Turnbull (2006b) found that standard (nonadaptive) group sequential tests with the first stage chosen optimally are nearly as efficient as their optimal adaptive tests.

Except for Jennison and Turnbull's optimal adaptive group sequential tests and the extensions of the sample size re-estimation approach to group sequential testing, the midcourse sample size re-estimation literature reviewed in Sect. 8.1 has focused on two-stage designs whose second-stage sample size is determined by the results from the first stage (internal pilot), following the seminal work of Stein (1945) in this area. Although this approach is intuitively appealing, it does not adjust for the uncertainty in the first-stage parameter estimates that are used to determine the second-stage sample size. Moreover, it considers primarily the special problem of comparing the means of the two normal populations, using the central limit theorem for extensions to more general situations. The case of unknown common variance at a prespecified alternative for the mean difference, which was considered first, is considered in Sect. 8.1.3.

A unified treatment, developed by Bartroff and Lai (2008a,b), of both cases in the general framework of multiparameter exponential families is presented in Sect. 8.2. It uses efficient generalized likelihood ratio (GLR) statistics in this framework and adds a third stage to adjust for the sampling variability of the first-stage parameter estimates that determine the second-stage sample size. The possibility of adding a third stage to improve two-stage designs dated back to Lorden (1983). Whereas Lorden used crude upper bounds for the type I error probability that are too conservative for practical applications, Bartroff and Lai (2008a) overcame this difficulty by modifying the numerical methods in Sect. 4.3 to compute the type I error probability and also extended the three-stage test to multiparameter and multi-armed settings, thus greatly broadening the scope of these efficient adaptive designs. Section 8.3 summarizes the simulation studies of Bartroff and Lai (2008a,b), comparing their approach to adaptive designs with other approaches in the literature.

The adaptive methods in Sect. 8.2 can be regarded as modifications of the group sequential GLR tests in Sect. 4.2. In Sect. 8.4 we give another modification, proposed by Lai et al. (2006c), for adaptive choice between superiority and non-inferiority objectives of a new treatment during interim analyses of a clinical trial to test the treatment's efficacy.

## 8.1 Internal Pilot, Midcourse Sample Size Re-estimation and Trial Extensions

As in group sequential designs considered in Sect. 4.1, most of the literature on adaptive designs focus on the prototypical problem of testing a normal mean when the variance is known. The case of unknown variance is also considered when the "internal pilot" is used to estimate the variance. The canonical problem considered in this section is testing the hypothesis $H_0 : \mu_X = \mu_Y$ versus the two-sided alternative $\mu_X \neq \mu_Y$ for the mean of two independent normal populations with common, unknown variance, and based on i.i.d. observations $X_1, X_2, \cdots \sim N(\mu_X, \sigma^2)$ and $Y_1, Y_2, \cdots \sim N(\mu_X, \sigma^2)$. Let $t_{\nu,\alpha}$ denote the upper $\alpha$-quantile of the $t$-distribution with $\nu$ degrees of freedom.

### *8.1.1 Stein's Two-Stage Procedure, with the First Stage to Estimate the Variance as an Internal Pilot*

After Dantzig's (1940) results showed that no fixed-sample test of $H_0$ can be guaranteed to achieve power at a given level if the variance $\sigma^2$ is unknown, Stein (1945) developed an elegant solution to the problem in a two-stage test whose power is independent of the variance. In its first stage, Stein's test samples $n_0$ observations from each of the two normal populations and computes the usual unbiased estimate $s_0^2$ of $\sigma^2$. In the second stage, the test samples up to

$$n_1 = n_0 \vee \left[ \left( t_{2n_0-2,\alpha/2} + t_{2n_0-2,\beta} \right)^2 \frac{2s_0^2}{\delta^2} \right] \tag{8.1}$$

observations from each population, where $\alpha$ is the prescribed type I error probability, and $1-\beta$ is the prescribed power at the alternatives satisfying $|\mu_X - \mu_Y| = \delta$. The null hypothesis $H_0 : \mu_x = \mu_Y$ is then rejected if

$$\frac{|\bar{X}_{n_1} - \bar{Y}_{n_1}|}{\sqrt{2s_0^2/n_1}} > t_{2n_0-2,\alpha/2}. \tag{8.2}$$

Stein (1945) showed that the use of the initial variance estimate $s_0^2$ in the final test statistic (8.2) ensures that the test has type I error probability $\alpha$ and power at least $1-\beta$. However, this feature also diminishes the practical appeal of the test. Denne and Jennison (1999) present a way of incorporating the variance estimate $s_1^2$ based on $2n_1$ observations into the final test statistic while increasing the degrees of freedom of the $t$-distribution so that the desired type I error probability and power constraints are not violated by much. Specifically, they propose the rejection rule

$$\frac{|\bar{X}_{n_1} - \bar{Y}_{n_1}|}{\sqrt{2s_1^2/n_1}} > t_{2n_0-2+2\varepsilon(n_1-n_0),\alpha/2}, \tag{8.3}$$

where $0 < \varepsilon < 1$ is a user-specified parameter. Denne and Jennison (2000) extend this test to a group sequential setting.

Many other modifications of Stein's initial idea have been proposed. Viewing the test statistic

$$\frac{|\bar{X}_{n_1} - \bar{Y}_{n_1}|}{\sqrt{2s_1^2/n_1}}$$

as a fixed-sample statistic based on a sample of size $n_1$ from each population, when $\mu_X - \mu_Y = \delta$, this test statistic has the noncentral $t$-distribution on $2n_1 - 2$ degrees of freedom with noncentrality parameter $\delta\sqrt{n_1/(2s_1^2)}$, which is a random variable if the random size (8.1) is used instead. Fixing $\alpha$, $\beta$, and $\delta$, let $n(\sigma^2)$ denote the smallest $n_1$ for which the probability exceeds $1-\beta$ that an observation from this distribution exceeds the critical value $t_{2n_1-2,\alpha/2}$ in (8.2). Based on a pretrial estimate $\sigma_0^2$ of $\sigma^2$, an estimate of the total desired sample size is $n(\sigma_0^2)$. Following a pilot study of size $n_0$ per arm, which results in the variance estimate $s_0^2$, the total sample size can be re-estimated as $n(s_0^2)$. At this point there are many options for how to proceed. Wittes and Brittain (1990) recommend taking the maximum of $n(\sigma_0^2)$ and $n(s_0^2)$ as the new total sample size, to safeguard against a low value of $s_0^2$ causing an increase in the type I error probability. Gould and Shih (1992) recommend retaining $n(\sigma_0^2)$ unless $n(s_0^2)$ is substantially larger, as well as truncating $n(s_0^2)$ to some practical value, like $2n(\sigma_0^2)$. Birkett and Day (1994) recommend taking the total sample size per treatment as the maximum of $n_0$ and $n(\sigma_0^2)$. Herson and Wittes (1993) propose a procedure, originally for binary data, in which the sample size is updated based on estimates from the control group alone. Shih (2001) and Whitehead et al. (2001) provide useful reviews of these and the other proposals for this problem.

### *8.1.2 Midcourse Sample Size Re-estimation*

From the viewpoint that the trial is intended as fixed sample size but at some intermediate time there may be a desire to re-estimate the total sample size in view of data accumulated so far, Fisher (1998) proposed a method that allows this while maintaining the original type I error rate in the setting where the variance is known, as follows. Without loss of generality, take $\sigma^2 = 1/2$ and let $\theta = \mu_X - \mu_Y$. If $n$ is the original per treatment sample size, then after $rn$ pairs of observations $(0 < r < 1)$, letting

$$S_1 = \sum_{i=1}^{rn} (X_i - Y_i),$$

we have

$$n^{-1/2} S_1 \sim N\left(r\theta\sqrt{n}, r\right).$$

If it is now desired to change the second-stage sample size from $(1-r)n$ to $\gamma(1-r)n$, for some $\gamma > 0$, then letting

$$S_2 = \sum_{i=rn+1}^{n^*} (X_i - Y_i),$$

where $n^* = rn + \gamma(1-r)n$ is the new total per treatment sample size, we have that given the first stage data,

$$(n\gamma)^{-1/2} S_2 \sim N\left((1-r)\theta\sqrt{\gamma n}, 1-r\right). \tag{8.4}$$

Note that under $H_0 : \theta = 0$, (8.4) has the $N(0, 1-r)$ distribution regardless of the (possibly data-dependent) choice of $\gamma$, thus Fisher's (1998) test statistic

$$n^{-1/2}\left(S_1 + \gamma^{-1/2} S_2\right) \tag{8.5}$$

has a $N(0,1)$ distribution under $H_0$. This test has been called the *variance spending test* because the variance $1-r$ of (8.4) is the remaining part of the total variance 1 not spent in the first stage, and thus the factor $\gamma^{-1/2}$ in (8.5) is in place to ensure the standard normal distribution of the test statistic (8.5). Shen and Fisher (1999) gave a multistage version of this procedure based on $S_1, S_2, \ldots, S_k$ in which the sample size update at each stage may be data dependent.

Denne (2001) proposed a test that also allows data-dependent updates of the total sample size but maintains the type I error probability by a seemingly different method. With the preceding notation, Denne's (2001) test chooses a critical value for $S_2$ that is a function of the first-stage data $S_1 = s_1$ by maintaining the conditional type I error rate

$$P_{\theta=0}\left(\frac{S_1 + S_2}{\sqrt{n}} > z_\alpha \,\middle|\, S_1 = s_1\right). \tag{8.6}$$

Jennison and Turnbull (2003) showed that this test, with no stopping after the first stage, is actually equivalent to Fisher's (1998) test: Since $S_2 \sim N(0, (1-r)n)$ under $H_0 : \theta = 0$, a simple calculation shows that (8.6) is equal to

$$A(s_1) = 1 - \Phi\left(\frac{z_\alpha}{\sqrt{(1-r)}} - \frac{s_1}{\sqrt{(1-r)n}}\right),$$

and thus Denne's (2001) rejection rule can be written as

$$\frac{S_2}{\sqrt{\gamma(1-r)n}} > \Phi^{-1}(1 - A(s_1)),$$

which simplifies to

$$n^{-1/2}\left(S_1 + \gamma^{-1/2} S_2\right) > z_\alpha,$$

which is Fisher's test statistic (8.5). Jennison and Turnbull (2003) also showed that the proposal of Cui et al. (1999) is exactly equivalent to Fisher's test, which they found to perform poorly in terms of efficiency and power in comparison to group sequential tests. Tsiatis and Mehta (2003) independently came to the same conclusion, attributing this inefficiency to the use of the non-sufficient "weighted" statistic (8.5).

Working in terms of the $z$-statistic that divides $S$ by its standard deviation, Proschan and Hunsberger (1995) noted that any nondecreasing function $C(z_1)$ with range $[0,1]$ can be used as a conditional type I error function to define a two-stage procedure, as long as it satisfies

$$\int_{-\infty}^{\infty} C(z_1)\phi(z_1)\, dz_1 = \alpha, \tag{8.7}$$

and suggested certain choices of $C$. Having observed the first-stage data $Z_1$, $H_0 : \theta = 0$ is rejected in favor of $\theta > 0$ after stage two if $Z_2 > \Phi^{-1}(1 - C(z_1))$. The condition (8.7) ensures that the type I error probability of any test of this form is $\alpha$. The tests proposed earlier by Bauer and Köhne (1994) can also be represented in this common framework, as noted by Posch and Bauer (1999).

### 8.1.3 Midcourse Modification of the Maximum Sample Size in a Group Sequential Trial

Cui et al. (1999) discussed the issue of increasing the maximum sample size after interim analysis in a group sequential trial. They cited a study protocol, which was reviewed by the Food and Drug Administration, involving a Phase III group sequential trial for evaluating the efficacy of a new drug to prevent myocardial infarction in patients undergoing coronary artery bypass graft surgery. During interim analysis, the observed incidence for the drug achieved a reduction that was only half of the target reduction assumed in the calculation of the maximum sample size $M$, resulting in a proposal to increase the maximum sample size to $\widetilde{M}$ ($N_{\max}$ in their notation). Cui et al. (1999) and Lehmacher and Wassmer (1999) extended the sample size re-estimation approach to adaptive group sequential trials by adjusting the test statistics as in Proschan and Hunsberger (1995) and allowing the future group sizes to be increased or decreased during interim analyses so that the overall sample size does not exceed $\widetilde{M}(> M)$ and the type I error probability is maintained at the prescribed level.

### 8.1.4 *Optimal Adaptive Group Sequential Designs via Dynamic Programming*

Jennison and Turnbull (2006a) introduced adaptive group sequential tests that choose the $j$th group size and stopping boundary on the basis of the cumulative sample size $n_{j-1}$ and the sample sum $S_{n_{j-1}}$ over the first $j-1$ groups and that are optimal in the sense of minimizing a weighted average of the expected sample sizes over a collection of parameter values subject to prescribed error probabilities at the null and a given alternative hypothesis. For example, they give the operating characteristics of the $k$-stage test minimizing

$$[E_0(T)+E_{\theta'}(T)+E_{2\theta'}(T)]/3, \tag{8.8}$$

where $T$ is the total sample size and $\theta'$ a specified alternative, among all $k$-stage tests with given maximum sample size, type I error probability, and power at a given alternative, for $k=2,3,4$. They also showed how the corresponding optimization problem can be solved numerically by using the backward induction algorithms for "optimal sequentially planned" designs developed by Schmitz (1993); this is in fact a special case of finite-horizon dynamic programming introduced in Sect. 3.6.1. Jennison and Turnbull (2006b) found that standard (nonadaptive) group sequential tests with the first stage chosen optimally are nearly as efficient as their optimal adaptive counterparts that are considerably more complicated.

## 8.2 Efficient Adaptive Design and GLR Tests

Instead of staying within the normal family, Bartroff and Lai (2008a,b) consider the more general framework of the multiparameter exponential family $f_{\boldsymbol{\theta}}(x) = \exp(\boldsymbol{\theta}^T x - \psi(\boldsymbol{\theta}))$ considered in Sect. 4.2.4. Let $\Lambda_{i,0}$ denote the right-hand side of (4.14), and let $\Lambda_{i,1}$ denote the right-hand side of (4.14) with $u_0$ replaced by $u_1$ that will be specified below.

### 8.2.1 *An Adaptive 3-Stage GLR Test*

Whereas Tsiatis and Mehta (2003) consider the case of simple null and alternative hypotheses $\boldsymbol{\theta} = \boldsymbol{\theta}_j$ $(j=0,1)$ for which likelihood ratio tests are most powerful even in their group sequential designs, Bartroff and Lai (2008a) use the GLR statistics $\Lambda_{i,0}$ and $\Lambda_{i,1}$ in an adaptive three-stage test of the composite null hypothesis $H_0: u(\boldsymbol{\theta}) \le u_0$, where $u$ is a smooth real-valued function such that

$$I(\boldsymbol{\theta},\boldsymbol{\lambda}) \text{ is increasing in } |u(\boldsymbol{\lambda})-u(\boldsymbol{\theta})| \text{ for every fixed } \boldsymbol{\theta}. \tag{8.9}$$

Let $n_1 = m$ be the sample size of the first stage and $n_3 = M$ be the maximum total sample size, both specified before the trial. Let $u_1 > u_0$ be the alternative implied by the maximum sample size $M$ and the reference type II error probability $\tilde{\alpha}$. That is, $u_1(> u_0)$ is the alternative where the fixed sample size (FSS) GLR test with type I error probability $\alpha$ and sample size $M$ has power $\inf_{\boldsymbol{\theta}:u(\boldsymbol{\theta})=u_1} P_{\boldsymbol{\theta}}\{\text{Reject } H_0\}$ equal to $1-\tilde{\alpha}$, as in Sect. 4.2.4. The three-stage test of $H_0 : u(\boldsymbol{\theta}) \le u_0$ stops and rejects $H_0$ at stage $i \le 2$ if

$$n_i < M, \quad u\left(\hat{\boldsymbol{\theta}}_{n_i}\right) > u_0, \quad \text{and} \quad \Lambda_{i,0} \ge b. \tag{8.10}$$

Early stopping for futility (accepting $H_0$) can also occur at stage $i \le 2$ if

$$n_i < M, \quad u\left(\hat{\boldsymbol{\theta}}_{n_i}\right) < u_1, \quad \text{and} \quad \Lambda_{i,1} \ge \tilde{b}. \tag{8.11}$$

The test rejects $H_0$ at stage $i = 2$ or 3 if

$$n_i = M, \quad u\left(\hat{\boldsymbol{\theta}}_M\right) > u_0, \quad \text{and} \quad \Lambda_{i,0} \ge c, \tag{8.12}$$

accepting $H_0$ otherwise. The sample size $n_2$ of the three-stage test is given by

$$n_2 = m \vee \left\{M \wedge \left\lceil (1+\rho_m)\, n\left(\hat{\boldsymbol{\theta}}_m\right)\right\rceil\right\}, \tag{8.13}$$

with

$$n(\boldsymbol{\theta}) = \min\left\{|\log \alpha| \Big/ \inf_{\boldsymbol{\lambda}:u(\boldsymbol{\lambda})=u_0} I(\boldsymbol{\theta},\boldsymbol{\lambda}),\ |\log \tilde{\alpha}| \Big/ \inf_{\boldsymbol{\lambda}:u(\boldsymbol{\lambda})=u_1} I(\boldsymbol{\theta},\boldsymbol{\lambda})\right\}, \tag{8.14}$$

where $I(\boldsymbol{\theta},\boldsymbol{\lambda})$ is the Kullback–Leibler information number and $\rho_m > 0$ is an inflation factor to adjust for uncertainty in $\widehat{\theta}_m$; see the examples in Sect. 8.3. Note that (8.14) is an asymptotic approximation to Hoeffding's lower bound (3.15). Letting $0 < \varepsilon, \tilde{\varepsilon} < 1$, define the thresholds $b, \tilde{b}$, and $c$ to satisfy the equations

$$\sup_{\boldsymbol{\theta}:u(\boldsymbol{\theta})=u_1} P_{\boldsymbol{\theta}}\{(8.11) \text{ occurs for } i = 1 \text{ or } 2\} = \tilde{\varepsilon}\tilde{\alpha}, \tag{8.15}$$

$$\sup_{\boldsymbol{\theta}:u(\boldsymbol{\theta})=u_0} P_{\boldsymbol{\theta}}\{(8.11) \text{ does not occur for } i \le 2,\ (8.10) \text{ occurs for } i = 1 \text{ or } 2\} = \varepsilon\alpha, \tag{8.16}$$

$$\sup_{\boldsymbol{\theta}:u(\boldsymbol{\theta})=u_0} P_{\boldsymbol{\theta}}\{(8.10) \text{ and } (8.11) \text{ do not occur for } i \le 2,\ (8.12) \text{ occurs}\} = (1-\varepsilon)\alpha. \tag{8.17}$$

The probabilities in (8.15)–(8.17) can be computed by using the normal approximation to the signed-root likelihood ratio statistic

$$\ell_{i,j} = \left\{\text{sign}\left(u(\hat{\boldsymbol{\theta}}_{n_i}) - u_j\right)\right\}(2n_i\Lambda_{i,j})^{1/2}$$

$(1 \le i \le 3$ and $j = 0,1)$ under $u(\boldsymbol{\theta}) = u_j$, as in Sect. 4.3.1. When $u(\boldsymbol{\theta}) = u_j$, $\ell_{i,j}$ is approximately normal with mean 0, variance $n_i$, and the increments $\ell_{i,j} - \ell_{i-1,j}$ are asymptotically independent. We can therefore approximate $\ell_{i,j}$ by a sum of independent standard normal random variables under $u(\theta) = u_j$ and thereby determine $b, \tilde{b}$, and $c$. Note that this normal approximation can also be used for the choice of $u_1$ implied by $M$ and $\tilde{\alpha}$. Computational details are given in Sect. 8.2.4.

A special multiparameter case of particular interest in clinical trials involves $K$-independent populations having density functions $\exp\{\theta_k x - \tilde{\psi}_k(\theta_k)\}$ so that $\boldsymbol{\theta}^T \boldsymbol{x} - \psi(\boldsymbol{\theta}) = \sum_{k=1}^K \{\theta_k x_k - \tilde{\psi}(\theta_k)\}$. In multiarmed trials, for which different numbers of patients are assigned to different treatments, the GLR statistic $\Lambda_{i,j}$ for testing the hypothesis $u(\theta_1, \dots, \theta_K) = u_j$ $(j = 0$ or $1)$ at stage $i$ has the form

$$\Lambda_{i,j} = \sum_{k=1}^K n_{ki} \left\{ \hat{\theta}_{k,n_{ki}} \bar{X}_{k,n_{ki}} - \tilde{\psi}\left(\hat{\theta}_{k,n_{ki}}\right) \right\} - \sup_{\boldsymbol{\theta}: u(\theta_1,\dots,\theta_K) = u_j} \sum_{k=1}^K n_{ki} \left\{ \theta_k \bar{X}_{k,n_{ki}} - \tilde{\psi}(\theta_k) \right\},$$

in which $n_{ki}$ is the total number of observations from the $k$th population up to stage $i$. Letting $n_i = \sum_{k=1}^K n_{ki}$, the normal approximation to the signed-root likelihood ratio statistic is still applicable when $n_{ki} = p_k n_i + O_p(n_i^{1/2})$, where $p_1, \dots, p_K$ are nonnegative constants that sum up to 1, as in random allocation of patients to the $K$ treatments (for which $p_k = 1/K$); see Lai and Shih (2004, p. 514).

### 8.2.2 Midcourse Modification of Maximum Sample Size

We now modify the adaptive designs in the preceding section to accommodate the possibility of midcourse increase of the maximum sample size from $M$ to $\widetilde{M}$. Let $u_2$ be the alternative implied by $\widetilde{M}$ so that the level-$\alpha$ GLR test with sample size $\widetilde{M}$ has power $1 - \tilde{\alpha}$. Note that $u_1 > u_2 > u_0$. Whereas the sample size $n_3$ is chosen to be $M$ in Sect. 8.2.1, we now define

$$\tilde{n}(\boldsymbol{\theta}) = \min\left\{ |\log \alpha| \Big/ \inf_{\boldsymbol{\lambda}: u(\boldsymbol{\lambda}) = u_0} I(\boldsymbol{\theta}, \boldsymbol{\lambda}),\ |\log \tilde{\alpha}| \Big/ \inf_{\boldsymbol{\lambda}: u(\boldsymbol{\lambda}) = u_2} I(\boldsymbol{\theta}, \boldsymbol{\lambda}) \right\},$$

$$n_3 = n_2 \vee \left\{ M' \wedge \left\lceil (1 + \rho_m) \tilde{n}\left(\hat{\boldsymbol{\theta}}_{n_2}\right) \right\rceil \right\},$$

where $M < M' \le \widetilde{M}$ and $n_2 = m \vee \{M \wedge (1 + \rho_m) \tilde{n}(\hat{\boldsymbol{\theta}}_m)\}$. We can regard the test as a group sequential test with four groups and $n_1 = m$, $n_4 = \widetilde{M}$, but with adaptively chosen $n_2$ and $n_3$. If the test does not end at the third stage, continue to the fourth and final stage with sample size $n_4 = \widetilde{M}$. Its rejection and futility boundaries are similar to those in Sect. 8.2.1. Extending our notation $\Lambda_{i,j}$ to $1 \le i \le 4$ and $0 \le j \le 2$, the test stops at stage $i \le 3$ and rejects $H_0$ if

$$n_i < \widetilde{M}, \quad u\left(\hat{\boldsymbol{\theta}}_{n_i}\right) > u_0, \quad \text{and} \quad \Lambda_{i,0} \ge b, \tag{8.18}$$

stops and accepts $H_0$ if

$$n_i < \widetilde{M}, \quad u(\hat{\boldsymbol{\theta}}_{n_i}) < u_2, \quad \text{and} \quad \Lambda_{i,2} \geq \tilde{b}, \tag{8.19}$$

and rejects $H_0$ at stage $i = 3$ or 4 if

$$n_i = \widetilde{M}, \quad u(\hat{\boldsymbol{\theta}}_{\widetilde{M}}) > u_0, \quad \text{and} \quad \Lambda_{i,0} \geq c, \tag{8.20}$$

accepting $H_0$ otherwise. The thresholds $b, \tilde{b}$, and $c$ can be defined by equations similar to (8.15)–(8.17) to insure the overall type I error probability to be $\alpha$. For example, in place of (8.15),

$$\sup_{\boldsymbol{\theta}: u(\boldsymbol{\theta}) = u_2} P_\theta\{(8.19) \text{ occurs for some } i \leq 3\} = \tilde{\varepsilon}\tilde{\alpha}. \tag{8.21}$$

The basic idea underlying (8.21) is to control the type II error probability at $u_2$ so that the test does not lose much power there in comparison with the GLR test that has sample size $\widetilde{M}$ (and therefore power $1 - \tilde{\alpha}$ at $u_2$).

### 8.2.3 *Asymptotic Theory*

For testing one-sided hypotheses $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta \geq \theta_1$ in a one-parameter exponential family, the idea of using a three-stage test to achieve the first-order asymptotic efficiency of Schwarz's fully sequential GLR test with stopping rule of the form (3.12) dated back to Lorden (1983). To prove asymptotic optimality as $\alpha + \tilde{\alpha} \to 0$, Lorden used crude upper bounds for the error probabilities that make the thresholds $A_0^{(n)}$ and $A_1^{(n)}$ too conservative for practical applications but suffice for the expected sample size to attain asymptotically Hoeffding's lower bound (3.15) at every fixed $\theta$. Bartroff and Lai (2008a,b) have overcome the practical difficulty due to the overly conservative stopping thresholds of Lorden's three-stage test by using a Haybittle–Peto-type boundary and developing numerical methods to compute the type I error probability as in Sect. 4.2 and have also extended the three-stage test to multiparameter and multiarmed settings, thus greatly broadening the scope of these adaptive designs. They establish the asymptotic optimality of the three-stage test in the following.

**Theorem 8.1.** *Let $N$ denote the sample size of the three-stage GLR test in Sect. 8.2.1, with $m$, $M$, and $m \vee [M \wedge \lceil (1+\rho_m) n(\hat{\boldsymbol{\theta}}_m) \rceil]$ being the possible values of $N$. Let $T$ be the sample size of any test of $H_0 : u(\boldsymbol{\theta}) \leq u_0$ versus $H_1 : u(\boldsymbol{\theta}) \geq u_1$, sequential or otherwise, which takes at least $m$ and at most $M$ observations and whose type I and type II error probabilities do not exceed $\alpha$ and $\tilde{\alpha}$, respectively. Assume that* $\log \alpha \sim \log \tilde{\alpha}$,

$$m/|\log \alpha| \to a, \quad M/|\log \alpha| \to A, \quad \rho_m \to 0 \ \textit{but} \ m^{1/2}\rho_m/(\log m)^{1/2} \to \infty \tag{8.22}$$

*as* $\alpha+\tilde{\alpha}\to 0$*, with* $0<a<A$*. Then for every fixed* $\boldsymbol{\theta}$*, as* $\alpha+\tilde{\alpha}\to 0$*,*

$$E_{\boldsymbol{\theta}}(N)\sim m\vee\left\{M\wedge|\log\alpha|/\left[\inf_{\boldsymbol{\lambda}:u(\boldsymbol{\lambda})=u_0}I(\boldsymbol{\theta},\boldsymbol{\lambda})\vee\inf_{\boldsymbol{\lambda}:u(\boldsymbol{\lambda})=u_1}I(\boldsymbol{\theta},\boldsymbol{\lambda})\right]\right\},\tag{8.23}$$

$$E_{\boldsymbol{\theta}}(T)\geq[1+o(1)]E_{\boldsymbol{\theta}}(N).\tag{8.24}$$

Since $m\sim a|\log\alpha|$ and $M\sim A|\log\alpha|$ and since the thresholds $b$, $\tilde{b}$, and $c$ are defined by (8.15)–(8.17), Bartroff and Lai (2008b) use an argument similar to the proof of Theorem 2(ii) of Lai and Shih (2004, p. 525) to show that (8.23) holds. They then use the following argument to prove (8.24). Let $\Theta_0=\{\boldsymbol{\theta}:u(\boldsymbol{\theta})\leq u_0\}$, $\Theta_1=\{\boldsymbol{\theta}:u(\boldsymbol{\theta})\geq u_1\}$. For $i=0,1$,

$$\inf_{\boldsymbol{\lambda}\in\Theta_i}I(\boldsymbol{\theta},\boldsymbol{\lambda})=I_i(\boldsymbol{\theta}),\quad\text{where } I_i(\boldsymbol{\theta})=\inf_{\boldsymbol{\lambda}:u(\boldsymbol{\lambda})=u_i}I(\boldsymbol{\theta},\boldsymbol{\lambda}).\tag{8.25}$$

Take any $\boldsymbol{\lambda}\in\Theta_0$ and $\tilde{\boldsymbol{\lambda}}\in\Theta_1$. From (8.25) and Hoeffding's lower bound (3.15), it follows that for a test that has error probabilities $\alpha$ and $\tilde{\alpha}$ at $\boldsymbol{\lambda}$ and $\tilde{\boldsymbol{\lambda}}$ and take at least $m$ and at most $M$ observations, its sample size $T$ satisfies

$$E_{\boldsymbol{\theta}}(T)\geq m\vee\left\{M\wedge\frac{[1+o(1)]|\log\alpha|}{I_0(\boldsymbol{\theta})\vee I_1(\boldsymbol{\theta})}\right\}\tag{8.26}$$

as $\alpha+\tilde{\alpha}\to 0$ such that $\log\alpha\sim\log\tilde{\alpha}$. The second-stage sample size of the adaptive test is a slight inflation of the Hoeffding-type lower bound (8.26) with $\boldsymbol{\theta}$ replaced by the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_m$ at the end of the first stage. The assumption $\rho_m\to 0$ but $\rho_m\succ m^{-1/2}(\log m)^{1/2}$ is used to accommodate the difference between $\boldsymbol{\theta}$ and its substitute $\hat{\boldsymbol{\theta}}_m$, which satisfies $P_{\boldsymbol{\theta}}\{\sqrt{m}\|\hat{\boldsymbol{\theta}}_m-\boldsymbol{\theta}\|\geq r(\log m)^{1/2}\}=o(m^{-1})$ if $m$ is sufficiently large, by standard exponential bounds involving moment generating functions.

As noted by Bartroff and Lai (2008b), the adaptive test in Sect. 8.2.2 can be regarded as a midcourse amendment of an adaptive test of $H_0:u(\boldsymbol{\theta})\leq u_0$ versus $H_1:u(\boldsymbol{\theta})\geq u_1$, with a maximum sample size of $M$, to that of $H_0$ versus $H_2:u(\boldsymbol{\theta})\geq u_2$, with a maximum sample size of $\widetilde{M}$. Whereas (8.26) provides an asymptotic lower bound for tests of $H_0$ versus $H_1$, any test of $H_0$ versus $H_2$ with error probabilities not exceeding $\alpha$ and $\tilde{\alpha}$ and taking at least $m$ and at most $\widetilde{M}$ observations likewise satisfies

$$E_{\boldsymbol{\theta}}(T)\geq m\vee\left\{\widetilde{M}\wedge\frac{[1+o(1)]|\log\alpha|}{I_0(\boldsymbol{\theta})\vee I_2(\boldsymbol{\theta})}\right\}\tag{8.27}$$

as $\alpha+\tilde{\alpha}\to 0$ such that $\log\alpha\sim\log\tilde{\alpha}$. Note that $\Theta_1=\{\boldsymbol{\theta}:u(\boldsymbol{\theta})\geq u_1\}\subset\Theta_2=\{\boldsymbol{\theta}:u(\boldsymbol{\theta})\geq u_2\}$ and therefore $I_2(\boldsymbol{\theta})\leq I_1(\boldsymbol{\theta})$. The four-stage test in Sect. 8.2.2, with $M'=\widetilde{M}$, attempts to attain the asymptotic lower bound in (8.26) prior to the third stage and the asymptotic lower bound in (8.27) afterwards. It replaces $I_1(\boldsymbol{\theta})$ in (8.26), which corresponds to early stopping for futility, by $I_2(\boldsymbol{\theta})$ that corresponds to rejection of $H_2$ (instead of $H_1$) in favor of $H_0$. Thus, the second-stage sample size $n_2$

corresponds to the lower bound in (8.26) with $\boldsymbol{\theta}$ replaced by $\hat{\boldsymbol{\theta}}_m$ and $I_1$ replaced by $I_2$, while the third-stage sample size corresponds to that in (8.27) with $\boldsymbol{\theta}$ replaced by $\hat{\boldsymbol{\theta}}_{n_2}$. The arguments used to prove the asymptotic optimality of the three-stage test in Theorem 8.1 can be readily modified to prove the following.

**Theorem 8.2.** *Let $N^*$ denote the sample size of the four-stage GLR test in Sect. 8.2.2, with $M' = \widetilde{M}$. Assume that $\log\alpha \sim \log\tilde{\alpha}$ as $\alpha + \tilde{\alpha} \to 0$, that (8.22) holds, and that $\widetilde{M}/|\log\alpha| \to \tilde{A}$ with $0 < a < A < \tilde{A}$. Then*

$$E_{\boldsymbol{\theta}}(N^*) \sim \begin{cases} m \vee [1+o(1)]|\log\alpha|/I_0(\boldsymbol{\theta}) & \textit{if } I_0(\boldsymbol{\theta}) > A^{-1}, \\ m \vee \left\{\widetilde{M} \wedge [1+o(1)]|\log\alpha|/[I_0(\boldsymbol{\theta}) \wedge I_2(\boldsymbol{\theta})]\right\} & \textit{if } I_0(\boldsymbol{\theta}) < A^{-1}. \end{cases}$$

### 8.2.4 Implementation via Normal Approximation or Monte Carlo

To begin with, suppose the $X_i$ are $N(\theta,1)$ and $u(\theta) = \theta$. We write $\theta_j$ instead of $u_j$, and, without loss of generality, we shall assume that $\theta_0 = 0$. The thresholds $b$, $\tilde{b}$, and $c$ of the three-stage test in Sect. 8.2.1 can be computed by solving in succession (8.15), (8.16), and (8.17). Univariate grid search or Brent's method (Press et al., 1992) can be used to solve each equation. Since $I(\theta,\lambda) = (\theta-\lambda)^2/2$, we can rewrite (8.15) as

$$\begin{aligned} &P_{\theta_1}\left\{S_m - m\theta_1 > -\left(2\tilde{b}m\right)^{1/2},\ S_{n_2} - n_2\theta_1 \le -\left(2\tilde{b}n_2\right)^{1/2}\right\} \\ &\qquad + P_{\theta_1}\left\{S_m - m\theta_1 \le -\left(2\tilde{b}m\right)^{1/2}\right\} = \tilde{\varepsilon}\tilde{\alpha}, \end{aligned} \tag{8.28}$$

and (8.16) and (8.17) as

$$\begin{aligned} &P_0\left\{S_m/(2m)^{1/2} \ge b^{1/2}\right\} \\ &\quad + P_0\left\{S_m - m\theta_1 > -(2\tilde{b}m)^{1/2}, S_m/(2m)^{1/2} < b^{1/2},\ S_{n_2}/(2n_2)^{1/2} \ge b^{1/2},\ n_2 < M\right\} = \varepsilon\alpha, \end{aligned} \tag{8.29}$$

$$\begin{aligned} &P_0\Big\{S_m - m\theta_1 > -(2\tilde{b}m)^{1/2}, S_m/(2m)^{1/2} < b^{1/2},\ n_2 < M, \\ &\quad S_{n_2} - n_2\theta_1 > -(2\tilde{b}n_2)^{1/2}, S_{n_2}/(2n_2)^{1/2} < b^{1/2},\ S_M/(2M)^{1/2} \ge c^{1/2}\Big\} \\ &\quad + P_0\left\{S_m - m\theta_1 > -(2\tilde{b}m)^{1/2}, S_m/(2m)^{1/2} < b^{1/2},\ n_2 = M,\ S_M/(2M)^{1/2} \ge c^{1/2}\right\} = (1-\varepsilon)\alpha. \end{aligned} \tag{8.30}$$

The probabilities involving $n_2$ can be computed by conditioning on the value of $S_m/m$, which completely determines the value of $n_2$, denoted by $k(x)$. For example, the probabilities under $\theta = 0$ can be computed via

$$\begin{aligned} &P_0\left\{S_{n_2} \geq (2bn_2)^{1/2} \,|\, S_m = mx\right\} \\ &\qquad = P\left\{N(0,1) \geq \left[(2bn_2)^{1/2} - mx\right] \Big/ [k(x)-m]^{1/2}\right\}, \qquad (8.31) \\ &P_0\left\{S_{n_2} \in dy, S_M \in dz \,|\, S_m = mx\right\} \\ &\qquad = \varphi_{k(x)-m}(y-mx)\varphi_{M-k(x)}(z-y)\,dy\,dz, \qquad (8.32) \end{aligned}$$

where $\varphi_v$ is the $N(0,v)$ density function, that is, $\varphi_v(w) = (2\pi v)^{-\frac{1}{2}}\exp(-w^2/2v)$. The probabilities under $\theta_1$ can be computed similarly. Hence standard recursive numerical integration algorithms can be used to compute the probabilities in (8.15)–(8.17); see Sect. 4.3.1. More generally, for the general multiparameter exponential family, this method can be used to compute the thresholds $b, \tilde{b}$, and $c$ for (8.10)–(8.12) since the problem can be approximated by that of testing a normal mean, as discussed in Sect. 8.2.1.

For midcourse modification of the maximum sample size in Sect. 8.2.2, the preceding recursive numerical algorithm can be modified to handle the randomness of $n_2$ and $n_3$. The basic idea is that conditional on $S_m/m = x$, the value of $n_2$ is completely determined as $k(x)$ and conditional on $S_m/m = x$ and $S_{n_2}/n_2 = y$, the value of $n_3$ is completely determined as $h(x,y)$. Therefore, analogous to (8.32), we now have

$$\begin{aligned} &P\{S_{n_3} \in du, S_{\widetilde{M}} \in dw \,|\, S_m/m = x, S_{n_2}/n_2 = y\} \\ &\qquad = \varphi_{h(x,y)-k(x)}(u - yk(x))\varphi_{\widetilde{M}-h(x,y)}(w-u)\,du\,dw \qquad (8.33) \end{aligned}$$

and can use bivariate recursive numerical integration. For the general exponential family, normal approximation to the signed-root likelihood ratio statistic can again be used.

An alternative to normal approximation is to use Monte Carlo similar to that used in bootstrap tests. As noted in Sect. 4.3.2, bootstrap theory suggests that we can simulate from the estimated distribution under the assumed hypothesis as the GLR statistic is an approximate pivot under that hypothesis. Since the "estimated distribution" needs data to arrive at the estimate, we make use of the first-stage data to determine $b$ and $\tilde{b}$ and then use the second-stage data to determine $c$ for the three-stage test in Sect. 8.2.1. Specifically, the Monte Carlo method to determine $b, \tilde{b}$, and $c$ proceeds as follows. At the end of the first stage, compute the maximum likelihood estimate $\hat{\theta}_{m,j}$ under the constraint $u(\theta) = u_j$, $j = 0, 1$. Determine $\tilde{b}, b$, and $c$ successively by solving

$$P_{\hat{\theta}_{m,1}}\{(8.11) \text{ occurs for } i = 1 \text{ or } 2\} = \tilde{\varepsilon}\tilde{\alpha}, \tag{8.34}$$

$$P_{\hat{\theta}_{m,0}}\{(8.11) \text{ does not occur for } i \le 2,\ (8.10) \text{ occurs for } i = 1 \text{ or } 2\} = \varepsilon\alpha, \tag{8.35}$$

$$P_{\hat{\theta}_{n_2,0}}\{(8.10)\text{–}(8.11) \text{ do not occur for } i \le 2,\ (8.12) \text{ occurs}\} = (1-\varepsilon)\alpha, \tag{8.36}$$

noting that $c$ does not have to be determined until after the second stage when $n_2$ observations become available for the updated estimate $\hat{\theta}_{n_2,0}$. The probabilities in (8.34)–(8.36) can be computed by Monte Carlo simulations, similarly, to determine thresholds $b, \tilde{b}$, and $c$ of the adaptive test in Sect. 8.2.2.

## 8.3 Comparison of Adaptive Designs

Bartroff and Lai (2008a,b) carried out comprehensive simulation studies of the performance, measured in terms of the expected sample size and power functions, of the adaptive tests in Sect. 8.2 and compared them with those in Sect. 8.1. In the case of normal mean with known variance and type I and II error constraints under the null and a given alternative hypothesis, they showed that the adaptive test in Sect. 8.2.2 is comparable to the benchmark optimal adaptive test of Jennison and Turnbull (2006a,b), which is superior to the existing two-stage adaptive designs. On the other hand, whereas the benchmark optimal adaptive test needs to assume a specified alternative, these adaptive two-stage tests and the adaptive tests in Sect. 8.2 do not require such assumptions as they consider the estimated alternative at the end of the first stage. In their recent survey of adaptive designs, Burman and Sonesson (2006) pointed out that previous criticisms of the statistical properties of two-stage adaptive designs may be unconvincing in some situations when flexibility and not having to specify parameters that are unknown at the beginning of a trial (like the relevant treatment effect or variance) are more imperative than efficiency or being powerful. The adaptive designs in Sect. 8.2 can therefore fulfill the seemingly disparate requirements of flexibility and efficiency on a design. Rather than achieving exact optimality at a specified collection of alternatives through dynamic programming, they achieve asymptotic optimality over the entire range of alternatives, resulting in near optimality in practice. They are based on efficient test statistics of the GLR type, which have an intuitively "adaptive" appeal via estimation of unknown parameters by maximum likelihood, ease of implementation, and freedom from having to specify the relevant alternative.

### *8.3.1 Normal Mean with Known Variance*

We consider the special case of normal $X_i$ with unknown mean $\theta$ and known variance 1 and compare a variety of adaptive tests of $H_0 : \theta \le 0$ in the literature with the tests proposed in Sect. 8.2. In this normal setting, $\hat{\theta}_n = \bar{X}_n$ and

$I(\theta,\lambda) = (\theta-\lambda)^2/2$. It is widely recognized that the performance of adaptive tests is difficult to evaluate and compare because it depends heavily on the choice of first-stage and maximum sample sizes, the number of groups (stages) allowed, and the parameter values at which the tests are evaluated. For this reason, the tests evaluated here use the same first-stage and maximum sample sizes, except for a few illustrative examples discussed below. In addition, we report a variety of operating characteristics for each test—power, mean number of stages, and the 25th, 50th, and 75th percentiles in addition to the mean of the sample size distribution—over a wide range of $\theta$ values. We also include the uniformly most powerful FSS test with the same maximum sample size and type I error probability $\alpha$, which provides the appropriate benchmark for the power of any test of $H_0$. Another relevant comparison—especially given their widespread use in clinical trials—made here is with standard (nonadaptive) group sequential tests having a similar number of stages as the adaptive test.

To test $H_0 : \theta \le 0$, Proschan and Hunsberger (1995) proposed a two-stage test, based on the conditional power criterion, which uses the usual $z$-statistic but with a data-dependent critical value to maintain the type I error at a prescribed level $\alpha$; see the last paragraph of Sect. 8.1.1. The test allows early stopping to accept (or reject) the null hypothesis if the test statistic is below a user-specified upper normal quantile $z_{p^*}$ (or above some level $k$) at the end of the first stage. Choosing a data-dependent critical value is tantamount to multiplying the $z$-statistic by a data-dependent factor and using a fixed critical value. Li and Shih (2002) proposed to use the $z$-statistic with a fixed critical value $c$ while still determining the second-stage sample size by conditional power and maintaining the type I error at $\alpha$. Their test stops after the first stage if the test statistic falls below $h$ or above $k$. For each $h$ and conditional power level, their test has a maximum allowable $k$, which they denote by $k_1^*(h)$. Fisher (1998) proposed a "variance spending" method for weighting the observations so that the type I error of his test does not exceed $\alpha$. To avoid a very large second-stage sample size if the first-stage estimate of $\theta$ lies near the null hypothesis, Shen and Fisher (1999) proposed early stopping due to futility whenever the upper $100(1-\alpha_0)\%$ confidence bound for $\theta$ falls below some specified alternative $\theta_1 > 0$.

Table 8.1 compares these tests, a FSS test, and two standard group sequential tests with the adaptive test described in Sect. 8.2. The values of the user-specified parameters of the tests are summarized in the list below. The user-specified parameters are chosen so that they have the same first-stage sample size $m = 40$ (except for the FSS test), maximum sample size $M = 120$ (except for SF$'$; see the last paragraph of this section), type I error not exceeding $\alpha = .025$, and nominal power (or conditional power level in the case of conditional power tests) equal to 0.9.

- ADAPT: The adaptive test described in Sect. 8.2.1 that uses $b = 3.26$, $\tilde{b} = 1.99$, and $c = 2.05$ corresponding to $\varepsilon = \tilde{\varepsilon} = 1/3$ in (8.15)–(8.17) and $\rho_m = 0.1$ (see Sect. 8.2.4 for details).
- FSS$_{120}$: The FSS test having sample size 120.

**Table 8.1** Power (italics), expected sample size (italics), sample size quantiles $T_q$, expected number of stages (bold), and efficiency ratio (at $\theta > 0$) with respect to ADAPT, of FSS, adaptive, and group sequential tests with maximum sample size $M = 120$ except for SF$'$ that uses $5M$

| Test | ADAPT | FSS$_{120}$ | OBF$_{PF}$ | OBF$_{SC}$ | PH | L | SF | SF$'$ |
|---|---|---|---|---|---|---|---|---|
| $\theta = -.03$ | *1.1%* | *1.0%* | *0.9%* | *1.0%* | *1.5%* | *1.3%* | *0.6%* | *0.9%* |
| | *68.5* | *120.0* | *72.3* | *91.7* | *40.8* | *41.4* | *41.3* | *72.3* |
| $T_{.25}$ | 40 | 120 | 40 | 80 | 40 | 40 | 40 | 40 |
| $T_{.5}$ | 40 | 120 | 80 | 80 | 40 | 40 | 40 | 40 |
| $T_{.75}$ | 120 | 120 | 80 | 120 | 40 | 40 | 40 | 40 |
| # | **1.53** | **1.00** | **1.81** | **2.29** | **1.01** | **1.03** | **1.03** | **1.14** |
| $\theta = 0$ | *2.5%* | *2.5%* | *2.3%* | *2.5%* | *2.4%* | *2.5%* | *1.2%* | *2.2%* |
| | *75.1* | *120.0* | *77.8* | *96.4* | *41.1* | *42.2* | *41.2* | *82.3* |
| $T_{.25}$ | 40 | 120 | 80 | 80 | 40 | 40 | 40 | 40 |
| $T_{.5}$ | 60 | 120 | 80 | 80 | 40 | 40 | 40 | 40 |
| $T_{.75}$ | 120 | 120 | 80 | 120 | 40 | 40 | 40 | 40 |
| # | **1.64** | **1.00** | **1.94** | **2.41** | **1.02** | **1.05** | **1.05** | **1.20** |
| $\theta = .15$ | *35.6%* | *37.6%* | *35.7%* | *37.1%* | *18.7%* | *20.9%* | *13.8%* | *36.1%* |
| | *98.6* | *120.0* | *98.9* | *110.2* | *44.5* | *48.3* | *47.2* | *115.3* |
| $T_{.25}$ | 71 | 120 | 80 | 120 | 40 | 40 | 40 | 40 |
| $T_{.5}$ | 120 | 120 | 120 | 120 | 40 | 40 | 40 | 47 |
| $T_{.75}$ | 120 | 120 | 120 | 120 | 40 | 40 | 40 | 146 |
| # | **2.05** | **1.00** | **2.47** | **2.76** | **1.09** | **1.22** | **1.22** | **1.53** |
| $R_\theta(T,N)$ | 100 | 78.5 | 99.5 | 86.4 | 332 | 289 | 358 | 84.5 |
| $\theta = .20$ | *57.2%* | *60.0%* | *57.9%* | *59.5%* | *30.2%* | *33.2%* | *24.8%* | *53.5%* |
| | *99.4* | *120.0* | *101.4* | *108.0* | *45.9* | *50.8* | *50.6* | *124.3* |
| $T_{.25}$ | 76 | 120 | 80 | 80 | 40 | 40 | 40 | 40 |
| $T_{.5}$ | 120 | 120 | 120 | 120 | 40 | 40 | 40 | 65 |
| $T_{.75}$ | 120 | 120 | 120 | 120 | 40 | 51 | 52 | 157 |
| # | **2.07** | **1.00** | **2.54** | **2.70** | **1.11** | **1.30** | **1.86** | **1.66** |
| $R_\theta(T,N)$ | 100 | 88.5 | 99.7 | 97.2 | 98.1 | 99.3 | 70.1 | 73.1 |
| $\theta = .26$ | *77.4%* | *80.0%* | *78.0%* | *79.6%* | *44.2%* | *47.5%* | *38.2%* | *67.5%* |
| | *95.2* | *120.0* | *99.8* | *102.0* | *46.6* | *52.7* | *52.9* | *120.2* |
| $T_{.25}$ | 59 | 120 | 80 | 80 | 40 | 40 | 40 | 41 |
| $T_{.5}$ | 120 | 120 | 80 | 120 | 40 | 40 | 40 | 68 |
| $T_{.75}$ | 120 | 120 | 120 | 120 | 40 | 60 | 60 | 145 |
| # | **2.00** | **1.00** | **2.47** | **2.55** | **1.13** | **1.38** | **1.47** | **1.78** |
| $R_\theta(T,N)$ | 100 | 84.7 | 96.8 | 98.6 | 91.4 | 88.4 | 67.4 | 62.7 |
| $\theta = \theta_1 = .3$ | *88.8%* | *90.0%* | *88.6%* | *89.5%* | *55.2%* | *58.0%* | *49.1%* | *75.5%* |
| | *89.2* | *120.0* | *94.5* | *96.4* | *46.8* | *53.3* | *54.0* | *111.8* |
| $T_{.25}$ | 40 | 120 | 80 | 80 | 40 | 40 | 40 | 43 |
| $T_{.5}$ | 118 | 120 | 80 | 80 | 40 | 40 | 42 | 65 |
| $T_{.75}$ | 120 | 120 | 120 | 120 | 40 | 62 | 62 | 128 |
| # | **1.91** | **1.00** | **2.36** | **2.41** | **1.14** | **1.42** | **1.58** | **1.85** |
| $R_\theta(T,N)$ | 100 | 77.5 | 93.8 | 94.7 | 82.6 | 77.5 | 61.5 | 55.6 |

(continued)

**Table 8.1** (continued)

| Test | ADAPT | $\text{FSS}_{120}$ | $\text{OBF}_{PF}$ | $\text{OBF}_{SC}$ | PH | L | SF | SF′ |
|---|---|---|---|---|---|---|---|---|
| $\theta = .33$ | *94.0%* | *95.0%* | *94.1%* | *94.7%* | *63.5%* | *66.5%* | *57.3%* | *80.8%* |
| | *83.0* | *120.0* | *90.1* | *91.1* | *46.7* | *53.3* | *54.3* | *103.5* |
| $T_{.25}$ | 40 | 120 | 80 | 80 | 40 | 40 | 40 | 43 |
| $T_{.5}$ | 89 | 120 | 80 | 80 | 40 | 40 | 44 | 61 |
| $T_{.75}$ | 120 | 120 | 120 | 120 | 40 | 62 | 62 | 114 |
| # | **1.81** | **1.00** | **2.25** | **2.28** | **1.13** | **1.44** | **1.65** | **1.89** |
| $R_\theta(T,N)$ | 100 | 72.8 | 92.6 | 94.3 | 76.4 | 71.8 | 56.9 | 52.0 |

- $\text{OBF}_{\text{PF}}$, $\text{OBF}_{\text{SC}}$: O'Brien and Fleming's (1979) one-sided group sequential tests having three groups of size 40. $\text{OBF}_{\text{PF}}$ uses power family futility stopping ($\Delta = 1$ in Jennison and Turnbull, 2000, Sect. 4.2), and $\text{OBF}_{\text{SC}}$ uses stochastic curtailment futility stopping ($\gamma = 0.9$ in Jennison and Turnbull, 2000, Sect. 10.2). Both $\text{OBF}_{\text{PF}}$ and $\text{OBF}_{\text{SC}}$ use reference alternative $\theta_1 = 0.3$; see below.
- PH: Proschan and Hunsberger's (1995) test that uses $p^* = 0.0436$ and $k = 2.05$.
- L: Li and Shih's (2002) test that uses $h = 1.63$ and $k = k_1^*(h) = 2.83$.
- SF, SF′: Two versions of Shen and Fisher's (1999) test; SF uses $\alpha_0 = 0.425$ and SF′ uses $\alpha_0 = 0.154$.

The tests are evaluated at the $\theta$ values where $\text{FSS}_{120}$ has powers 0.01, 0.025, 0.6, 0.8, 0.9, and 0.95, and at $\theta = 0.15$, the midpoint of $\theta = 0$ and $\theta = \theta_1 = 0.3$, the alternative implied by $M = 120$ since $\text{FSS}_{120}$ has power $1 - \tilde{\alpha} = 0.9$ there. This is also the alternative used by the OBF tests for futility stopping. Each entry in Table 8.1 is computed by Monte Carlo simulation with 100,000 replications. To compare tests $T$, $T'$ with type I error probability $\alpha$ but with different type II error probabilities $\tilde{\alpha}_T(\theta)$, $\tilde{\alpha}_{T'}(\theta)$ and expected sample sizes $E_\theta T$, $E_\theta T'$ at $\theta > 0$, Jennison and Turnbull (2006a) defined the efficiency ratio of $T$ to $T'$:

$$R_\theta(T,T') = \frac{(z_\alpha + z_{\tilde{\alpha}_T(\theta)})^2 / E_\theta T}{(z_\alpha + z_{\tilde{\alpha}_{T'}(\theta)})^2 / E_\theta T'} \times 100, \tag{8.37}$$

noting that $(z_\alpha + z_{\tilde{\alpha}_T(\theta)})^2/\theta^2$ is the sample size of the FSS test with the same type I error probability and power as $T$. Table 8.1 contains $R_\theta(T,N)$ for all tests $T$ and $\theta > 0$, where $N$ is the sample size of ADAPT.

ADAPT has power comparable to $\text{FSS}_{120}$ at all values of $\theta$ while achieving substantial savings in sample size, as shown by the percentiles and mean of the sample size. The three-stage OBF tests have power comparable to ADAPT and $\text{FSS}_{120}$, but ADAPT has sample size savings over the OBF tests, especially for larger $\theta > 0$, reflected by the efficiency ratio. The mean number of stages (denoted by #) reveals that although ADAPT allows for the possibility of three stages, most frequently it uses only one or two stages.

The conditional power tests PH, L, SF, and SF′ are underpowered at values of $\theta > 0$ in Table 8.1. In particular, PH, L, and SF all have power less than 0.6 at $\theta_1 = 0.3$, where ADAPT, $\text{FSS}_{120}$, and the OBF tests have power around 0.9. The lack

**Table 8.2** Expected sample size of ADAPT, optimal adaptive, and group sequential tests, with $\alpha = 0.025$, power 0.8 at $\theta'$, first group size $m$, and maximum sample size $M = 120$ for normal data with known variance

| Test | $m$ | $E_0 N$ | $E_{\theta'} N$ | $E_{2\theta'} N$ |
|---|---|---|---|---|
| ADAPT | 29 | 58.1 | 81.2 | 41.5 |
| $T_3^*$ | 29 | 54.9 | 78.9 | 39.5 |
| OGS(3) | 34 | 58.2 | 78.1 | 43.0 |
| $T_2^*$ | 43 | 64.0 | 85.3 | 49.0 |
| OGS(2) | 43 | 64.6 | 86.2 | 48.9 |
| $T_4^*$ | 24 | 50.9 | 75.2 | 36.0 |
| OGS(4) | 29 | 55.1 | 74.8 | 39.8 |

of power of PH, L, and SF shown by Table 8.1 is caused by stopping too early for futility. For example, the PH test stops for futility after the first stage if $S_m/\sqrt{m}$ falls below $z_{p^*} = 1.71$. But $P_{\theta_1}\{S_m/\sqrt{m} < 1.71\} = 0.44$, well exceeding the nominal type II error of 0.1. On the other hand, such stringent futility stopping is necessary to control the sample size of conditional power tests. For example, the 0.025-level PH test that stops for futility only when $\hat{\theta}_m \le 0$ (i.e., with $p^* = 0.5$) has expected sample size greater than $10^7$ at all values of $\theta$ in Table 8.1, yet power less than 0.9 at $\theta_1$. SF and SF$'$ provide another example of this behavior. Since these tests stop for futility at the first stage when $S_m/m \le \theta_1 - z_{\alpha_0}/\sqrt{m}$, the choice of $\alpha_0$ determines the maximum sample size. For maximum sample size $M = 120$, SF uses $\alpha_0 = 0.425$, a high rate of first-stage futility stopping which results in small expected sample sizes, low power, and a reduced type I error of 0.012, which is $\alpha = 0.025$ in the absence of futility stopping. In contrast, SF$'$ uses less stringent futility stopping with $\alpha_0 = 0.154$ that corresponds to maximum sample size $5M = 600$, which results in a type I error closer to 0.025 and better power, though it is still underpowered and its expected sample size exceeds 120 at $0.2 \le \theta \le 0.26$. The smallest $\alpha_0$ that does not perturb the type I error of 0.025 of Shen and Fisher's test is $\alpha_0 = 0.039$, but the resultant test has expected sample size 1856 at $\theta = 0$ and maximum sample size 52341.

The efficiency ratios relative to ADAPT in Table 8.1 are all less than 100 with the exception of PH, L, and SF at $\theta = 0.15$, but it is not clear that the efficiency ratio has much meaning in this case where the power of these tests is so low. For the other cases, it is natural to ask if much more improvement is possible. A benchmark for answering this question is provided by the optimal adaptive tests of Jennison and Turnbull (2006a,b) that minimize the expected sample size averaged over a collection of $\theta$ values, subject to a given type I error probability and power level at a prespecified alternative $\theta'$. Table 8.2 contains the expected sample size of $T_k^*$, the $k$-stage test minimizing

$$[E_0(T) + E_{\theta'}(T) + E_{2\theta'}(T)]/3 \tag{8.38}$$

among all $k$-stage tests with maximum sample size $M = 120$, type I error probability $\alpha = 0.025$ and power 0.8 at $\theta'$, and the alternative where FSS$_{100}$ has power 0.8, from Jennison and Turnbull (2006b, Table III). To this benchmark, we compare ADAPT with the same first group size $m = 29$ as $T_3^*$, $M = 120$, $\theta_1$ fixed at $\theta'$, and $b = 2.94$, $\tilde{b} = 0.7$, and $c = 2.05$ corresponding to $\varepsilon = 1/2$, $\tilde{\varepsilon} = 3/4$. Also included in Table 8.2 is the optimal $k$-stage "$\rho$-family" group sequential test (denoted by OGS($k$)) with $M = 120$, groups $2,\dots,k$ of size $(M-m)/(k-1)$, and with $m$ and $\rho$ chosen to minimize (8.38). Jennison and Turnbull (2006b) concluded that OGS($k$) is a computationally easier alternative to $T_k^*$, and Table 8.2 shows that their expected sample sizes are close at $\theta = 0,\ \theta',\ 2\theta'$. Note that ADAPT has expected sample size close to OGS(3) and $T_3^*$ even though the probability that ADAPT uses only one or two stages is 96.4%, 83.1%, and 98.4% for $\theta = 0,\ \theta'$, and $2\theta'$, respectively, showing that ADAPT very often behaves like a two-stage test. ADAPT has substantially smaller expected sample size than $T_2^*$ and OGS(2), however. On the other hand, $T_4^*$ is more efficient than ADAPT, but this is due in part to its smaller first group of $m = 24$, afforded by its additional stage. Here, we have matched the first group $m$=29 of ADAPT to that of $T_3^*$ for the purpose of comparison, but in practice there is flexibility in its choice of $m$. The $T_k^*$ and OGS($k$) tests, on the other hand, are rigid in their choice of $m$ that is determined by dynamic programming from the prespecified alternative $\theta'$, about which there may be some uncertainty before the trial.

### 8.3.2 Difference of Means with Unknown Variances

Let $X_1, X_2, \dots$ and $Y_1, Y_2, \dots$ be independent normal observations with unknown means $\mu_X, \mu_Y$ and variances $\sigma_X^2, \sigma_Y^2$, respectively. Table 8.3 reports a simulation study of Bartroff and Lai (2008b) to compare the performance of the adaptive test in Sect. 8.2.1 (denoted by ADAPT) with Stein's test, denoted by S, and the modified versions of Wittes and Brittain (1990, denoted by WB), Birkett and Day (1994, denoted by BD), and Denne and Jennison (1999, denoted by DJ) in the context of a Phase II hypercholesterolemia treatment efficacy trial described by Facey (1992). In this trial, patients were randomized into treatment and placebo groups, and serum cholesterol level reductions, $X_i$ and $Y_i$, assumed to be normally distributed, were measured after 4 weeks of treatment. A difference in reductions of serum cholesterol levels, in mmol/liter, between the treatment and placebo groups of 1.2 was of clinical interest. Based on previous studies, it was anticipated that the standard deviation of the reductions would be about 0.7 for both groups. If the standard deviation were known to be $\sigma_0 = 0.7$, the size of the fixed sample $t$-test with error probabilities $\alpha = \tilde{\alpha} = 0.05$ at mean difference 0 and $\delta = 1.2$ is 9 per group. Following Denne and Jennison (1999), we assume a first-stage per-group sample size of $m = 5$, approximately half of 9. If the standard deviation were in fact $2\sigma_0 = 1.4$, the per-group sample size of the same $t$-test is 31, which we take as a reasonable maximum sample size $M$ for our three-stage test with $\rho_m = 0.1$ and $\varepsilon = \tilde{\varepsilon} = 1/3$. Table 8.3 contains the power and per-group expected sample size of ADAPT and the aforementioned procedures in the literature, evaluated by 100,000 simulations at various values of $\mu_X - \mu_Y \in [0, \delta]$ and $\sigma = \sigma_X = \sigma_Y$. Whereas the Stein-type tests require this assumption of equal variances, the three-stage tests defined in Sect. 8.2.1 do not, so for comparison we also include in Table 8.3 the three-stage

**Table 8.3** Power and per-group expected sample size of tests of $H_0 : \mu_X \leq \mu_Y$

| $(\mu_X - \mu_Y, \sigma)$<br>$I$ | S | WB | BD | DJ | ADAPT | ADAPT$_{\neq}$ |
|---|---|---|---|---|---|---|
| $(0, \sigma_0/2)$ | 5.0% | 4.9% | 5.0% | 5.0% | 1.6% | 1.8% |
| $I = 0$ | 5.0 | 9.0 | 5.0 | 5.0 | 5.0 | 5.0 |
| $(0, \sigma_0)$ | 5.0% | 5.4% | 6.0% | 5.6% | 4.0% | 4.1% |
| $I = 0$ | 10.1 | 10.1 | 10.1 | 10.3 | 9.4 | 10.2 |
| $(\delta/2, \sigma_0)$ | 53.0% | 59.5% | 55.4% | 57.3% | 65.0% | 68.1% |
| $I = 0.169$ | 10.1 | 10.1 | 10.1 | 10.3 | 15.5 | 13.7 |
| $(\delta, \sigma_0)$ | 96.6% | 98.0% | 95.8% | 96.4% | 97.8% | 98.5% |
| $I = 0.551$ | 10.1 | 10.1 | 10.1 | 10.3 | 9.4 | 8.0 |
| $(0, 2\sigma_0)$ | 5.0% | 5.5% | 5.5% | 4.6% | 5.0% | 5.3% |
| $I = 0$ | 38.2 | 38.2 | 38.2 | 30.7 | 22.1 | 22.7 |
| $(\delta/2, 2\sigma_0)$ | 50.3% | 50.0% | 49.7% | 53.8% | 44.0% | 44.5% |
| $I = 0.045$ | 38.1 | 38.2 | 38.2 | 30.7 | 25.5 | 26.0 |
| $(\delta, 2\sigma_0)$ | 95.2% | 89.1% | 91.3% | 92.9% | 91.9% | 91.4% |
| $I = 0.169$ | 38.1 | 38.2 | 38.2 | 30.7 | 22.1 | 20.2 |
| $(0, 3\sigma_0)$ | 5.0% | 5.3% | 5.3% | 4.6% | 5.1% | 5.2% |
| $I = 0$ | 85.2 | 85.2 | 85.3 | 67.7 | 26.3 | 26.7 |
| $(0, 5\sigma_0)$ | 5.0% | 5.4% | 5.4% | 4.7% | 5.2% | 5.1% |
| $I = 0$ | 236 | 235 | 236 | 186 | 27.6 | 28.6 |
| $(0, 10\sigma_0)$ | 5.0% | 5.4% | 3.8% | 4.9% | 5.1% | 5.1% |
| $I = 0$ | 942 | 940 | 943 | 754 | 28.7 | 29.3 |

test that does not assume $\sigma_X = \sigma_Y$, which we denote by ADAPT$_{\neq}$. The Kullback–Leibler information number $I = \min\{I((\mu_X, \mu_Y, \sigma^2), (\tilde{\mu}_X, \tilde{\mu}_Y, \tilde{\sigma}^2)) : \tilde{\mu}_X - \tilde{\mu}_Y = 0\}$ is also reported in the first column. When the true standard deviations $\sigma_X$ and $\sigma_Y$ are equal to the specified value $\sigma_0$, ADAPT and ADAPT$_{\neq}$ have similar power but smaller expected sample size than the other tests for values of $\mu_X - \mu_Y$ near 0 and $\delta$. When the standard deviations $\sigma_X$ and $\sigma_Y$ are larger than the specified value $\sigma_0$, the adaptive tests have much smaller expected sample sizes than the Stein-type tests, whose second-stage sample size increases without bound as a function of the first-stage sample variance; in particular, see the last three rows of Table 8.3.

An alternative approach to Stein-type designs has been used by Proschan and Hunsberger (1995) and Li and Shih (2002), who simply replace the $\sigma^2$ in their two-stage tests that assume known variance with its current estimate at each stage. To compare ADAPT with these tests, which rely on stable variance estimates, we allow a larger first-stage sample size of $m = 20$. Table 8.4 contains the power and per-group expected sample size of Proschan and Hunsberger's test (denoted by PH), two choices of the early stopping boundaries $(h, k)$ in Table 1 of Li and Shih (2002) for their test, which we denote by L1 and L2, and our three-stage test ADAPT, for various values of $\mu_X - \mu_Y$ and $\sigma$, each entry being the result of 100,000 replications. To compare these tests on equal footing, we have chosen the

**Table 8.4** Maximum sample size $M$, power, and per-group expected sample size for the tests L1 and L2 of Li et al., Proschan and Hunsberger (PH), and ADAPT

| $(\mu_X - \mu_Y, \sigma)$ | | | | |
|---|---|---|---|---|
| $I$ | L1 | L2 | PH | ADAPT |
| $(0,1)$ | 5.5% | 5.3% | 5.5% | 4.8% |
| $I = 0$ | 26.3 | 25.5 | 25.9 | 56.5 |
| $(0,2)$ | 5.3% | 5.3% | 5.4% | 5.4% |
| $I = 0$ | 26.2 | 26.3 | 25.9 | 93.5 |
| $(1/4,1)$ | 29.9% | 29.3% | 29.0% | 48.3% |
| $I = 0.016$ | 32.8 | 31.0 | 31.6 | 76.1 |
| $(3/8,1)$ | 49.5% | 48.7% | 48.3% | 77.5% |
| $I = 0.035$ | 34.5 | 32.7 | 33.1 | 73.5 |
| $(1/2,1)$ | 67.8% | 66.4% | 66.4% | 92.8% |
| $I = 0.061$ | 34.3 | 32.8 | 32.7 | 63.3 |
| $(1/2,2)$ | 12.0% | 29.9% | 28.9% | 56.1% |
| $I = 0.016$ | 29.1 | 32.9 | 31.7 | 98.7 |
| $(3/4,2)$ | 49.8% | 48.5% | 48.1% | 85.6% |
| $I = 0.035$ | 34.5 | 32.7 | 33.2 | 87.0 |

maximum sample size $M = 121$ for ADAPT because this is the maximum sample size of L1 and is quite close to the maximum sample sizes of PH and L2, which are 122 and 104, respectively. The PH, L1, and L2 tests are designed to achieve type I error probability 0.05, and they choose the sample size of their second stage based on a conditional power level of 80%. The threshold values $b = 2.68$, $\tilde{b} = 1.75$, and $c = 1.75$ used by ADAPT are thus computed using $\alpha = 0.05$ and $\tilde{\alpha} = 0.20$. The results in Table 8.4 show that the true power of L1, L2, and PH falls well below their nominal conditional power level of 80%. When $\sigma = 2$, the L1, L2, and PH tests have power less than 50% for all values of $\mu_X - \mu_Y$ considered, which is caused by stopping prematurely for futility at the end of the first stage; see in particular the rows in Table 8.4 that correspond to $\mu_X - \mu_Y = 0$. Since the conditional power criterion is not valid when the estimated difference of means is near zero, the L and PH tests must stop for futility when this occurs even though the true difference of means may be substantially greater than zero.

### 8.3.3 Comparison of Tests Allowing Midcourse Modification of Maximum Sample Size

As pointed out in Sect. 8.1.3, Cui et al. (1999) have proposed a method to modify the group size of a given group sequential test of $H_0 : \theta \leq 0$ in response to protocol amendments during interim analyses. In the example considered by Cui et al. (1999, p. 854), the maximum sample size is initially $M = 125$ for detecting $\theta_1 = 0.29$ with power 0.9 and $\alpha = 0.025$ but can be subsequently increased up to $\widetilde{M} = 500$; their sample sizes are twice as large because they consider variance 2. They consider modifying the group size at the end of a given stage $L$ if the ratio of conditional power at the observed alternative $\hat{\theta}_{n_L}$ to the conditional power at $\theta_1$ is greater than 1 or less than 0.8, in which case the group size is then modified so that the new maximum sample size is

$$\widetilde{M} \wedge M \left(\theta_1 / \hat{\theta}_{n_L}\right)^2 . \tag{8.39}$$

If (8.39) is less than the already sampled $n_L$, error spending can be used to end the trial. The crux of this method is that the original critical values can be used for the weighted test statistic without changing the type I error probability regardless of how the sample size is changed. Table 8.5 compares their proposed adaptive group sequential tests with FSS tests, standard (nonadaptive) group sequential tests, and the adaptive test described in Sect. 8.2.2. Each result is based on 100,000 simulations. All adaptive tests in Table 8.5 use the first-stage sample size $m = 25$, maximum sample size initially $M = 125$ with the possibility of extension up to $\widetilde{M} = 500$, and type I error probability not exceeding $\alpha = 0.025$, matching the setting considered in Sect. 2 of Cui et al. (1999). Since the maximum sample size can vary between $M = 125$ and $\widetilde{M} = 500$, the two relevant implied alternatives are $\theta_1 = 0.29$, where $\text{FSS}_{125}$ has power $1 - \tilde{\alpha} = 0.9$, and $\theta_2 = 0.15$, where $\text{FSS}_{500}$ has power 0.9. The values of the user-specified parameters of the tests in Table 8.5 are summarized as follows:

- ADAPT: The adaptive test, described in Sect. 8.2.2, that uses $b = 3.48$, $\tilde{b} = 2.1$, and $c = 2.31$ corresponding to $\varepsilon = \tilde{\varepsilon} = 1/2$, $\rho_m = 0.1$, and $M' = 250$.
- $\text{FSS}_{125}$, $\text{FSS}_{500}$: The FSS tests having sample sizes $M = 125$ and $\widetilde{M} = 500$, respectively.
- $\text{OBF}^5_{\text{SC}}$: A one-sided O'Brien–Fleming group sequential test having five groups of size 100 and that uses stochastic curtailment futility stopping ($\gamma = 0.9$ in Sect. 10.2 of Jennison and Turnbull, 2000) with reference alternative $\theta_2 = 0.15$; see below.
- $\text{C}^4$, $\text{C}^5$: Two versions of the adaptive group sequential test of Cui et al. (1999) that adjusts the group size at the end of the first stage; $\text{C}^4$ uses four stages and $\text{C}^5$ uses five stages.
- $\text{C}^5_{\text{SC}}$, $\text{C}^5_{\text{PF}}$: Two modifications of $\text{C}^5$ to allow for futility stopping; $\text{C}^5_{\text{SC}}$ uses stochastic curtailment futility stopping ($\gamma = 0.9$ in Sect. 10.2 of Jennison and Turnbull, 2000) and $\text{C}^5_{\text{PF}}$ uses power family futility stopping ($\Delta = 1$ in Sect. 4.2 of Jennison and Turnbull, 2000). Both $\text{C}^5_{\text{SC}}$ and $\text{C}^5_{\text{PF}}$ use reference alternative $\theta_2 = 0.15$.

Since $\text{OBF}^5_{\text{SC}}$, $\text{C}^5_{\text{SC}}$, and $\text{C}^5_{\text{PF}}$ have maximum sample size $\widetilde{M} = 500$, the futility stopping boundaries of these tests are designed to have power 0.9 at $\theta_2$. We have also included $\text{C}^4$ because our adaptive test uses no more than four stages. The tests are evaluated at the $\theta$ values where $\text{FSS}_{125}$ has power 0.01, 0.025, 0.7, 0.8, and 0.9 and where $\text{FSS}_{500}$ has power 0.7, 0.8, and 0.9.

Even though the C tests have maximum sample size $\widetilde{M} = 500$, they are underpowered at $0 < \theta \leq \theta_2$, the alternative implied by $\widetilde{M}$, when compared with

**Table 8.5** Power (italics), expected sample size (italics), sample size quantiles $T_q$, expected number of stages (bold), and efficiency ratio (at $\theta > 0$) with respect to ADAPT, of FSS, adaptive, and group sequential tests with maximum sample size $M = 120$ except for SF$'$, which uses $5M$

| Test | ADAPT | FSS$_{120}$ | OBF$_{\text{EF}}$ | OBF$_{\text{SC}}$ | PH | L | SF | SF$'$ |
|---|---|---|---|---|---|---|---|---|
| $\theta = -0.03$ | *1.1%* | *1.0%* | *0.9%* | *1.0%* | *1.5%* | *1.3%* | *0.6%* | *0.9%* |
| | *68.5* | *120.0* | *72.3* | *91.7* | *40.8* | *41.4* | *41.3* | *72.3* |
| $T_{.25}$ | 40 | 120 | 40 | 80 | 40 | 40 | 40 | 40 |
| $T_{.5}$ | 40 | 120 | 80 | 80 | 40 | 40 | 40 | 40 |
| $T_{.75}$ | 120 | 120 | 80 | 120 | 40 | 40 | 40 | 40 |
| # | **1.53** | **1.00** | **1.81** | **2.29** | **1.01** | **1.03** | **1.03** | **1.14** |
| $\theta = 0$ | *2.5%* | *2.5%* | *2.3%* | *2.5%* | *2.4%* | *2.5%* | *1.2%* | *2.2%* |
| | *75.1* | *120.0* | *77.8* | *96.4* | *41.1* | *42.2* | *41.2* | *82.3* |
| $T_{.25}$ | 40 | 120 | 80 | 80 | 40 | 40 | 40 | 40 |
| $T_{.5}$ | 60 | 120 | 80 | 80 | 40 | 40 | 40 | 40 |
| $T_{.75}$ | 120 | 120 | 80 | 120 | 40 | 40 | 40 | 40 |
| # | **1.64** | **1.00** | **1.94** | **2.41** | **1.02** | **1.05** | **1.05** | **1.20** |
| $\theta = 0.15$ | *35.6%* | *37.6%* | *35.7%* | *37.1%* | *18.7%* | *20.9%* | *13.8%* | *36.1%* |
| | *98.6* | *120.0* | *98.9* | *110.2* | *44.5* | *48.3* | *47.2* | *115.3* |
| $T_{.25}$ | 71 | 120 | 80 | 120 | 40 | 40 | 40 | 40 |
| $T_{.5}$ | 120 | 120 | 120 | 120 | 40 | 40 | 40 | 47 |
| $T_{.75}$ | 120 | 120 | 120 | 120 | 40 | 40 | 40 | 146 |
| # | **2.05** | **1.00** | **2.47** | **2.76** | **1.09** | **1.22** | **1.22** | **1.53** |
| $R_\theta(T,N)$ | *100* | *78.5* | *99.5* | *86.4* | *332* | *289* | *358* | *84.5* |
| $\theta = 0.20$ | *57.2%* | *60.0%* | *57.9%* | *59.5%* | *30.2%* | *33.2%* | *24.8%* | *53.5%* |
| | *99.4* | *120.0* | *101.4* | *108.0* | *45.9* | *50.8* | *50.6* | *124.3* |
| $T_{.25}$ | 76 | 120 | 80 | 80 | 40 | 40 | 40 | 40 |
| $T_{.5}$ | 120 | 120 | 120 | 120 | 40 | 40 | 40 | 65 |
| $T_{.75}$ | 120 | 120 | 120 | 120 | 40 | 51 | 52 | 157 |
| # | **2.07** | **1.00** | **2.54** | **2.70** | **1.11** | **1.30** | **1.86** | **1.66** |
| $R_\theta(T,N)$ | *100* | *88.5* | *99.7* | *97.2* | *98.1* | *99.3* | *70.1* | *73.1* |
| $\theta = 0.26$ | *77.4%* | *80.0%* | *78.0%* | *79.6%* | *44.2%* | *47.5%* | *38.2%* | *67.5%* |
| | *95.2* | *120.0* | *99.8* | *102.0* | *46.6* | *52.7* | *52.9* | *120.2* |
| $T_{.25}$ | 59 | 120 | 80 | 80 | 40 | 40 | 40 | 41 |
| $T_{.5}$ | 120 | 120 | 80 | 120 | 40 | 40 | 40 | 68 |
| $T_{.75}$ | 120 | 120 | 120 | 120 | 40 | 60 | 60 | 145 |
| # | **2.00** | **1.00** | **2.47** | **2.55** | **1.13** | **1.38** | **1.47** | **1.78** |
| $R_\theta(T,N)$ | *100* | *84.7* | *96.8* | *98.6* | *91.4* | *88.4* | *67.4* | *62.7* |
| $\theta = \theta_1 = 0.3$ | *88.8%* | *90.0%* | *88.6%* | *89.5%* | *55.2%* | *58.0%* | *49.1%* | *75.5%* |
| | *89.2* | *120.0* | *94.5* | *96.4* | *46.8* | *53.3* | *54.0* | *111.8* |
| $T_{.25}$ | 40 | 120 | 80 | 80 | 40 | 40 | 40 | 43 |
| $T_{.5}$ | 118 | 120 | 80 | 80 | 40 | 40 | 42 | 65 |
| $T_{.75}$ | 120 | 120 | 120 | 120 | 40 | 62 | 62 | 128 |
| # | **1.91** | **1.00** | **2.36** | **2.41** | **1.14** | **1.42** | **1.58** | **1.85** |
| $R_\theta(T,N)$ | *100* | *77.5* | *93.8* | *94.7* | *82.6* | *77.5* | *61.5* | *55.6* |

(continued)

**Table 8.5** (continued)

| Test | ADAPT | $\text{FSS}_{120}$ | $\text{OBF}_{\text{EF}}$ | $\text{OBF}_{\text{SC}}$ | PH | L | SF | SF$'$ |
|---|---|---|---|---|---|---|---|---|
| $\theta = 0.33$ | *94.0%* | *95.0%* | *94.1%* | *94.7%* | *63.5%* | *66.5%* | *57.3%* | *80.8%* |
| | *83.0* | *120.0* | *90.1* | *91.1* | *46.7* | *53.3* | *54.3* | *103.5* |
| $T_{.25}$ | 40 | 120 | 80 | 80 | 40 | 40 | 40 | 43 |
| $T_{.5}$ | 89 | 120 | 80 | 80 | 40 | 40 | 44 | 61 |
| $T_{.75}$ | 120 | 120 | 120 | 120 | 40 | 62 | 62 | 114 |
| # | **1.81** | **1.00** | **2.25** | **2.28** | **1.13** | **1.44** | **1.65** | **1.89** |
| $R_\theta(T,N)$ | *100* | *72.8* | *92.6* | *94.3* | *76.4* | *71.8* | *56.9* | *52.0* |
| $T_{.25}$ | 40 | 120 | 80 | 80 | 40 | 40 | 40 | 43 |
| $T_{.5}$ | 89 | 120 | 80 | 80 | 40 | 40 | 44 | 61 |
| $T_{.75}$ | 120 | 120 | 120 | 120 | 40 | 62 | 62 | 114 |
| # | **1.81** | **1.00** | **2.25** | **2.28** | **1.13** | **1.44** | **1.65** | **1.89** |
| $R_\theta(T,N)$ | *100* | *72.8* | *92.6* | *94.3* | *76.4* | *71.8* | *56.9* | *52.0* |

ADAPT, $\text{FSS}_{500}$, and $\text{OBF}^5_{\text{SC}}$. In particular, the C tests have power less than 0.65 at $\theta_2$. Since $\text{C}^4$ and $\text{C}^5$ use no futility stopping, this suggests that their updated maximum sample size (8.39) (with $L = 1$) has contributed to the power loss. The large expected sample sizes of $\text{C}^4$ and $\text{C}^5$ at $\theta \leq 0$ reveal another problem with this sample size updating rule: It does not consider the sign of $\hat{\theta}_m$; a negative value of $\hat{\theta}_m$ could result in the same sample size modification as a positive one, causing a large increase in the group size when it should be decreased toward futility stopping. ADAPT has only a slight loss of power in comparison with $\text{FSS}_{500}$ and the five-stage $\text{OBF}^5_{\text{SC}}$ at $\theta > 0$, with substantially smaller expected sample size. The mean number of stages of ADAPT at $\theta_1 = 0.29$ shows that it behaves like a two- or three-stage test there. $\text{OBF}^5_{\text{SC}}$, on the other hand, has the largest expected sample size at $\theta \geq 0$ of the tests in Table 8.5 other than $\text{FSS}_{125}$.

### 8.3.4 *Coronary Intervention Study*

The National Heart, Lung and Blood Institute (NHLBI) type II Coronary Intervention Study (Brensike et al., 1982) was designed to investigate the cholesterol-lowering effects of cholestyramine on patients with type II hyperlipoproteinemia and coronary artery disease. Patients were randomized into cholestyramine and placebo groups, and coronary angiography was performed before and after five years of treatment. It was found that the disease had progressed in 20 of 57 in the placebo group and 15 of 59 in the cholestyramine group. Proschan and Hunsberger (1995) and Li and Shih (2002) have considered how this study could have been extended by using their two-stage tests for the difference in two normal means with common unknown variance. To apply these tests to the NHLBI study, they assumed the first-stage sample size to be $58 = (57 + 59)/2$ for the normal problem and used the arcsine transformation so that the difference between the transformed binomial frequencies, $p_1$ for the placebo group and $p_2$ for the treatment group, is approximately normally distributed; details are given in the next paragraph. As an alternative we apply the three-stage test in Sect. 8.2.1 to two binomial populations. In the notation of Sect. 8.2, to test $H_0 : p_2 \leq p_1$ we have $\theta = (p_1, p_2)^T$, $u(\theta) = p_2 - p_1$, $u_0 = 0$, and the test statistic $\inf_{\theta:u(\theta)=\delta} nI(\hat{\theta}_n, \theta)$ takes the form

$$n\left\{\hat{p}_{1,n}\log\left(\frac{\hat{p}_{1,n}}{p_{\delta,n}}\right) + \hat{q}_{1,n}\log\left(\frac{\hat{q}_{1,n}}{1-p_{\delta,n}}\right)\right.$$
$$\left. + \hat{p}_{2,n}\log\left(\frac{\hat{p}_{2,n}}{p_{\delta,n}+\delta}\right) + \hat{q}_{2,n}\log\left(\frac{\hat{q}_{2,n}}{1-p_{\delta,n}-\delta}\right)\right\}, \qquad (8.40)$$

where $\hat{p}_{i,n}$ is the maximum likelihood estimator of $p_i$ based on $n$ observations, $\hat{q}_{i,n} = 1 - \hat{p}_{i,n}$, and $p_{\delta,n}$ is the maximum likelihood estimator of $p_1$ under the assumption $p_2 - p_1 = \delta$. The treatment and placebo groups are assumed to have the same per-group sample size during interim analyses, following Proschan and Hunsberger (1995) and Li and Shih (2002).

Letting $S_n$ denote the sum of independent normal random variables with mean $\mu$ and variance 1, following a pilot study of size $m$ resulting in $S_m = s_m$, Proschan and Hunsberger's (1995) test chooses $n_2$ and critical value $c$ to satisfy the conditional power criterion

$$P\left\{S_{n_2}/n_2^{1/2} > c \,\middle|\, S_m = s_m, \mu = s_m/m^{1/2}\right\} \geq 1 - \tilde{\alpha} \qquad (8.41)$$

and type I error constraint

$$P_0\left\{S_{n_2}/n_2^{1/2} > c \,\middle|\, S_m = s_m\right\} = \alpha. \qquad (8.42)$$

In order to solve for $n_2$ and $c$, a parametric form for the probability in (8.42) is assumed, which contains a user-specified futility boundary $h$ and critical value $k$ for the internal pilot. Li and Shih (2002) introduce a modification of Proschan and Hunsberger's (1995) test in which the critical value $c$ is specified before the internal pilot study but $h$, $k$, and $n_2$ are chosen to satisfy (8.41) and (8.42) after the internal pilot study. This modification allows approximations to the probabilities in (8.41) and (8.42) to be used in lieu of a specific parametric form. For the coronary intervention study, Proschan and Hunsberger (1995) and Li and Shih (2002) propose using these tests with the variance-stabilizing transformation $S_n = (2n)^{1/2}\{\arcsin(\hat{p}_{1,n}^{1/2}) - \arcsin(\hat{p}_{2,n}^{1/2})\}$.

Table 8.6 gives the power, per-group expected sample size, and efficiency ratio (8.37), using the normal approximation, relative to ADAPT (for alternatives $p_2 > p_1$) of the following tests for various values of $p_1, p_2$ near $15/59 = 0.254$ and $20/57 = 0.351$, the values observed in the NHLBI study (Brensike et al., 1982).

- L: Li and Shih's (2002) test with $h = 1.036$, $k = 1.82$, $c = 1.7$, $\alpha = .05$, conditional power level 0.8, and first-stage size $m = 58$.

**Table 8.6** Power, expected sample size, and efficiency ratio (in parentheses and at $p_2 > p_1$) of the tests of $H_0 : p_2 \leq p_1$

| $p_1$ | $p_2$ | L | PH | ADAPT |
|---|---|---|---|---|
| 0.20 | 0.15 | 0.7% | 0.7% | 0.3% |
| | | 63.4 | 63.0 | 98.6 |
| | 0.20 | 5.2% | 5.2% | 5.0% |
| | | 75.8 | 74.5 | 158.2 |
| | 0.30 | 53.0% | 51.8% | 81.8% |
| | | 102.0 (89.7) | 97.2 (90.8) | 206.1 (100) |
| | 0.35 | 77.1% | 76.2% | 97.4% |
| | | 95.3 (73.3) | 90.7 (75.1) | 160.5 (100) |
| 0.25 | 0.20 | 0.8% | 1.0% | 0.4% |
| | | 64.7 | 64.5 | 111.2 |
| | 0.25 | 5.2% | 5.1% | 5.0% |
| | | 77.3 | 75.8 | 171.2 |
| | 0.35 | *48.3%* | *47.0%* | *79.2%* |
| | | *97.7 (90.5)* | *93.3 (91.9)* | *213.1 (100)* |
| | 0.40 | 72.7% | 71.7% | 96.7% |
| | | 94.1 (74.1) | 89.7 (76.3) | 170.3 (100) |
| 0.30 | 0.25 | 0.9% | 0.9% | 0.4% |
| | | 65.5 | 64.7 | 122.2 |
| | 0.30 | 5.1% | 5.0% | 5.0% |
| | | 75.1 | 73.7 | 177.0 |
| | 0.40 | 45.3% | 44.3% | 76.6% |
| | | 96.4 (92.7) | 92.0 (95.1) | 218.3 (100) |
| | 0.45 | 70.9% | 69.9% | 96.2% |
| | | 96.1 (75.2) | 91.4 (77.6) | 176.9 (100) |

The italicized numbers represent those at the NHLBI parameter values, where L and PH are markedly under-powered

- PH: Proschan and Hunsberger's (1995) test with $h = 1.036$, $k = 1.82$, $\alpha = 0.05$, conditional power level 0.8, and first-stage size $m = 58$.
- ADAPT: The adaptive test described in a previous paragraph with $m = 58$, $M = 302$ (the maximum sample size of L), and thresholds $b = 2.36$, $\tilde{b} = 1.1$, and $c = 1.55$ corresponding to $\alpha = 0.05$, $\tilde{\alpha} = 0.2$, and $\varepsilon = \tilde{\varepsilon} = 1/2$.

All three tests use the same first-stage size $m = 58$. ADAPT matches the maximum sample size $M = 302$ of L, and the parameters of PH determine its maximum sample size to be slightly larger at 354. The actual power of L and PH is around 50% for the values of $p_1$ and $p_2$ in Table 8.6 with $p_2 - p_1 = 0.1$, and is less than 50% when $p_1 = 0.254$ and $p_2 = 0.351$ where they were designed to have conditional power 80%. This is caused in part by premature stopping for futility at the end of the first stage. Indeed, L and PH use the same futility boundary and their probability of stopping at the end of the first stage when $p_1 = 0.254$ and $p_2 = .0351$ is 0.47, well exceeding the nominal type II error probability 0.2. One might ask if a conditional power test can avoid this phenomenon by using a larger first-stage sample size so that the estimate $\hat{p}_2 - \hat{p}_1$ is near 0 less often after the first stage when the true difference $p_2 - p_1$ is substantially greater than 0. If the first-stage sample size of L is raised to 162 (raising the maximum sample size to 1331), the resultant test has power 79% when $p_1 = 0.254$ and $p_2 = 0.351$, approximately equal the power of ADAPT. However, the expected sample size of this version of L is 264 at this alternative, compared to the expected sample size 213.1 of ADAPT. Similar oversampling also occurs for the values of $p_1$ and $p_2$ in Table 8.6 with $p_2 - p_1 > 0.1$, where the power of L and PH is closer to the nominal conditional power level of 80%, but the efficiency ratio drops to around 75%.

## 8.4 Adaptive Choice Between Superiority and Non-inferiority Objectives via a Flexible Group Sequential Design

In the design of controlled clinical trials comparing a new treatment with an active control, one often has to choose between two different study objectives: either a superiority or a non-inferiority hypothesis that the new treatment is more effective, or no worse (within certain indifference limits) than, the active control. The following example concerning the clinical trial design of a new antimicrobial drug in Lai et al. (2006c) illustrates some of the issues in the choice between these two study objectives, at the design stage when there is not enough information to decide on which objective has a better chance of success. Let $p_1$ and $p_2$ denote the response rates of the new and control drugs, respectively. The null hypothesis is $H_0 : p_1 - p_2 \leq -\gamma$ for a non-inferiority trial and is $H_0' : p_1 - p_2 \leq 0$ for a superiority trial. The $\gamma$ is chosen by the FDA that requires two trials to prove non-inferiority and only one trial to establish superiority. The pharmaceutical company developing the new drug does not have a good feel of the magnitude of $p_1 - p_2$ to decide whether it should perform two non-inferiority trials or one superiority trial and would like to have a flexible design which can adapt to the information about $p_1 - p_2$ during interim analyses so that it can switch from the superiority to the non-inferiority objective, if needed.

For the simpler problem of choosing between a superiority and only one (as opposed to two) non-inferiority group sequential tests after specifying $H_1 : p_1 - p_2 > -\gamma$ (or $H_1' : p_1 - p_2 > 0$) as the alternative hypothesis of non-inferiority (or superiority), a different adaptive group sequential strategy has been proposed by Wang et al. (2001). Their approach, described in Sect. 8.4.2, uses a conditional power criterion to decide whether one should switch from the superiority to non-inferiority alternative during interim analyses and involves modifying the studentized test statistics along the lines introduced by Cui et al. (1999) to avoid inflation of the type I error due to such data-dependent switch; see Sect. 8.1.3. In this section we present a more efficient procedure, proposed by Lai et al. (2006c), that uses the group sequential GLR tests in Sect. 4.2. The basic idea is first described in exponential families and then specialized to the Bernoulli case, for which it is compared with the test of Wang et al. (2001).

As noted by Jennison and Turnbull (2006b), standard group sequential tests with the first stage chosen optimally are nearly as efficient as their optimal adaptive tests. Instead of using an adaptive design of the type in Sect. 8.2, we use a group sequential design that conveniently combines a sequential GLR test of superiority with that of non-inferiority. We begin by considering the problem of adaptively choosing between testing $H_0 : \theta \leq \theta_0$ and the larger null hypothesis $H_0' : \theta \leq \theta_0'$ in a one-parameter exponential family $f_\theta = \exp(\theta x - \psi(x))$, with $\theta_0' > \theta_0$ and significance level $\alpha$. For the special case of $\theta_0 = -\delta$ and $\theta_0' = 0$ related to a normal mean $\theta$, the alternative hypotheses $H_1 : \theta > -\delta$ and $H_1' : \theta > 0$ are often used as the non-inferiority and superiority alternatives. Let $M$ be the maximum allowable sample size for testing non-inferiority (i.e., $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$), with $k$ interim analyses so that $n_k = M$. Let $M' = n_{k'}$ be the maximum allowable sample size for testing superiority ($H_0'$ versus $H_1'$), with $k'$ interim analyses.

### 8.4.1 Adaptation and Group Sequential Tests of $H_0$ or $H_0'$ When $M' \leq M$

In practice the non-inferiority margin, which is prescribed by the regulatory agency, is often smaller than the distance between $\theta_0'$ and the alternative under which superiority is considered for sample size determination, resulting in $M' < M$ and therefore also $k' < k$. In this case, we can modify the group sequential test in Sect. 4.2.2 as follows to allow concurrent testing of (i) $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ and (ii) $H_0' : \theta \leq \theta_0'$ versus $H_1' : \theta > \theta_0'$. We start the study by testing superiority in the first $k'$ stages. If the superiority null hypothesis $H_0'$ is not rejected during the first $k'$ stages, the test switches to testing non-inferiority ($H_0$ versus $H_1$) thereafter. During these $k'$ stages and after stage $k'$, the test can also terminate early with acceptance of $H_0$ if the lower futility boundary of $H_0$ versus the alternative $\theta = \theta_0'$ in $H_1$ is crossed.

Specifically, the test uses the GLR statistics $n_i I(\hat{\theta}_{n_i}, \theta_0)$ and $n_i I(\hat{\theta}_{n_i}, \theta_0')$ in conjunction with a stopping rule of the form

$$\hat{\theta}_{n_i} > \theta_0' \text{ and } n_i I\left(\hat{\theta}_{n_i}, \theta_0'\right) \geq b' \text{ if } 1 \leq i \leq k'-1, \tag{8.43a}$$

$$\text{or} \quad S_{n_{k'}} \geq c' \quad \text{or} \quad \hat{\theta}_{n_{k'}} > \theta_0 \text{ and } n_{k'} I\left(\hat{\theta}_{n_{k'}}, \theta_0'\right) \geq b, \tag{8.43b}$$

$$\text{or} \quad \hat{\theta}_{n_i} > \theta_0 \text{ and } n_i I\left(\hat{\theta}_{n_i}, \theta_0\right) \geq b \text{ if } k' < i < k, \tag{8.43c}$$

$$\text{or} \quad \hat{\theta}_{n_i} < \theta_0' \text{ and } n_i I\left(\hat{\theta}_{n_i}, \theta_0'\right) \geq \tilde{b} \text{ if } 1 \leq i < k. \tag{8.43d}$$

If (8.43a) holds, the test rejects $H_0'$ (in favor of the superiority alternative) upon stopping. If stopping occurs at the $k'$th interim analysis, the test rejects $H_0'$ (in favor of the superiority alternative) when $S_{n_{k'}} \geq c'$, and rejects $H_0$ (in favor of the non-inferiority alternative) when $S_{n_{k'}} < c'$ and the other event in (8.43b) occurs. If stopping occurs with (8.43c), the test rejects $H_0$ (in favor of the non-inferiority alternative) upon stopping. Early stopping (due to futility) with acceptance of $H_0$ occurs with (8.43d). If stopping does not occur during the first $k-1$ interim analyses, the test rejects $H_0$ (in favor of the non-inferiority alternative) when $S_{n_k} \geq c$ at the $k$th interim analysis.

The thresholds $\tilde{b}$, $b'$, $c'$, $b$, and $c$ in the preceding test are determined as follows to ensure that the test has probability no larger than $\alpha$ for wrongly claiming superiority when $H_0'$ is true (and equal to $\alpha$ at $\theta_0'$) and for wrongly claiming either non-inferiority or superiority when $H_0$ is true (and equal to $\alpha$ at $\theta_0$). Let $0 \leq \varepsilon < \frac{1}{2}$ and define $\tilde{b}$, $b'$, and then $c'$ by

$$P_{\theta_0'}\left\{\hat{\theta}_{n_i} < \theta_0' \text{ and } n_i I\left(\hat{\theta}_{n_i}, \theta_0'\right) \geq \tilde{b} \text{ for some } i \leq k-1\right\} = \varepsilon\tilde{\alpha}, \tag{8.44}$$

$$P_{\theta_0'}\left\{\hat{\theta}_{n_i} > \theta_0' \text{ and } n_i I\left(\hat{\theta}_{n_i}, \theta_0'\right) \geq b' \text{ for some } i \leq k'-1\right\} = \varepsilon\alpha, \tag{8.45}$$

$$P_{\theta_0'}\left\{\text{Test terminates with (8.43a) for some } i \leq k'-1 \text{ or } S_{n_{k'}} \geq c'\right\} = \alpha. \tag{8.46}$$

Note that the test switches the null hypothesis from $H_0$ to $H_0'$ at the $k'$th interim analysis if $H_0'$ is not rejected and stopping has not occurred by that time. Letting $A_{k'-1} = \{n_i I(\hat{\theta}_{n_i}, \theta_0')1_{\{\hat{\theta}_{n_i} > \theta_0'\}} < b' \text{ and } n_i I(\hat{\theta}_{n_i}, \theta_0')1_{\{\hat{\theta}_{n_i} < \theta_0'\}} < \tilde{b} \text{ for all } i \leq k'-1\}$, define $b$ and $c$ by the equations

$$\begin{aligned} &P_{\theta_0}\{\text{Test rejects } H_0'\} + \sum_{j=k'}^{k-1} P_{\theta_0}\Big[\left\{S_{n_{k'}} < c'\right\} \cap A_{k'-1} \\ &\quad \cap \Big\{\hat{\theta}_{n_j} > \theta_0 \text{ and } n_j I\left(\hat{\theta}_{n_j}, \theta_0\right) \geq b,\ n_i I\left(\hat{\theta}_{n_i}, \theta_0\right) 1_{\{\hat{\theta}_{n_i} > \theta_0\}} < b \\ &\quad \text{and } n_i I\left(\hat{\theta}_{n_i}, \theta_0'\right) 1_{\{\hat{\theta}_{n_i} < \theta_0'\}} < \tilde{b} \text{ for } k' \leq i < j\Big\}\Big] = \varepsilon\alpha, \end{aligned} \tag{8.47}$$

$$\begin{aligned} &P_{\theta_0}\Big[\left\{S_{n_{k'}} < c'\right\} \cap A_{k'-1} \cap \Big\{S_{n_k} \geq c,\ n_i I\left(\hat{\theta}_{n_i}, \theta_0\right) 1_{\{\hat{\theta}_{n_i} > \theta_0\}} < b \\ &\quad \text{and } n_i I\left(\hat{\theta}_{n_i}, \theta_0'\right) 1_{\{\hat{\theta}_{n_i} < \theta_0'\}} < \tilde{b} \text{ for } k' \leq i \leq k-1\Big\}\Big] = (1-\varepsilon)\alpha. \end{aligned} \tag{8.48}$$

From (8.45)–(8.48), it follows that $P_{\theta_0'}(\text{Test rejects } H_0') = \alpha$ and $P_{\theta_0}(\text{Test rejects } H_0' \text{ or } H_0) = \alpha$. By monotonicity, the test has probability no larger than $\alpha$ for wrongly claiming superiority when $H_0'$ is true, and its probability of wrongly claiming a positive result (either superiority or non-inferiority) when $H_0$ holds also does not exceed $\alpha$. The lower futility boundary in (8.43d) is chosen in (8.44) so that the power to reject the non-inferiority null hypothesis $H_0$ at the alternative $\theta = \theta_0'$ does not differ much from $1 - \tilde{\alpha}$, which is the power of the fixed sample size test (with sample size $M$). Note that early stopping due to futility is incorporated in the determination of $c'$, $b$, and $c$ via (8.46)–(8.48).

The preceding adaptive strategy assumes a larger maximum sample size $M$ for testing non-inferiority than the corresponding $M'$ for testing superiority. It is still applicable to the case $M = M'$ (and therefore $k = k'$), for which the stopping rule has the simpler form:

$$\hat{\theta}_{n_i} > \theta_0' \text{ and } n_i I\left(\hat{\theta}_{n_i}, \theta_0'\right) \geq b' \text{ if } 1 \leq i \leq k-1, \tag{8.49a}$$

or

$$\hat{\theta}_{n_i} < \theta_0' \text{ and } n_i I\left(\hat{\theta}_{n_i}, \theta_0'\right) \geq \tilde{b} \text{ if } 1 \leq i \leq k-1. \tag{8.49b}$$

Note that (8.49a) corresponds to (8.43a) with $k' = k$, while (8.49b) corresponds to (8.43d). If (8.49a) holds, the test rejects $H_0'$ (in favor of the superiority alternative) upon stopping. Early stopping (due to futility) with acceptance of $H_0$ occurs with (8.49b). If stopping does not occur in the first $k-1$ interim analyses, then at the final analysis, the test rejects $H_0'$ (in favor of superiority) if $S_{n_k} > c'$, rejects $H_0$ (in favor of non-inferiority) if $\hat{\theta}_{n_k} > \theta_0$ and $n_k I(\hat{\theta}_{n_k}, \theta_0) \geq c$, or accepts $H_0$ otherwise. Letting $0 < \varepsilon < \frac{1}{2}$ and defining $A_{k'-1}$ as before but with $k' = k$, the thresholds $\tilde{b}$, $b'$, $c'$, and $c$ are determined by

$$P_{\theta_0'}\left\{\hat{\theta}_{n_i} < \theta_0' \text{ and } n_i I\left(\hat{\theta}_{n_i}, \theta_0'\right) \geq \tilde{b} \text{ for some } i \leq k-1\right\} = \varepsilon\tilde{\alpha}, \tag{8.50a}$$

$$P_{\theta_0'}\left\{\hat{\theta}_{n_i} > \theta_0' \text{ and } n_i I\left(\hat{\theta}_{n_i}, \theta_0'\right) \geq b' \text{ for some } i \leq k-1\right\} = \varepsilon\alpha, \tag{8.50b}$$

$$P_{\theta_0'}\left\{\text{Test terminates with (8.49a) for some } i \leq k-1 \text{ or } S_{n_k} \geq c'\right\} = \alpha, \tag{8.50c}$$

$$\begin{aligned} &P_{\theta_0}\{\text{Test rejects } H_0'\} \\ &\quad + P_{\theta_0}\left[A_{k-1} \cap \{S_{n_k} < c'\} \cap \left\{\hat{\theta}_{n_k} > \theta_0 \text{ and } n_k I\left(\hat{\theta}_{n_k}, \theta_0\right) \geq c\right\}\right] = \alpha. \end{aligned} \tag{8.50d}$$

As noted in Sect. 4.2.4, the group sequential GLR test of the form (8.43) can be readily extended to multiparameter and multiarmed settings. Therefore, the group sequential test of $H_0$ or $H_0'$ described above can be similarly extended to the multiparameter exponential family and to multiarmed clinical trials. It can also be modified for the case $M < M'$ (and therefore $k < k'$), as shown by Lai et al. (2006c, p. 1159–1161).

### *8.4.2 Binary Responses and a Comparative Study*

Suppose there are $m = 2$ treatment groups and the responses $X_{i,n}$ are Bernoulli random variables with $P\{X_{i,n} = 1\} = p_i = 1 - P\{X_{i,n} = 0\}$, $i = 1, 2$. In this case, $\theta_i = \log\{p_i/(1-p_i)\}$ and

$$I(\theta, \theta') = p\log(p/p') + (1-p)\log\left\{(1-p)/(1-p')\right\}. \tag{8.51}$$

Letting $p_1$ denote the response probability of the experimental treatment and $p_2$ that of the control treatment, testing for superiority (respectively non-inferiority) of the experimental treatment involves the null hypothesis $H_0' : d \leq 0$ (respectively $H_0 : d \leq -\gamma$), where $d = p_1 - p_2$ and $\gamma > 0$ denotes a prescribed non-inferiority margin. In this case, the GLR statistic for testing $p_1 - p_2 = \delta$ at the $j$th interim analysis can be expressed explicitly as

$$\Lambda_j(\delta) = \sum_{i=1}^{2} n_{ij} \left\{ \hat{p}_{i,j} \log\left(\hat{p}_{i,j}/\bar{p}_{i,j}(\delta)\right) + (1-\hat{p}_{i,j}) \log\left[(1-\hat{p}_{i,j})/(1-\bar{p}_{i,j}(\delta))\right] \right\}, \tag{8.52}$$

where $\hat{p}_{i,j} = \bar{X}_{i,n_{ij}}$, $\bar{p}_{1,j}(\delta) = p + \delta$, and $\bar{p}_{2,j}(\delta) = p$, in which $p$ is the minimizer of (8.52) (over such values of $\bar{p}_{1,j}(\delta)$ and $\bar{p}_{2,j}(\delta)$). In particular, $\bar{p}_{1,j}(0) = \bar{p}_{2,j}(0) = (\sum_{i=1}^{2} n_{ij}\bar{X}_{i,n_{ij}})/(n_{1j} + n_{2j})$. For $\delta = -\gamma$, the minimization problem leads to a nonlinear equation in $p$. Since $\gamma$ is typically small, we can replace it by the linear approximation

$$\bar{p}_{2,j}(-\gamma) = \left\{ n_{2j}\bar{X}_{2,n_{2j}} + n_{1j}\left(\bar{X}_{1,n_{1j}} + \gamma\right) \right\} / (n_{2j} + n_{1j}). \tag{8.53}$$

The GLR statistics (8.52) with $\delta = 0, -\gamma$ can be applied to the group sequential test of superiority and non-inferiority.

Wang et al. (2001) also assume that $M' < M$ and that the treatment responses are Bernoulli random variables. Instead of GLR statistics, they use the studentized statistics $Z_j = \hat{\triangle}_j/\hat{\sigma}_j$ prior to switching from the superiority to the non-inferiority alternative, where $\hat{\triangle}_j = \hat{p}_{1,j} - \hat{p}_{2,j}$ and

$$\hat{\sigma}_j^2 = \hat{p}_{1,j}\left(1 - \hat{p}_{1,j}\right)/n_{1j} + \hat{p}_{2,j}\left(1 - \hat{p}_{2,j}\right)/n_{2j}. \tag{8.54}$$

Moreover, in place of the modified Haybittle–Peto stopping boundaries, they use the O'Brien–Fleming error-spending function (see Sect. 4.1.3). A major difference between their approach and ours lies in how they switch from the superiority to the non-inferiority alternative during the course of the trial. Whereas our procedure makes the switch at the $k'$th interim analysis, where $2M' = 2n_{k'}$ is the maximum sample size for testing the superiority alternative $p_1 - p_2 = \tilde{\gamma}(> 0)$ at which the fixed sample size (FSS) GLR test attains some prescribed power, they make the switch at the first interim analysis at which the conditional power in favor of the non-inferiority alternative $p_1 - p_2 > -\gamma$ exceeds that of the superiority alternative $p_1 - p_2 > 0$. Specifically, at the $j$th interim analysis with $j < k$, consider the conditional probability $\mathrm{CP_S}(\triangle)$ at $p_1 - p_2 = \triangle$, given $(\hat{p}_{1,j}, \hat{p}_{2,j})$, that the FSS test with sample size $2M'$ rejects $H_0' : p_1 - p_2 \leq 0$. Also compute the conditional probability $\mathrm{CP_{NI}}(\triangle)$ at $p_1 - p_2 = \triangle$, given $(\hat{p}_{1,j}, \hat{p}_{2,j})$, that the FSS test with sample size $2M$ rejects $H_0 : p_1 - p_2 \leq -\gamma$ but accepts $H_0'$. The conditional power approach computes these conditional probabilities at $\triangle = \hat{\triangle}_j$, and the adaptive strategy in Wang et al. (2001) switches from the superiority alternative (with maximum sample size $2M'$) to the non-inferiority alternative (with maximum sample size $2M$) when $\mathrm{CP_S}(\hat{\triangle}_j) < \mathrm{CP_{NI}}(\hat{\triangle}_j)$. To circumvent the possibility of an inflated type I error probability due to such data-dependent switch, Wang et al. (2001, p. 1907–1908) modify both the times of interim analyses and the test statistics after the interim analysis at which such switch is made. Making use of previous work of Cui et al. (1999), they have shown that this modification indeed yields a type I error probability that is asymptotically no larger than $\alpha$. The modified statistics, however, are not sufficient statistics and are similar to those in Cui et al. (1999) which have been found to be inefficient. In contrast, the procedures of Lai et al. (2006c) use efficient GLR statistics for the switching, stopping, and terminal decision rules.

Wang et al. (2001) reported a simulation study demonstrating the advantages of their adaptive strategy over the FSS and some other group sequential methods. The superiority alternative in their study is at $p_2 = 0.25$, $p_1 = 0.35$, at which the level $\alpha = 0.025$ FSS test of $H_0' : p_1 - p_2 \leq 0$ with power 0.8 requires a sample size of $M' = 330$ from each population. The non-inferiority alternative is at $p_1 = p_2 = 0.25$, at which the level $\alpha = 0.025$ FSS test of $H_0 : p_1 - p_2 \leq -0.05$ with power 0.8 requires a sample size of $M = 1200$ from each population. Their group sequential tests involve $k = 5$ interim analyses, and of particular interest is their adaptive group sequential design that uses the above conditional power criterion to switch from the superiority to non-inferiority objective.

The simulation results of Wang et al. (2001) show that at the prespecified superiority alternative $(p_2 = 0.25, p_1 = 0.35)$, their adaptive procedure has power 0.947 and an expected sample size of 428 (which exceeds 330 for the FSS test) from each population. For comparison, we have computed corresponding operating characteristics of our group sequential test by Monte Carlo involving 50,000 simulations. The test of Lai et al. (2006c) (described in the first paragraph of this section with $\varepsilon = 1/3$) also uses $k = 5$ analyses, with $k' = 3$, $n_j - n_{j-1} = 110$ (for $j \leq 3$) or 435 (for $j = 4, 5$). It has type I error probability 0.024 of falsely rejecting $H_0'$ at $p_1 = p_2 = 0.25$, which is close to the corresponding value of 0.0258 for the test of Wang et al. (2001). At the superiority alternative $p_2 = 0.25$, $p_1 = 0.35$, it has power 0.785 (which is close to the target power 0.8, whereas the adaptive test of Wang et al. (2001) is substantially overpowered) and expected sample size of 299 from each population. At the prespecified non-inferiority alternative $(p_1 = p_2 = 0.25)$, the adaptive test of Wang et al. (2001) has power 0.768 and expected sample size 1144 (from each population), whereas that of Lai et al. (2006c) has power 0.784 and expected sample size 1013. The type I error probability of falsely rejecting $H_0 : p_1 - p_2 \leq -0.05$ at $p_2 = 0.25$, $p_1 = 0.2$ is 0.024 for the test of Lai et al. and 0.0253 for that of Wang et al. (2001).

### 8.4.3 Adaptive Choice Between One Superiority and Two Non-inferiority Trials

To begin with, we explain the background for the clinical trial mentioned in the first paragraph of Sect. 8.4. A pharmaceutical company that developed a new antimicrobial drug had planned to conduct two independent non-inferiority trials, as required by the FDA, with the same (fixed) sample size to demonstrate the drug's non-inferiority relative to an active control. This plan had to be reconsidered when the FDA required a substantially smaller non-inferiority margin than what was assumed in the plan. A major issue was whether the increased total sample size for the two non-inferiority trials due to the decreased non-inferior margin would already suffice to establish superiority of the drug since the FDA only required a single trial to demonstrate superiority. Following the notation of Sect. 8.4.2, the null hypothesis is $H_0 : p_1 - p_2 \leq -\gamma$ to test for non-inferiority and is $H_0' : p_1 - p_2 \leq 0$ to test for superiority, where $p_2$ is the response probability of the active control and $p_1$ is that of the new drug. The narrower non-inferiority margin required by the FDA was $\gamma = 0.1$, and the response rate $p_2$ of the active control was estimated to be 0.7 from previous studies. Thus, to have 90% power at the alternative $p_1 = p_2 (= 0.7)$, a level $\alpha = 0.025$ test of $H_0$ requires a sample size of 882 (i.e., 441 per treatment arm), which means a total sample size of 1764 for two non-inferiority trials.

Because it could only make rough a priori guesses of $p_1$ and was also concerned that the estimate 0.7 might differ substantially from the actual $p_2$, the pharmaceutical company was unable to decide whether it should perform two non-inferiority trials or one superiority trial with the same number of subjects. It would prefer to make that decision during the course of the trial when accumulating data would provide information on the feasibility of demonstrating non-inferiority or superiority of the new drug at the end of the trial. How should this be done without inflating the overall type I error? Moreover, 1764 was already considered to be somewhat too large for the total sample size because of the eligibility criterion that made it difficult to enroll subjects. The company, therefore, would also like to be able to terminate the study if interim analysis of the data would suggest "futility" of a trial with a maximum sample size of 1764. This led Lai et al. (2006c) to develop a group sequential design that can "self-tune" to the unknown $(p_1, p_2)$ and thereby choose adaptively among testing for superiority at level $\alpha$ (with no more than 1764 subjects), testing for non-inferiority (with two independent trials each of level $\alpha$, as required by the FDA), and early termination due to futility. The test statistics are the GLR statistics $\Lambda_j(\delta)$ in (8.52), with $\delta = 0$ for superiority and $\delta = -0.1$ for non-inferiority. The sequential design involves $k = 5$ groups with $n_1 = 300$ (or 150 patients per arm), $n_2 = 600$, $n_3 = 882$ (which is the fixed sample size of the non-inferiority trial), $n_4 = 1320$, and $n_5 = 1764$. The group sequential trial terminates early at the $j$th analysis with the superiority claim (rejecting $H_0'$) for $j \leq 4$ if

$$\hat{p}_{1,j} > p_{2,j} \quad \text{and} \quad \Lambda_j(0) \geq b'. \tag{8.55}$$

It can also terminate during the first two interim analyses due to futility (accepting $H_0$) if for $j = 1$ or 2,

$$\hat{p}_{1,j} < p_{2,j} \quad \text{and} \quad \Lambda_j(0) \geq \tilde{b}. \tag{8.56}$$

At the third interim analysis (with total sample size 882), assuming termination has not occurred, if $H_0'$ is not rejected, continue the trial if

$$\hat{p}_{1,3} - \hat{p}_{2,3} > -0.1 \quad \text{and} \quad \Lambda_3(-0.1) \geq b, \tag{8.57}$$

otherwise accept $H_0$ and stop (for futility). At the final analysis ($j = 5$), reject $H_0'$ (claiming superiority) if

$$(\hat{p}_{1,5} - \hat{p}_{2,5}) \big/ \hat{\sigma}_5 \geq c', \tag{8.58}$$

where $\hat{\sigma}_j^2$ is defined in (8.54), otherwise reject $H_0$ (claiming non-inferiority) only if

$$(\hat{p}_1 - \hat{p}_2 + 0.1) \big/ \hat{\sigma} \geq c, \tag{8.59}$$

where $\hat{p}_1 = \sum_{i=883}^{882+n_D} X_{1i}/n_D$, $\hat{p}_2 = \sum_{i=883}^{882+n_C} X_{2i}/n_C$, $\hat{\sigma}^2 = \hat{p}_1(1-\hat{p}_1)/n_D + \hat{p}_2(1-\hat{p}_2)/n_C$, and $n_D(n_C)$ denote the number of subjects receiving the new drug (active control) between the third and fifth analyses (representing the second non-inferiority trial required by the FDA) so that $n_D + n_C = 882$. Letting $0 < \varepsilon < \frac{1}{2}$, the thresholds $b'$, $\tilde{b}$, and then $b$, $c'$, and $c$ are determined by the equations

$$P_0\{(8.55) \text{ holds for some } j \leq 4\} = \varepsilon\alpha, \tag{8.60a}$$

$$P_0\{(8.56) \text{ holds for some } j \leq 2\} = \varepsilon\tilde{\alpha}, \tag{8.60b}$$

$$\begin{aligned} P_{-0.1}\{&\text{Test terminates at } j\text{th analysis with (8.55) for some } j \leq 3 \\ &\text{or continues at the third analysis with (8.57)}\} = \alpha, \end{aligned} \tag{8.60c}$$

$$P_0\{\text{Test terminates with (8.55) for some } j \leq 4, \text{ or (8.58) holds}\} = \alpha, \tag{8.60d}$$

$$P_{-0.1}\{(8.59) \text{ holds, or (8.55) holds for } j = 4\} = \alpha. \tag{8.60e}$$

From (8.60c)–(8.60e), it follows that $P_0(\text{Test claims superiority}) = \alpha$ and $P_{-0.1}(\text{Test claims either non-inferiority or superiority}) = \alpha$. Note that early stopping due to futility is incorporated in (8.60c)–(8.60e) for the determination of $b$, $c'$, and $c$. The major difference between this group sequential design and that in Sect. 8.4.2, where only one trial is needed to establish non-inferiority, is that the design has to allow the option of two independent non-inferiority trials required by the FDA for a non-inferiority claim, under the maximum sample size constraint of 1764. In this connection, note that (8.59) represents the rejection region of the second non-inferiority trial involving a new set of 882 subjects accrued after the third interim analysis. Since superiority testing (which is based on all subjects that have entered the trial) can still proceed after the third interim analysis, (8.60) gives the probability of a false positive claim in the second non-inferiority trial.

**Table 8.7** Type I error (in boldface) and power of FSS tests and the group sequential test that chooses adaptively between one superiority and two non-inferiority trials

| | | ModHP | | | | $\text{FSS}_1$ | $\text{FSS}_2$ |
|---|---|---|---|---|---|---|---|
| $p_2$ | $p_1$ | P(NI) | P(S) | P(+) | $E(T)$ | P(S) | P($\text{NI}^2$) |
| 0.70 | 0.70 | 0.786 | **0.025** | 0.811 | 1664 | **0.025** | 0.810 |
| | 0.60 | 0.001 | 0.000 | **0.001** | 693 | 0.000 | **0.001** |
| | 0.65 | 0.124 | 0.000 | 0.124 | 1144 | 0.000 | 0.125 |
| | 0.73 | 0.706 | 0.275 | 0.981 | 1680 | 0.287 | 0.980 |
| | 0.75 | 0.363 | 0.634 | 0.997 | 1529 | 0.653 | 0.998 |
| | 0.77 | 0.094 | 0.906 | 1.000 | 1251 | 0.916 | 1.000 |
| | 0.80 | 0.002 | 0.998 | 1.000 | 792 | 0.998 | 1.000 |
| 0.60 | 0.60 | 0.709 | **0.025** | 0.734 | 1626 | **0.025** | 0.736 |
| | 0.50 | 0.001 | 0.000 | **0.001** | 713 | 0.000 | **0.001** |
| | 0.55 | 0.108 | 0.000 | 0.108 | 1130 | 0.000 | 0.107 |
| | 0.63 | 0.712 | 0.244 | 0.956 | 1679 | 0.253 | 0.956 |
| | 0.65 | 0.429 | 0.563 | 0.992 | 1569 | 0.584 | 0.992 |
| | 0.67 | 0.150 | 0.849 | 0.999 | 1339 | 0.864 | 0.999 |
| | 0.70 | 0.008 | 0.992 | 1.000 | 911 | 0.993 | 1.000 |
| 0.80 | 0.80 | 0.895 | **0.025** | 0.920 | 1715 | **0.025** | 0.922 |
| | 0.70 | 0.001 | 0.000 | **0.001** | 646 | 0.000 | **0.001** |
| | 0.75 | 0.183 | 0.000 | 0.183 | 1198 | 0.000 | 0.186 |
| | 0.83 | 0.645 | 0.351 | 0.997 | 1662 | 0.368 | 0.997 |
| | 0.85 | 0.225 | 0.775 | 1.000 | 1425 | 0.791 | 1.000 |
| | 0.87 | 0.025 | 0.975 | 1.000 | 1039 | 0.978 | 1.000 |
| | 0.90 | 0.000 | 1.000 | 1.000 | 584 | 1.000 | 1.000 |

Table 8.7 gives the results of a simulation study on this group sequential design with $\alpha = 0.025$ and $\varepsilon = 1/3$. Each result is based on 50,000 simulations. The expected sample size $E(T)$ and the probabilities $P(\text{S})$ of claiming superiority, $P(\text{NI})$ of claiming non-inferiority (but not superiority), and $P(+) = P(\text{S}) + P(\text{NI})$ of a positive claim for the new drug are given for a variety of parameter configurations. Note that $P(\text{NI}) = P\{\text{Test terminates at the fifth analysis with rejection of } H_0 \text{ because of (8.59)}\}$, so the non-inferiority claim is supported by the first set of 882 subjects who result in (8.57) at the third analysis and by a second set of 882 subjects yielding (8.59) at the fifth analysis. Also given in Table 8.7 are (a) the power $P(\text{S})$ of a level-$\alpha$ FSS test of $H_0'$ with sample size 1764, denoted by $\text{FSS}_1$, and (b) the power $P(\text{NI}^2)$ of two independent level-$\alpha$ FSS tests of $H_0$, denoted by $\text{FSS}_2$, with sample size 882 for each test so that $\text{NI}^2$ represents non-inferiority claims for both tests. Table 8.7 shows that our group sequential design has markedly smaller expected sample size than 1764 while having better power than $\text{FSS}_1$ or $\text{FSS}_2$. Although $P(+)$ and $P(\text{NI}^2)$ seem to differ little, note that $P(+) = P(\text{S}) + P(\text{NI})$ and that the probability $P(\text{S})$ of a superiority claim by the group sequential test can be quite high, whereas $\text{FSS}_2$ can only claim non-inferiority with probability $P(\text{NI}^2)$.

### 8.4.4 Discussion

A major drawback of the commonly used conditional power approach to two-stage designs is pointed out in Sect. 8.3.2. The actual power can be much lower than the conditional power since the estimated alternative at the end of the first stage can be quite different from the actual alternative. In particular, if the estimated alternative falls in the region of the null hypothesis and misleads one to stop for futility, there can be substantial loss of power. On the other hand, early stopping for futility is critical for keeping the sample size of a conditional power test within a manageable bound $M$. The three-stage test in Sect. 8.2.1 makes use of $M$ to come up with an implied alternative which is used to choose the rejection and futility boundaries appropriately so that the test does not lose much power in comparison with the (most powerful) fixed sample size test of the null hypothesis versus the implied alternative. This idea underlying (8.15)–(8.17) that define the stopping boundaries of three-stage tests is similar to that underlying efficient group sequential tests in Sect. 4.2.

## 8.5 Supplements and Problems

1. Verify that if $X_1, X_2, \ldots$ are i.i.d. $N(\theta, 1)$ random variables, then equations (8.15)–(8.17) can be written as (8.28)–(8.30), respectively.
2. Verify formulas (8.31), (8.32), and (8.33).
3. Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be independent normal observations with unknown means $\mu_X, \mu_Y$ and common unknown variance $\sigma^2$.

   (a) Find an expression for the Kullback–Leibler information number

   $$I\left((\mu_X, \mu_Y, \sigma^2), (\tilde{\mu}_X, \tilde{\mu}_Y, \tilde{\sigma}^2)\right)$$

   for arbitrary $\mu_X, \mu_Y, \sigma^2, \tilde{\mu}_X, \tilde{\mu}_Y, \tilde{\sigma}^2$.

   (b) For a given constant $\delta$, find an expression for

   $$\inf_{(\tilde{\mu}_X, \tilde{\mu}_Y, \tilde{\sigma}): \tilde{\mu}_X - \tilde{\mu}_Y = \delta} I\left((\mu_X, \mu_Y, \sigma^2), (\tilde{\mu}_X, \tilde{\mu}_Y, \tilde{\sigma}^2)\right)$$

   for arbitrary $\mu_X, \mu_Y, \sigma^2$.

4. In the setting of two binomial populations in Sect. 8.3.4, verify that the test statistic $\inf_{\boldsymbol{\theta}: u(\boldsymbol{\theta}) = \delta} nI(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\theta})$ takes the form (8.40).
5. *Adaptation beyond sample size re-estimation*

   Chapter 5 of Berry et al. (2011) gives a comprehensive overview of Bayesian design and analysis of confirmatory Phase III trials. It uses the posterior probability of the trial ending with a beneficial claim for the treatment, based on the data available at interim analysis, to determine adaptively the additional sample size, which may be 0 if the posterior probability is sufficiently high (for efficacy stopping) or low (for futility stopping). Thus, it is similar to the conditional power approach reviewed in Sect. 6.2.3, except that it uses posterior probability instead of conditional power. The Bayesian approach via posterior probabilities is also used to select arms in an adaptive multiarmed trial that starts with multiple treatment arms and makes a midcourse decision concerning which arm is appropriate to carry forward for confirmatory testing. "This type of confirmatory trial is referred to as a seamless Phase II/III trial," and the Bayesian design is "prospectively" adaptive in the sense that the design is completely specified in the protocol before the start of the trial, and although the interim results can change the trial's features, the changes are "by design, not ad hoc retrospective changes." (Berry et al. 2011, p. 194–195). Monte Carlo simulations of the frequentist type I error probabilities are used to determine the threshold for early stopping to claim efficacy of the new treatment. These simulations are carried out at certain parameter values belonging to the null hypothesis. However, as pointed out in Sect. 1.5, there is no guarantee that the type I error is actually maintained since the null hypothesis is highly composite. Frequentist approaches to adaptive designs of confirmatory trials with interim arm selection have also appeared in the past decade; see Bretz et al. (2006) for a review. Like adaptive sample size re-estimation reviewed in Sect. 8.1.2, these methods control the type I error probability by using inefficient test statistics that are similar to (8.5). By making use of techniques from multiarmed bandit theory which Sect. 1.5 has alluded to, Lai and Liao (2012) have recently developed asymptotic lower bounds for the expected sample sizes from the respective arms, subject to type I error and power constraints, and have developed adaptive allocation rules and sequential GLR tests that achieve these bounds. Fully sequential rules are used, similar to those in Chap. 3, and their group sequential or multistage modifications, similar to those in Sects. 4.2 and 8.2, are under investigation.

# References

Andersen PK, Borgan Ø, Gill RD, Keiding N (1993) Statistical models based on counting processes. Springer Series in Statistics, Springer-Verlag, New York, DOI 10.1007/978-1-4612-4348-9

Anscombe F (1952) Large-sample theory of sequential estimation. Math Proc Cambridge Philosoph Soc 48(04):600–607

Armitage P (1960) Sequential medical trials. Charles C. Thomas, Springfield, Ill

Armitage P, McPherson CK, Rowe BC (1969) Repeated significance tests on accumulating data. J Roy Statist Soc Ser A 132:235–244

Azriel D (2012) Optimal sequential designs in phase I studies. Comput Stat Data Anal To appear

Babb J, Rogatko A, Zacks S (1998) Cancer phase I clinical trials: efficient dose escalation with overdose control. Stat Med 17(10):1103–1120

Barnard GA (1959) Control charts and stochastic processes (with discussion). J Roy Statist Soc Ser B 21:239–271

Bartroff J, Lai TL (2008a) Efficient adaptive designs with mid-course sample size adjustment in clinical trials. Stat Med 27(10):1593–1611, DOI 10.1002/sim.3201

Bartroff J, Lai TL (2008b) Generalized likelihood ratio statistics and uncertainty adjustments in efficient adaptive design of clinical trials. Sequent Anal 27(3):254–276

Bartroff J, Lai TL (2010) Approximate dynamic programming and its applications to the design of phase I cancer trials. Statist Sci 25:245–257

Bartroff J, Lai TL (2011) Incorporating individual and collective ethics into phase I cancer trial designs. Biometrics 67(2):596–603, DOI 10.1111/j.1541-0420.2010.01471.x

Bauer P (1989) Multistage testing with adaptive designs. Biometrie und Informatik in Medizin und Biologie 20:130–148, with discussion

Bauer P, Köhne K (1994) Evaluation of experiments with adaptive interim analyses. Biometrics 50(4):1029–1041, DOI 10.2307/2533441

Bayard DS (1991) A forward method for optimal stochastic nonlinear and adaptive control. IEEE Trans Automat Contrl 36(9):1046–1053, DOI 10.1109/9.83535

Beal, S.L. and Sheiner, L.B., eds. (1992), NONMEM User.s Guide, University of California, San Francisco, NONMEM Project Group.

Berry SM, Carlin BP, Lee JJ, Müller P (2011) Bayesian adaptive methods for clinical trials. Chapman & Hall/CRC biostatistics series, vol 38. CRC, Boca Raton

Bertsekas DP (2007) Dynamic programming and optimal control, vol II, 3rd edn. Athena Scientific Belmont, Masssachusetts

Beta-Blocker Heart Attack Trial Research Group (1982) A randomized trial of propranolol in patients with acute myocardial infarction: 1. mortality results. J Am Med Assoc 247(12):1707–1714

J. Bartroff et al., *Sequential Experimentation in Clinical Trials,*
Springer Series in Statistics 298, DOI 10.1007/978-1-4614-6114-2,

Beta-Blocker Heart Attack Trial Research Group (1984) Beta-blocker heart attack trial: design, methods, and baseline results. Contr Clin Trial 5(4):382–437, DOI 10.1016/S0197-2456(84)80017-3

Birkett M, Day S (1994) Internal pilot studies for estimating sample size. Stat Med 13(23–24):2455–2463, DOI 10.1002/sim.4780132309

Boyd S. and Vandenberghe L. (2004), “Convex Optimization”, Cambridge University Press, New York.

Brensike JF, Kelsey SF, Passamani ER, Fisher MR, Richardson JM, Loh IK, Stone NJ, Aldrich RF, Battaglini JW, Moriarty DJ, Myrianthopoulos MB, Detre KM, Epstein SE, Levy RI (1982) National heart lung and blood institute type ii coronary intervention study: design, methods and baseline characteristics. Contr Clin Trial 3(2):91–111, DOI 10.1016/0197-2456(82)90038-1

Breslow NE, Clayton DG (1993) Approximate inference in generalized linear mixed models. J Am Statist Assoc 88(421):9–25

Bretz F, Schmidli H, König F, Racine A, Maurer W (2006) Confirmatory seamless phase II/III clinical trials with hypothesis selection at interim: general concepts. Biomet J 4:623–634, DOI 10.1002/bimj.200510232

Brookmeyer R, Crowley J (1982) A confidence interval for the median survival time. Biometrics 38:29–41

Burman CF, Sonesson C (2006) Are flexible designs sound? Biometrics 62(3):664–669, DOI 10.1111/j.1541-0420.2006.00626.x

Chan HP, Lai TL (2000) Asymptotic approximations for error probabilities of sequential or fixed sample size tests in exponential families. Ann Statist 28(6):1638–1669

Chang M (2007) Adaptive design theory and implementation using SAS and R. Chapman & Hall/CRC, Boca Raton

Chen RT, Davis RL, Rhodes PH (2005) Special methodological issues in pharmacoepidemiology studies of vaccine safety. In: Strom B (ed) Pharmacoepidemiology, 4th edn. Wiley, New York, pp 455–486

Chernoff H (1961) Sequential tests for the mean of a normal distribution. In: Proceedings of the 4th Berkeley symposium on mathematical statistics and probability, vol I, University of California Press, Berkeley, California, pp 79–91

Chernoff H (1965) Sequential tests for the mean of a normal distribution III (small $t$). Ann Math Statist 36:28–54

Chernoff H (1972) Sequential analysis and optimal design. Society for Industrial and Applied Mathematics, Philadelphia

Choi S, Smith P, Becker D (1985) Early decision in clinical trials when the treatment differences are small: experience of a controlled trial in head trauma. Contr Clin Trial 6(4):280–288, DOI 10.1016/0197-2456(85)90104-7

Chow SC, Chang M (2006) Adaptive design methods in clinical trials. Chapman & Hall/CRC, Boca Raton

Chuang CS, Lai TL (1998) Resampling methods for confidence intervals in group sequential trials. Biometrika 85(2):317–332

Chuang CS, Lai TL (2000) Hybrid resampling methods for confidence intervals. Statist Sinica 10(1):1–50, with discussion and rejoinder by the authors

Cook A, Tiwari R, Wellman R, Heckbert S, Li L, Heagerty P, Marsh T, Nelson J (2012) Statistical approaches to group sequential monitoring of postmarket safety surveillance data: current state of the art for use in the mini-sentinel pilot. Pharmacoepidemiol Drug Safety 21:72–81

Cox DR (1972) Regression models and life-tables. J Roy Statist Soc Ser B 34:187–220, with discussion and a reply by the author

Cox DR (1975) Partial likelihood. Biometrika 62(2):269–276

Cui L, Hung HMJ, Wang SJ (1999) Modification of sample size in group sequential clinical trials. Biometrics 55(3):853–857

Dantzig GB (1940) On the non-existence of tests of “student’s” hypothesis having power functions independent of $\sigma$. Ann Math Statist 11:186–192

Davidian M, Giltinan DM (1995) Nonlinear models for repeated measurement data. Monographs on statistics & applied probability. Chapman and Hall/CRC, Boca Raton

Davis R, Kolczak M, Lewis E, Nordin J, Goodman M, Shay D, Platt R, Black S, Shinefield H, Chen R (2005) Active surveillance of vaccine safety – a system to detect early signs of adverse events. Epidemiology 16(3):336–341, DOI 10.1097/01.ede.0000155506.05636.a4

Denne JS (2001) Sample size recalculation using conditional power. Stat Med 20(17–18):2645–2660, DOI 10.1002/sim.734

Denne JS, Jennison C (1999) Estimating the sample size for a $t$-test using an internal pilot. Stat Med 18(13):1575–1585

Denne JS, Jennison C (2000) A group sequential $t$-test with updating of sample sizes. Biometrika 87(1):125–134, DOI 10.1093/biomet/87.1.125

Dixon WJ, Mood AM (1948) A method for obtaining and analyzing sensitivity data. J Am Statist Assoc 43(241):109–126

Durham SD, Flournoy N (1994) Random walks for quantile estimation. In: Statistical decision theory and related topics, V (West Lafayette, IN, 1992). Springer, New York, pp 467–476

Dvoretzky A, Kiefer J, Wolfowitz J (1953) Sequential decision problems for processes with continuous time parameter. Ann Math Statist 24:403–415

Eales JD, Jennison C (1992) An improved method for deriving optimal one-sided group sequential tests. Biometrika 79(1):13–24

Eales JD, Jennison C (1995) Optimal two-sided group sequential tests. Sequent Anal 14(4):273–286

Efron B (1987) Better bootstrap confidence intervals. J Am Statist Assoc 82(397):171–200, with comments and a rejoinder by the author

Eisenhauer EA, O'Dwyer PJ, Christian M, Humphrey JS (2000) Phase I clinical trial design in cancer drug development. J Clin Oncol 18:684

Elfving G (1952) Optimum allocation in linear regression theory. Ann Math Statist 23:255–262

Ellenberg SS, Foulkes MA, Midthun K, Goldenthal KL (2005) Evaluating the safety of new vaccines: summary of a workshop. Am J Public Health 95(5):800–807, DOI 10.2105/AJPH.2004.039438

Emerson SS, Fleming TR (1989) Symmetric group sequential test designs. Biometrics 45(3):905–923

Emerson SS, Fleming TR (1990) Parameter estimation following group sequential hypothesis testing. Biometrika 77(4):875–892, DOI 10.1093/biomet/77.4.875

Facey KM (1992) A sequential procedure for a phase II efficacy trial in hypercholesterolemia. Contr Clin Trial 13(2):122–133, DOI 10.1016/0197-2456(92)90018-U

Fan J (1991) On the optimal rates of convergence for nonparametric deconvolution problems. Ann Statist 19(3):1257–1272, DOI 10.1214/aos/1176348248

Fedorov VV (1972) Theory of optimal experiments. In: Studden WJ, Klimko EM (eds) Probability and mathematical statistics, vol 12. Academic, New York, translated from the Russian

Fisher LD (1998) Self-designing clinical trials. Stat Med 17(14):1551–1562

Gehan EA (1965) A generalized wilcoxon test for comparing arbitrarily singly censored samples. Biometrika 52:203–223

Gould AL, Shih WJ (1992) Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. Commun Statist A – Theor Method 21(10):2833–2853, DOI 10.1080/03610929208830947

Greene SK, Kulldorff M, Yin R, Yih WK, Lieu TA, Weintraub ES, Lee GM (2011) Near real-time vaccine safety surveillance with partially accrued data. Pharmacoepidemiol Drug Safety 20:583–590

Gross ST, Lai TL (1996) Bootstrap methods for truncated and censored data. Statist Sinica 6(3):509–530

Gu M, Lai TL (1991) Weak convergence of time-sequential censored rank statistics with applications to sequential testing in clinical trials. Ann Statist 19(3):1403–1433

Gu M, Lai TL (1998) Repeated significance testing with censored rank statistics in interim analysis of clinical trials. Statist Sinica 8(2):411–428

Gu M, Lai TL (1999) Determination of power and sample size in the design of clinical trials with failure-time endpoints and interim analyses. Contr Clin Trials 20(5):423–438, DOI 10.1016/S0197-2456(99)00021-5

Gu M, Lai TL, Lan KKG (1991) Rank tests based on censored data and their sequential analogues. Am J Math Manag Sci 11(1–2):147–176

Haines LM, Perevozskaya I, Rosenberger WF (2003) Bayesian optimal designs for phase I clinical trials. Biometrics 59(3):591–600, DOI 10.1111/1541-0420.00069

Hall P (1992) The Bootstrap and edgeworth expansion. Springer, New York

Hall WJ, Yakir B (2003) Inference about a secondary process following a sequential trial. Biometrika 90(3):597–611, DOI 10.1093/biomet/90.3.597

Han J, Lai TL, Spivakovsky V (2006) Approximate policy optimization and adaptive control in regression models. Comput Econ 27:433–452

Harrington DP, Fleming TR (1982) A class of rank test procedures for censored survival data. Biometrika 69(3):553–566, DOI 10.1093/biomet/69.3.553

Haybittle J (1971) Repeated assessment of results in clinical trials of cancer treatment. Brit J Radiol 44(526):793–&

He P, Lai TL, Liao OY (2012) Futility stopping in clinical trials. Stat Interface To appear

Herson J (1979) Predictive probability early termination plans for phase II clinical trials. Biometrics 35(4):775–783, DOI 10.2307/2530109

Herson J, Wittes JT (1993) The use of interim analysis in sample size adjustment. Drug Inf J 27:753–760

Heyse JF, Kuter BJ, Dallas MJ, Heaton P, REST Study Team (2008) Evaluating the safety of a rotavirus vaccine: the REST of the story. Clin Trials 5(2):131–139, DOI 10.1177/1740774508090507

Hoeffding W (1960) Lower bounds for the expected sample size and the average risk of a sequential procedure. Ann Math Statist 31:352–368

Huang X, Ning J, Li Y, Estey E, Issa JP, Berry DA (2009) Using short-term response information to facilitate adaptive randomization for survival clinical trials. Stat Med 28(12):1680–1689, DOI 10.1002/sim.3578

Inoue LYT, Thall PF, Berry DA (2002) Seamlessly expanding a randomized Phase II trial to Phase III. Biometrics 58:823–831

Jennison C, Turnbull BW (1990) Statistical approaches to interim monitoring of medical trials: a review and commentary. Statist Sci 5(3):299–317

Jennison C, Turnbull BW (2000) Group sequential methods with applications to clinical trials. Chapman & Hall/CRC, Boca Raton

Jennison C, Turnbull BW (2003) Mid-course sample size modification in clinical trials based on the observed treatment effect. Stat Med 22(6):971–993, DOI 10.1002/sim.1457

Jennison C, Turnbull BW (2006a) Adaptive and nonadaptive group sequential tests. Biometrika 93(1):1–21, DOI 10.1093/biomet/93.1.1

Jennison C, Turnbull BW (2006b) Efficient group sequential designs when there are several effect sizes under consideration. Stat Med 25(6):917–932, DOI 10.1002/sim.2251

Johnson RA (1970) Asymptotic expansions associated with posterior distributions. Ann Math Statist 41:851–864

Julious S, Swank D (2005) Moving statistics beyond the individual clinical trial: applying decision science to optimize a clinical development plan. Pharmaceut Statist 4(1):37–46, DOI 10.1002/pst.149

Katzung BG (1995) Basic and clinical pharmacology, 6th edn. Prentice Hall, Englewood, Cliffs

Kiefer J (1974) General equivalence theory for optimum designs (approximate theory). Ann Statist 2:849–879

Kiefer J, Weiss L (1957) Some properties of generalized sequential probability ratio tests. Ann Math Statist 28:57–74

Kim K, DeMets DL (1987) Design and analysis of group sequential tests based on the type I error spending rate function. Biometrika 74(1):149–154

Korn EL, Midthune D, Chen TT, Rubinstein LV, Christian MC, Simon RM (1994) A comparison of two phase I trial designs. Stat Med 13(18):1799–1806, DOI 10.1002/sim.4780131802

Kulldorff M, Davis R, Kolczak M, Lewis E, Lieu T, Platt R (2011) A maximized sequential probability ratio test for drug and vaccine safety surveillance. Sequent Anal 30(1):58–78

Lai TL (1973) Optimal stopping and sequential tests which minimize the maximum expected sample size. Ann Statist 1:659–673

Lai TL (1988) Nearly optimal sequential tests of composite hypotheses. Ann Statist 16(2):856–886

Lai TL (1995) Sequential changepoint detection in quality control and dynamical systems. J Roy Statist Soc Ser B 57(4):613–658, with discussion and a reply by the author

Lai TL (1997) On optimal stopping problems in sequential hypothesis testing. Statist Sinica 7(1):33–51

Lai TL (1998) Information bounds and quick detection of parameter changes in stochastic systems. IEEE Trans Inform Theory 44:2917–2929

Lai TL (2001) Sequential analysis: some classical problems and new challenges. Statist Sinica 11(2):303–408, with comments and a rejoinder by the author

Lai TL (2004) Likelihood ratio identities and their applications to sequential analysis. Sequent Anal 23(4):467–497

Lai TL, Li W (2006) Confidence intervals in group sequential trials with random group sizes and applications to survival analysis. Biometrika 93(3):641–654

Lai TL, Liao OY (2012) Efficient adaptive randomization and stopping rules in multi-arm clinical trials for testing a new treatment. Sequent Anal To appear

Lai TL, Robbins H (1979) Adaptive design and stochastic approximation. Ann Statist 7(6):1196–1221

Lai TL, Shih MC (2003a) A hybrid estimator in nonlinear and generalised linear mixed effects models. Biometrika 90(4):859–879

Lai TL, Shih MC (2003b) Nonparametric estimation in nonlinear mixed effects models. Biometrika 90(1):1–13

Lai TL, Shih MC (2004) Power, sample size and adaptation considerations in the design of group sequential clinical trials. Biometrika 91(3):507–528

Lai TL, Su Z (2006) Confidence intervals for survival quantiles in the cox regression model. Lifetime Data Anal 12:407–419

Lai TL, Wang JQ (1994) Asymptotic expansions for the distributions of stopped random walks and first passage times. Ann Probab 22(4):1957–1992

Lai TL, Xing H (2010) Sequential change-point detection when the pre- and post-change parameters are unknown. Sequent Anal 29:162–175

Lai TL, Zhang L (1994) A modification of schwarz's sequential likelihood ratio tests in multivariate sequential analysis. Sequent Anal 13(2):79–96

Lai T, Su Z, Chuang C (2006a) Bias correction and confidence intervals following sequential tests. In: Lecture Notes-Monograph Series, JSTOR, pp 44–57

Lai TL, Shih MC, Wong S (2006b) A new approach to modeling covariate effects and individualization in population pharmacokinetics-pharmacodynamics. J Pharmacok Pharmacodynam 33:49–74, DOI 10.1007/s10928-005-9000-2

Lai TL, Shih MC, Zhu G (2006c) Modified Haybittle-Peto group sequential designs for testing superiority and non-inferiority hypotheses in clinical trials. Stat Med 25(7):1149–1167

Lai TL, Shih MC, Su Z (2009) Tests and confidence intervals for secondary endpoints in sequential clinical trials. Biometrika 96:903–915, DOI 10.1093/biomet/asp063

Lai TL, Lavori PW, Shih MC (2012a) Sequential design of phase II–III cancer trials. Stat Med 31(18):1944–1960, DOI 10.1002/sim.5346

Lai TL, Lavori PW, Shih MC, Sikic BI (2012b) Clinical trial designs for testing biomarker-based personalized therapies. Clin Trials 9(2):141–154, DOI 10.1177/1740774512437252

Lai TL, Liao OY, Zhu G (2012c) Adaptation in clinical development plans and adaptive clinical trial designs. Stat Interface To appear

Lan KKG, DeMets DL (1983) Discrete sequential boundaries for clinical trials. Biometrika 70(3):659–663

Lan KKG, Wittes J (1988) The B-value: a tool for monitoring data. Biometrics 44(2):579–585, DOI 10.2307/2531870

Lan KKG, Simon R, Halperin M (1982) Stochastically curtailed tests in long-term clinical trials. Comm Statist C—Sequent Anal 1(3):207–219

Lehmacher W, Wassmer G (1999) Adaptive sample size calculations in group sequential trials. Biometrics 55(4):1286–1290, DOI 10.1111/j.0006-341X.1999.01286.x

Li G, Shih WJ (2002) A sample size adjustment procedure for clinical trials based on conditional power. Biostatistics 3(2):277–287, DOI 10.1093/biostatistics/3.2.277

Li L (2009) A conditional sequential sampling procedure for drug safety surveillance. Stat Med 28(25):3124–3138, DOI 10.1002/sim.3689

Li L, Kulldorff M (2010) A conditional maximized sequential probability ratio test for pharmacovigilance. Stat Med 29(2):284–295

Lieu TA, Kulldorfif M, Davis RL, Lewis EM, Weintraub E, Yih K, Yin R, Brown JS, Platt R (2007) Real-time vaccine safety surveillance for the early detection of adverse events. Med Care 45(10, Suppl. 2):S89–S95

Lindstrom MJ, Bates DM (1990) Nonlinear mixed effects models for repeated measures data. Biometrics 46(3):673–687, DOI 10.2307/2532087

Liu A, Hall WJ (1999) Unbiased estimation following a group sequential test. Biometrika 86(1):71–78, DOI 10.1093/biomet/86.1.71

Liu A, Tan M, Boyett JM, Xiong XP (2000) Testing secondary hypotheses following sequential clinical trials. Biometrics 56(2):640–644

Liu Q, Pierce DA (1994) A note on Gauss-Hermite quadrature. Biometrika 81(3):624–629, DOI 10.1093/biomet/81.3.624

Lorden G (1971) Procedures for reacting to a change in distribution. Ann Math Statist 42:1897–1908

Lorden G (1976) 2-SPRT's and the modified Kiefer-Weiss problem of minimizing an expected sample size. Ann Statist 4(2):281–291

Lorden G (1983) Asymptotic efficiency of three-stage hypothesis tests. Ann Statist 11(1):129–140

Mantel N (1966) Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemotherap Rep 50(3):163–170

Müller HG, Wang JL (1994) Hazard rate estimation under random censoring with varying kernels and bandwidths. Biometrics 50:61–76

Murphy T, Gargiullo P, Massoudi M, Nelson D, Jumaan A, Okoro C, Zanardi L, Setia S, Fair E, LeBaron C, Schwartz B, Wharton M, Livingood J (2001) Intussusception among infants given an oral rotavirus vaccine. N Engl J Med 344(8):564–572

Murphy T, Smith P, Gargiullo P, Schwartz B (2003) The first rotavirus vaccine and intussusception: epidemiological studies and policy decisions. J Infec Dis 187(8):1309–1313

Nelson J, Cook A, Yu O, Dominguez C, Zhao S, Greene S, Fireman B, Jacobsen S, Weintraub E, Jackson L (2012) Challenges in the design and analysis of sequentially monitored postmarket safety surveillance evaluations using electronic observational health care data. Pharmacoepidemiol Drug Safety 21:62–71

Newlands E, Blackledge G, Slack J, Rustin G, Smith D, Stuart N, Quarterman C, Hoffman R, Stevens M, Brampton M, Gibson A (1992) Antitumor imidazotetrazines .26. phase I trial of temozolomide (ccrg-81045, m-and-b 39831, nsc-362856). Br J Cancer 65(2):287–291, DOI 10.1038/bjc.1992.57

Nguyen M, Ball R, Midthun K, Lieu T (2012) The food and drug administration's postlicensure rapid immunization safety monitoring program: strengthening the federal vaccine safety enterprise. Pharmacoepidemiol Drug Safety 21:291–297

O'Brien PC, Fleming TR (1979) Multiple testing procedure for clinical trials. Biometrics 35(3):549–556

O'Quigley J (2002) Continual reassessment designs with early termination. Biostatistics 3(1):87–99, DOI 10.1093/biostatistics/3.1.87

O'Quigley J, Pepe M, Fisher L (1990) Continual reassessment method: a practical design for phase I clinical trials in cancer. Biometrics 46(1):33–48, DOI 10.2307/2531628

Page ES (1954) Continuous inspection schemes. Biometrika 41:100–115

Pampallona S, Tsiatis AA (1994) Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. J Statist Plan Infer 42(1–2):19–35

Pepe MS, Anderson GL (1992) Two-stage experimental designs: early stopping with a negative result. J Roy Statist Soc Ser C 41(1):181–190, DOI 10.2307/2347627

Peto R, Peto J (1972) Asymptotically efficient rank invariant test procedures. J Roy Statist Soc Ser A 135(2):185–207

Peto R, Pike M, Armitage P, Breslow N, Cox D, Howard S, Mantel N, McPherson K, Peto J, Smith P (1976) Design and analysis of randomized clinical trials requiring prolonged observation of each patient 1. Introduction and design. Br J Cancer 34(6):585–612

Pinheiro JC, Bates DM (1995) Approximations to the log-likelihood function in the nonlinear mixed-effects model. J Comput Graph Stat 4(1):12–35

Platt R, Carnahan R, Brown J, Chrischilles E, Curtis L, Hennessy S, Nelson J, Racoosin J, Robb M, Schneeweiss S et al (2012) The us food and drug administration's mini-sentinel program: status and direction. Pharmacoepidemiol Drug Safety 21:1–8

Pocock SJ (1977) Group sequential methods in the design and analysis of clinical trials. Biometrika 64(2):191–199

Pollak M (1985) Optimal detection of a change in distribution. Ann Statist 13:206–227

Posch M, Bauer P (1999) Adaptive two stage designs and the conditional error function. Biom J 41(6):689–696

Prentice RL (1978) Linear rank tests with right censored data. Biometrika 65(1):167–179, DOI 10.1093/biomet/65.1.167

Press NH, Flannery BP, Teukolsky SA, Vitterling WT (1992) Numerical recipes in C: the art of scientific computing, 2nd edn. Cambridge University Press, Cambridge

Proschan MA, Hunsberger SA (1995) Designed extension of studies based on conditional power. Biometrics 51(4):1315–1324, DOI 10.2307/2533262

Proschan MA, Lan KG, Wittes JT (2006) Statistical monitoring of clinical trials: a unified approach. Springer, New York, DOI 10.1007/978-0-387-44970-8

Reiner E, Paoletti X, O'Quigley J (1999) Operating characteristics of the standard phase I clinical trial design. Comput Stat Data Anal 30(3):303–315, DOI 10.1016/S0167-9473(98)00095-4

Robbins H, Monro S (1951) A stochastic approximation method. Ann Math Statist 22:400–407

Roberts SW (1966) A comparison of some control chart procedures. Technometrics 8:411–430

Rosner GL, Tsiatis AA (1988) Exact confidence-intervals following a group sequential trial – a comparison of methods. Biometrika 75(4):723–729

Rowland M, Tozer TN (1989) Clinical pharmacokinetics: concepts and applications, 2nd edn. Lea and Febiger Phildelphia

Rubinstein L, Crowley J, Ivy P, LeBlanc M, Sargent D (2009) Randomized phase II designs. Clin Cancer Res 15:1883–1890, DOI 10.1158/1078-0432.CCR-08-2031

Sacks J (1958) Asymptotic distribution of stochastic approximation procedures. Ann Math Statist 29:373–405

Scharfstein DO, Tsiatis AA (1998) The use of simulation and bootstrap in information-based group sequential studies. Stat Med 17(1):75–87

Scharfstein DO, Tsiatis AA, Robins JM (1997) Semiparametric efficiency and its implication on the design and analysis of group-sequential studies. J Am Statist Assoc 92(440):1342–1350, DOI 10.2307/2965404

Schmitz N (1993) Optimal sequentially planned decision procedures. Lecture notes in statistics, vol 79. Springer, New York

Schwarz G (1962) Asymptotic shapes of Bayes sequential testing regions. Ann Math Statist 33:224–236

Shen Y, Fisher L (1999) Statistical inference for self-designing clinical trials with a one-sided hypothesis. Biometrics 55(1):190–197, DOI 10.1111/j.0006-341X.1999.00190.x

Shewhart WA (1931) Economic control of manufactured products. Van Nostrand Reinhold, New York

Shih MC, Lai TL, Heyse JF, Chen J (2010) Sequential generalized likelihood ratio tests for vaccine safety evaluation. Stat Med 29(26):2698–2708, DOI 10.1002/sim.4036

Shih WJ (2001) Sample size re-estimation: Journey for a decade. Stat Med 20(4):515–518, DOI 10.1002/sim.532

Siegmund D (1978) Estimation following sequential tests. Biometrika 65(2):341–349, DOI 10.2307/2335213

Siegmund D (1985) Sequential analysis. Springer, New York

Siegmund D, Venkatraman ES (1995) Using the generalized likelihood ratio statistics for sequential detection of a change-point. Ann Statist 23:255–271

Silvey SD (1980) Optimal design. Monographs on applied probability and statistics. Chapman & Hall, London

Simon R (1989) Designs for efficient clinical trials. Oncology 3(7):43–53 with discussion

Simon R (1989) Optimal 2-stage designs for phase II clinical trials. Contr Clin Trial 10(1):1–10, DOI 10.1016/0197-2456(89)90015-9

Slud E, Wei LJ (1982) Two-sample repeated significance tests based on the modified Wilcoxon statistic. J Am Statist Assoc 77(380):862–868

Sonesson C, Bock D (2003) A review and discussion of prospective statistical surveillance in public health. J Roy Statist Soc Ser A 166(1):5–21, DOI 10.1111/1467-985X.00256

Spiegelhalter D, Freedman L, Blackburn P (1986) Monitoring clinical trials: Conditional or predictive power? Contr Clin Trial 7(1):8–17, DOI 10.1016/0197-2456(86)90003-6

Stein C (1945) A two-sample test for a linear hypothesis whose power is independent of the variance. Ann Math Statist 16:243–258

Storer B (1989) Design and analysis of phase I clinical trials. Biometrics 45(3):925–937, DOI 10.2307/2531693

Susarla V, Van Ryzin J (1976) Nonparametric Bayesian estimation of survival curves from incomplete observations. J Am Statist Assoc 71(356):897–902

Todd S, Whitehead J (1996) Point and interval estimation following a sequential clinical trial. Biometrika 83(2):453–461

Tsiatis AA (1981) The asymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time. Biometrika 68(1):311–315, DOI 10.1093/biomet/68.1.311

Tsiatis AA, Mehta C (2003) On the inefficiency of the adaptive design for monitoring clinical trials. Biometrika 90(2):367–378, DOI 10.1093/biomet/90.2.367

Vickers AJ, Ballen V, Scher HI (2007) Setting the bar in phase II trials: the use of historical data for determining “go/no go” decision for definitive phase III testing. Clin Cancer Res 13(3):972–976, DOI 10.1158/1078-0432.CCR-06-0909

Von Hoff DD, Turner J (1991) Response rates, duration of response, and dose-response effects in phase I studies of antineoplastics. Invest New Drugs 9(1):115–122

Vonesh EF (1996) A note on the use of Laplace’s approximation for nonlinear mixed-effects models. Biometrika 83(2):447–452, DOI 10.1093/biomet/83.2.447

Wald A (1945) Sequential tests of statistical hypotheses. Ann Math Statist 16:117–186

Wald A, Wolfowitz J (1948) Optimum character of the sequential probability ratio test. Ann Math Statist 19:326–339

Wallis WA (1980) The statistical research group, 1942–1945. J Am Statist Assoc 75(370):320–335, with comments by F.J. Anscombe and William H. Kruskal and a reply by the author

Wang S, Hung H, Tsong Y, Cui L (2001) Group sequential test strategies for superiority and non-inferiority hypotheses in active controlled clinical trials. Stat Med 20(13):1903–1912

Wang SK, Tsiatis AA (1987) Approximately optimal one-parameter boundaries for group sequential trials. Biometrics 43(1):193–199

Wassmer G, Vandemeulebroecke M (2006) A brief review on software developments for group sequential and adaptive designs. Biom J 48(4):732–737, DOI 10.1002/bimj.200510233

Whitehead J (1986) Supplementary analysis at the conclusion of a sequential clinical trial. Biometrics 42(3):461–471

Whitehead J (1992) The design and analysis of sequential clinical trials, 2nd edn. Wiley, New York

Whitehead J, Stratton I (1983) Group sequential clinical trials with triangular continuation regions. Biometrics 39(1):227–236

Whitehead J, Todd S, Hall WJ (2000) Confidence intervals for secondary parameters following a sequential test. J Roy Statist Soc Ser B 62(4):731–745, DOI 10.1111/1467-9868.00260

Whitehead J, Whitehead A, Todd S, Bolland K, Sooriyarachchi M (2001) Mid-trial design reviews for sequential clinical trials. Stat Med 20(2):165–176

Willsky AS, Jones HG (1976) A generalized likelihood ratio approach to detection and estimation of jumps in linear systems. IEEE Trans Automat Contr 21:108–112

Wilson EB, Worcester J (1943) The determination of L.D.50 and its sampling error in bio-assay. Proc Nat Acad Sci U S A 29(2):79–85

Wittes J, Brittain E (1990) The role of internal pilot: studies in increasing the efficiency of clinical trials. Stat Med 9(1–2):65–72, DOI 10.1002/sim.4780090113

Wolfinger R (1993) Laplace's approximation for nonlinear mixed models. Biometrika 80(4):791–795, DOI 10.1093/biomet/80.4.791

Wolfinger R, O'Connell M (1993) Generalized linear mixed models: a pseudo-likelihood approach. J Stat Comput Sim 48(3–4):233–243, DOI 10.1080/00949659308811554

Woodroofe M (1986) Very weak expansions for sequential confidence levels. Ann Statist 14(3):1049–1067, DOI 10.1214/aos/1176350049

Woodroofe M (1992) Estimation after sequential testing: a simple approach for a truncated sequential probability ratio test. Biometrika 79(2):347–353, DOI 10.1093/biomet/79.2.347

Wu CFJ (1985) Efficient sequential designs with binary data. J Am Statist Assoc 80(392):974–984

Yafune A, Takebe M, Ogata H (1998) A use of Monte Carlo integration for population pharmacokinetics with multivariate population distribution. J Pharmacok Biopharm 26(1):103–123, DOI 10.1023/A:1023280909207

Yin G, Yuan Y (2009) Bayesian model averaging continual reassessment method in phase I clinical trials. J Am Statist Assoc 104(487):954–968, DOI 10.1198/jasa.2009.ap08425

# Index

J. Bartroff et al., *Sequential Experimentation in Clinical Trials,*
Springer Series in Statistics 298, DOI 10.1007/978-1-4614-6114-2,