Springer Series in Pharmaceutical Statistics

Gernot Wassmer
Werner Brannath

Group Sequential and Confirmatory Adaptive Designs in Clinical Trials

Springer

Springer Series in Pharmaceutical Statistics

More information about this series at http://www.springer.com/series/15122

Gernot Wassmer • Werner Brannath

Group Sequential and Confirmatory Adaptive Designs in Clinical Trials

Springer

Gernot Wassmer
University of Cologne
Cologne, Germany

Werner Brannath
University of Bremen
Bremen, Germany

ISSN 2366-8695 ISSN 2366-8709 (electronic)
Springer Series in Pharmaceutical Statistics
ISBN 978-3-319-32560-6 ISBN 978-3-319-32562-0 (eBook)
DOI 10.1007/978-3-319-32562-0

Library of Congress Control Number: 2016945375

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Preface

During the last two decades, there was an intensive discussion of a statistical methodology for clinical trials that generalized the use and conduct of interim analyses. Using a specific methodology, under control of the Type I error rate, it was made possible to redesign the trial for the forthcoming stages in relevant details including sample size, considered treatment arms, subgroups of patients, and others. The use of such multi-stage adaptive designs was introduced in an article by Peter Bauer entitled "Multi-stage testing with adaptive designs." It appeared 1989 in the German journal *Biometrie und Informatik in Medizin und Biologie*. Especially the publication by Bauer and Köhne 1994 in *Biometrics* raised many controversial discussions. In the meantime, the underlying concept is regarded as a relevant and important generalization of the "classical" group sequential design methodology. It has even found its way into a reflection paper entitled "Methodological Issues in Confirmatory Clinical Trials Planned With an Adaptive Design" from the European Medicines Agency (EMA 2007), a draft guidance on "Adaptive Clinical Trials for Drug and Biologics" from the US Food and Drug Administration (FDA 2010), and a draft guidance on "Adaptive Designs for Medical Device Clinical Studies," also from the US Food and Drug Administration (FDA 2015).

This book describes the adaptive design methodology from the standpoint of confirmatory testing of statistical hypotheses. That is, the problem is to derive procedures that control, at least approximately, the Type I error irrespective of the data-driven redesign of the trial. We do not consider other very relevant "adaptive topics" such as adaptive dose finding, response adaptive randomization, Bayesian design methodology, or adaptive regression modeling. Here, Type I error rate control is not the major issue, and the methodology is often applied in the exploratory case only. This is said without diminishing its importance and relevance for clinical trials. We also explicitly exclude the discussion of blinded sample size reassessment which is one of the simplest adaptive designs and usually consists of two stages in which the sample size for the second stage is determined based on the first-stage data in a blinded way (for a review of these methods, see Friede and Kieser 2006). We also do not explicitly discuss the regulatory perspectives and operation of data monitoring committees (for this, see DeMets et al. 2006; Ellenberg et al. 2003; He et al. 2014; Herson 2009) although we briefly discuss some practical aspects for adaptive designs at the end of the book.

The most relevant, comprehensive, and modern books on group sequential methodology in clinical trials are Whitehead (1997), Jennison and Turnbull (2000), Moyé (2006), Proschan et al. (2006), and Bartroff et al. (2013); we also refer to overview articles of Whitehead (2001), Todd (2007), Mazumbdar and Bang (2008), and Wassmer (2009). Books and collections on adaptive designs are Chang (2014), Chin (2012), Chow and Chang (2006), He et al. (2014), Pong and Chow (2011), Schmitz (1993), Vandemeulebroecke (2008), and Wassmer (2010). Recently, Bauer et al. (2016) provide a critical overview and perspectives of 25 years of adaptive designs.

In this monograph, we give an overview and a basic introduction into both, the group sequential and the adaptive design methodology. Wassmer (1999c) already gave an early attempt (in German) to summarize the approaches; the current book is partly developed from the translation and the extension of this small book. It is also Proschan et al. (2006), who already explicitly considered the adaptive extension within a unified approach; in our book there will be the focus on the adaptive designs.

The monograph consists of three parts. The first part provides the group sequential methods that are necessary for the understanding and the application of the generalization to the adaptive methodology supplied in Parts II and III of the book. We try to provide the necessary details as well as an insight into the complexity and richness of these designs including the estimation problem and more flexible designs when using this “classical” approach. It is by no means intended to give a complete survey of group sequential designs (or even sequential theory). Especially the books of Jennison and Turnbull (2000) and Proschan et al. (2006) provide a much deeper insight into this methodology, but we hope to provide an elementary introduction in this area.

Part II is the core of the monograph. It derives the adaptive design methodology as it was introduced in the mid-1990s of the last century. The two basic approaches of Bauer and Köhne (1994) and Proschan and Hunsberger (1995), which are the *combination testing principle* and the *conditional error function approach*, respectively, are presented in Chap. 6 in a unified framework, together with some relevant generalizations of the principles. Chapter 7 describes decision tools that can be used for adaptive decision making where here we only consider the case of sample size reassessment based on conditional power and related tools. In Chap. 8, we describe solutions for the estimation problem in adaptive designs and also discuss some open issues. Adaptive designs with survival data involve some difficulties which are discussed in Chap. 9.

Primarily, the focus of flexible designs was on sample size reassessment although from the very beginning it was already emphasized that the principle allows much more general design adaptations including a data-driven selection of hypotheses within a set of hypotheses. This is the topic of Part III where we describe the closed testing principle and its use in adaptive multi-stage designs with multiple hypotheses. It is the nowadays highlighted topic. The most prominent applications are adaptive multi-arm trials or adaptive population enrichment designs. We describe these in Chap. 11 together with some clinical trial examples. There is still ongoing research in this area, and we hope to give an up-to-date review of the approaches that were proposed in the literature. We admit that we were not able to include all the different aspects in the considerations of these or even the more elementary adaptive designs. Particularly, the references cannot consider all papers that were published in this exciting area of the statistical methodology. Prophylactically, we apologize that we missed to cite some important work. We were simply not able to cite all the numerous articles that have been published in the last two decades.

We thank the many colleagues that went along with us in the last 20 years. 1995 and 1998 were the years when we (GW and WB, respectively) started research in the adaptive design methodology. It is too much to explicitly name all of the colleagues who contributed and discussed the topics with us; to name *all* of them is easier and less problematic with respect to forgetting someone. We especially thank Frank Bretz and Silke Jörgens for providing input on relevant topics for the book. We also thank the publisher for his support on the manuscript. Last but not least, we want to thank our wives, Anne and Stela, and our children for accepting and supporting our work.

Cologne, Germany	Gernot Wassmer
Bremen, Germany	Werner Brannath
October 2015

Contents

Part I Group Sequential Designs

List of Tables

List of Figures

Part I
Group Sequential Designs

Chapter 1
Repeated Significance Tests

In this chapter, we describe the repeated significance tests approach which is the conceptual background of the "classical" group sequential tests. These tests were partly made feasible through a specific (recursive) integration formula, which is a consequence of the independent increment structure of the underlying process of data accumulation. This formula is due to Armitage et al. (1969) and McPherson and Armitage (1971). We will briefly introduce it in Sect. 1.4. Note that this section is quite technical and can be omitted for the first reading.

We start with a brief historical survey of group sequential tests. The notation and the construction of statistical tests that are based on the repeated significance testing approach are described in the following section. We then address the basic issues—power and average sample size—for the assessment of these tests in order to discuss the properties of group sequential test procedures that were proposed in the literature.

1.1 Introduction

A statistical significance test is given by the decision regions for rejecting and not rejecting the null hypothesis H_0, respectively. After a preplanned number of observations a test statistic is calculated. If the test statistic exceeds a critical value, H_0 is rejected, otherwise H_0 cannot be rejected. If H_0 is rejected, the alternative hypothesis H_1 is "statistically significant." The probability of erroneously rejecting H_0 (which is the Type I error rate) is bounded by the significance level α of the test. In randomized controlled clinical trials hypotheses tests are mandatory since they ensure that the probability of a false positive test result does not exceed a pre-specified level α and hence, in the long run, only a small proportion of actually ineffective treatments were applied. This frequentist hypothesis testing principle

G. Wassmer, W. Brannath, *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*, Springer Series in Pharmaceutical Statistics,
DOI 10.1007/978-3-319-32562-0_1

is widely used in practice, in particular, for clinical trials in later phases of drug development. Practically, the calculation of p-values in standard statistical software packages enables a very easy way to perform these tests: H_0 is rejected if $p \leq \alpha$. But this is only one of the attractive features of the p-value. Another reason for the widespread application of p-values in clinical trials is its explorative and descriptive nature that enables the presentation of results in a very convenient manner.

In sequential sampling schemes, the finally achieved sample size is not fixed. Typically, the study is continued as long as the underlying test statistic does not exceed a specified boundary. If one uses the decision boundaries of a fixed sample size test, the actual level of the test will be greater than α due to the fact that several tests at significance level α are conducted on the accumulating data. The concept of repeated significance testing, that is due to Armitage et al. (1969), is based on this fundamental property. It involves finding appropriate adjusted decision regions such that the actual significance level is not greater than α. This is possible if a fixed schedule of inspections is determined in advance. In order to control a specified significance level α somewhat larger critical values than for fixed sample size significance tests must be used.

In group sequential sampling schemes the schedule usually consists of accumulating data in groups of observations, and fixing the maximum number of stages. This is in contrast to fully sequential plans with continuous monitoring of the accumulating data. Although Peter Armitage was the first who adapted the sequential testing approach to clinical research (Armitage 1975; first published 1960), Pocock (1977, 1982) and O'Brien and Fleming (1979) gave the major impetus for the development of group sequential test procedures that are nowadays widely used, especially in clinical research. Their use in quality control goes back to the work of Dodge and Romig (1929) and Shewhart (1931). For historical reviews of the early work about group sequential procedures see Jennison and Turnbull (1991b, 2000), §1.3, and Todd (2007). Ghosh (1991) provides a comprehensive history of sequential analysis taking into account the developments beginning in the seventeenth century.

The repeated significance testing approach is conceptually different from approaches where typically Type I and Type II error probabilities are used to determine the stopping boundaries of the sequential test procedure (see, for example, Bauer et al. 1986; Ghosh 1970; Siegmund 1985; Wald 1947; Wetherill 1975; Whitehead 1997). The comprehensive theoretical development of these procedures is much owing to the optimality of the sequential probability ratio test (SPRT), and to the derivation of analytic expressions for the decision regions and certain test characteristics. The SPRT is optimal in the sense that the average sample size is minimum under both the null and alternative hypothesis. Theoretical research on repeated significance tests was also done. A series of papers were concerned with finding a bound for the critical value and approximations for the power and the average sample size. Much of this work is due to the research group of David Siegmund where many of the theoretical results were obtained from renewal theory. Essential developments and results are presented in Siegmund (1985).

An important part in the development of sequential designs was concerned with "triangular plans" where the stopping region is defined by two straight lines which cross by the end of the trial. These are "closed" or "truncated" plans where a decision to stop the trial is fixed at some maximum amount of information. Much of the theoretical development is concerned with the overshoot problem that occurs when groups of observations are considered rather than continuous monitoring. In this case it typically happens that the decision is not on the boundary. Mathematical sophistication has led to the "Christmas tree correction" which turns out to be quite accurate in many cases of practical importance (Whitehead and Stratton 1983). Whitehead (1997) provides a comprehensive overview of sequential methods with special reference to the triangular plans in clinical trials.

In a review article, Whitehead (2001) argues against the distinction of group sequential methods from the wider family of sequential methods. Certainly this is true, but the development and investigation of group sequential designs was in some sense separated from the rigorous mathematical derivation of results within the sequential theory. This is also due to the rapid development of computer power which made the computations of the recursive numerical integral introduced by Armitage et al. (1969) possible. Easy to use computer programs are available today to investigate the characteristics of the procedures numerically.

In clinical research there is a great interest in interim analyses for ethical, economical, and organizational reasons. A sequential design offers the possibility to stop a trial early with a statistically significant test result. Hence, this trial is likely to need less patients than the trial with a fixed sample size where a decision can be made only by the end of the trial. If a therapy was shown to be superior to another one, the inferior therapy can be replaced by the better one and the superior therapy can then be applied earlier. In interim analyses one can also assess the quality of the performance of the trial and possibly improve it when necessary. The observation of serious adverse events can lead to an early stopping of the trial. The appointment of a Safety and Data Monitoring Committee in a clinical trial is a recommended constitution to perform these issues according to generally accepted standards of good clinical practice (GCP) (see Armitage 1991; McPherson 1990; Sankoh 1999). Another issue is the redesign of the trial (for example, sample size reassessment, dropping treatment arms, and selecting subpopulations), but this was not intended originally nor is it generally possible with "classical" group sequential test designs. In Parts II and III of this book the methods specifically designed for these purposes will be presented.

Pocock (1977) proposed two-sided tests for normal responses with known variance assuming that the number of looks at the data is known in advance, and the number of observations is equal between the stages. In a group sequential test design the hypothesis concerning the mean can be rejected if the p-value is smaller or equal than a value α' where α' is determined such that the overall probability of a false rejection of H_0 in the sequential scheme does not exceed α. Clearly, α' depends on α and on the number of planned stages. Pocock's design is characterized by assuming constant adjusted significance levels α' over the stages of the trial. O'Brien and Fleming (1979) suggested an approach that requires more conservative bounds for very early stages and proposed an increasing sequence of adjusted significance levels. As a consequence the final stage adjusted significance level comes near to the unadjusted value α. It is interesting that in their original paper the adjusted bounds were approximated using simulations and the testing procedure was proposed for comparing two treatments with regard to a dichotomous response.

For the exact numerical calculation of the adjusted critical levels it was shown that the quite restrictive assumption of normality and, more important, equal stage sizes can be relaxed. The size and power requirement is fulfilled if these assumptions are not grossly violated. Nevertheless, the importance of more flexible procedures is obvious. Several approaches were proposed that do not require these restrictive assumptions, of which the most prominent is the α-spending function or use function approach (Lan and DeMets 1983). The idea is to specify the amount of significance level spent up to an interim analysis rather than the shape of the adjusted critical levels. Particularly, the sample size per stage may be an unknown quantity that will be observed when the interim analysis is performed. With this approach even the number of interim analyses needs not be fixed in advance. Instead, a maximum amount of information must be specified which, in the simplest case, is the maximum sample size of the trial. When conducting an interim analysis a certain information, relative to the maximum amount of information, is observed. Through the use of a specified use function, the significance level spent up to this information is fixed in advance. This enables the calculation of the adjusted levels. The use function approach is outstandingly attractive if the interim analyses are planned at specific time points rather than after a specific number of observations. An important application of group sequential designs is in trials where the endpoint is the time to an event, for example, survival data. It was shown by several authors that the usual log-rank test can be embedded into the group sequential design. The information here is the observed number of events and the use function approach turns out to be a very useful and flexible instrument for analyzing such trials in a sequential design.

An important field of research was concerned with the parameter estimation in group sequential test designs. Through the use of a stopping rule, i.e., the possibility of early stopping a trial with the rejection (or acceptance) of the null hypothesis, point estimates that are derived for the fixed sample size case (for example, maximum likelihood estimates) are biased. In the long run, hence, one is faced with the over- or underestimation of the true parameter. Point estimates were proposed that correct for the estimation bias through numerical methods (for example, Emerson 1993; Emerson and Fleming 1990; Emerson and Kittelson 1997; Liu and Hall 1999; Pinheiro and DeMets 1997; Todd et al. 1996; Whitehead 1986). These estimates try to correct for the overall estimation bias but do not take into account the bias per stage of the test procedure. The correction for the (conditional) bias per stage was proposed and investigated by Troendle and Yu (1999) and Coburger and Wassmer (2001, 2003).

Two conceptually different methods for the calculation of confidence intervals were considered in the literature. The first method enables the calculation after the trial has stopped and a final decision of rejection or acceptance of the null hypothesis was reached (for example, Chang and O'Brien 1986; Coad and Woodroofe 1996; Kim and DeMets 1987a; Rosner and Tsiatis 1988; Todd et al. 1996; Tsiatis et al. 1984). This approach requires the strict adherence to the stopping rule and depends on the ordering in the sample space. This means, it has to be decided if, for example, an observed effect leading to the rejection of the null hypothesis in the first interim is "more extreme" than an effect that is larger but observed in the second interim analysis. This ordering involves the earlier the stopping the stronger the effect, but there are other orderings that are reasonable choices. The second method merely takes into account the multiplicity that arises from the repeated looks at the data. The resulting intervals are called *Repeated Confidence Intervals* which were introduced by Jennison and Turnbull (1984) and Lai (1984). The application of repeated confidence intervals was thoroughly discussed by Jennison and Turnbull (1989). These confidence intervals are independent of the stopping rule and can also therefore be calculated if the study is going on. If desired, they can be calculated at each interim analysis and presented to the data monitoring committee.

1.2 Basics

In the non-sequential case, the sample size, n_f, is fixed and one comes to a decision after observing the complete sample. Consider independent and normally distributed observations $X_1, \ldots, X_{n_f}$ with unknown mean μ and known variance σ^2. Let the null hypothesis to be tested at significance level α be given by

$$H_0 : \mu = \mu_0 \ .$$

This model is used, for example, for paired comparisons, but we will see in Chap. 5 that the resulting group sequential test designs can be used as a "prototype" for many different testing situations. The test decision is based on the statistic

$$Z = \frac{\bar{X} - \mu_0}{\sigma} \sqrt{n_f} \ ,$$

which is standard normal under H_0. Let H_0 be tested against the two-sided alternative, which is given by

$$H_1 : \ \mu \neq \mu_0 \ .$$

H_0 is rejected if the observed absolute value $|z|$ of the test statistic Z exceeds the critical value $u = \Phi^{-1}(1-\alpha/2)$, where Φ^{-1} denotes the inverse standard normal cumulative distribution function (cdf). Since

$$P_{H_0}(|Z| \geq u) = \alpha \ ,$$

the test has level α.

Alternatively, given the observed value z of Z, one can compute the p-value which is defined by

$$p = P_{H_0}(|Z| \geq |z|) = \min\{(2\,(1-\Phi(|z|)), 1\} \ ,$$

where $\Phi(\cdot)$ is the standard normal cdf. If $p \leq \alpha$, the null hypothesis H_0 can be rejected. That is, the test decision can be based on the observed z-value as well as on the p-value, the latter with the advantage of having an attractive interpretation: The lower the p-value the stronger the sample indicates evidence against H_0. Although this interpretation depends on the sample size, n_f, it has gained widespread acceptance in the (medical) literature.

In the group sequential setting, inspections are made after groups of observations. Given a maximum number of inspections, K, let the sample sizes in the K sequences of observations be given by $n_1, \ldots, n_K$. $N = \sum_{k=1}^{K} n_k$ is the maximum sample size of the test procedure. The observations X_{ki} are indexed by an additional index, k, that refers to the stage in which the data were observed. With this notation, the independent observations are given by

$$\underbrace{X_{11}, \ldots, X_{1n_1}}_{n_1 \text{ observations}}, \underbrace{X_{21}, \ldots, X_{2n_2}}_{n_2 \text{ observations}}, \ \ldots \ , \underbrace{X_{K1}, \ldots, X_{Kn_K}}_{n_K \text{ observations}} .$$

The cumulative mean of observations up to stage k is given by

$$\bar{X}^{(k)} = \frac{1}{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}} \sum_{\tilde{k}=1}^{k} n_{\tilde{k}} \bar{X}_{\tilde{k}} \ , \tag{1.1}$$

where $\bar{X}_k$ denotes the mean of observations at stage k, $k = 1, \ldots, K$, i.e.,

$$\bar{X}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} X_{ki} \ .$$

At stage k, a reasonable test statistic for testing H_0 is

$$Z_k^* = \frac{\bar{X}^{(k)} - \mu_0}{\sigma} \sqrt{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}} \ , \tag{1.2}$$

which is standard normal under H_0. One might as well consider the stage-wise test statistic

$$Z_k = \frac{\bar{X}_k - \mu_0}{\sigma}\sqrt{n_k}\ ,$$

and calculate (1.2) through

$$Z_k^* = \frac{\sum_{\tilde{k}=1}^{k}\sqrt{n_{\tilde{k}}}Z_{\tilde{k}}}{\sqrt{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}}}\ . \tag{1.3}$$

At stage k, the test statistic that summarizes the data up to stage k is then a weighted sum of the stage-wise test statistics Z_k. It is important to realize, however, that the weights in (1.3) are not proportional to the sample sizes but to their square roots. Furthermore, they do not sum up to one. The test statistic is thus not—in a classical sense—a weighted average of the stage-wise Z_k.

Obviously, the statistics $Z_1^*, \ldots, Z_K^*$ are stochastically dependent. The covariance between Z_k^* and $Z_{k'}^*$ $(k < k')$ is given by

$$\begin{aligned}\mathrm{Cov}(Z_k^*, Z_{k'}^*) &= \frac{1}{\sqrt{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}}\sqrt{\sum_{\tilde{k}=1}^{k'} n_{\tilde{k}}}}\,\mathrm{Cov}\left(\sum_{\tilde{k}=1}^{k}\sqrt{n_{\tilde{k}}}Z_{\tilde{k}}, \sum_{\tilde{k}=1}^{k'}\sqrt{n_{\tilde{k}}}Z_{\tilde{k}}\right)\\ &= \frac{1}{\sqrt{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}}\sqrt{\sum_{\tilde{k}=1}^{k'} n_{\tilde{k}}}}\sum_{\tilde{k}=1}^{k} n_{\tilde{k}} = \frac{\sqrt{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}}}{\sqrt{\sum_{\tilde{k}=1}^{k'} n_{\tilde{k}}}}\ . \end{aligned}\tag{1.4}$$

Since $\mathrm{Var}(Z_k^*) = 1$, $k = 1, \ldots, K$, the quantity (1.4) is also the correlation coefficient between Z_k^* and $Z_{k'}^*$.

The random vector $\boldsymbol{Z}^* = (Z_1^*, \ldots, Z_K^*)^T$ is multivariate normal with mean vector $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_K)^T$ where

$$\vartheta_k = E(Z_k^*) = \frac{\mu - \mu_0}{\sigma}\sqrt{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}} = \delta\sqrt{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}},\ k = 1, \ldots, K, \tag{1.5}$$

and $\delta = (\mu - \mu_0)/\sigma$ is the standardized effect size. Because of (1.4), the elements $\boldsymbol{\Sigma}_{kk'}$ of the covariance (correlation) matrix $\boldsymbol{\Sigma}$ are given by

$$\boldsymbol{\Sigma}_{kk'} = \mathrm{Cov}(Z_k^*, Z_{k'}^*) = \frac{\sqrt{\sum_{\tilde{k}=1}^{\min\{k,k'\}} n_{\tilde{k}}}}{\sqrt{\sum_{\tilde{k}=1}^{\max\{k,k'\}} n_{\tilde{k}}}}\ . \tag{1.6}$$

$\boldsymbol{Z}^*$ is multivariate normally distributed since the random vector can be written in the form

$$\boldsymbol{Z}^* = \boldsymbol{A}\boldsymbol{U} + \boldsymbol{a} \,, \tag{1.7}$$

where $\boldsymbol{U}$ is a random vector consisting of K independent and standard normally distributed random variables, $\boldsymbol{A}$ is a $K \times K$ matrix, and $\boldsymbol{a}$ is a K-dimensional vector. This transformation yields, by definition, a multivariate normally distributed random variable with mean vector and covariance matrix given by

$$\boldsymbol{\vartheta} = \boldsymbol{a} \quad \text{and} \quad \boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^T,$$

respectively. In our case, $\boldsymbol{A}$ and $\boldsymbol{a}$ are given by

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ \frac{\sqrt{n_1}}{\sqrt{n_1+n_2}} & \frac{\sqrt{n_2}}{\sqrt{n_1+n_2}} & 0 & \ldots & 0 \\ & \ldots\ldots\ldots\ldots\ldots\ldots & & & \\ \frac{\sqrt{n_1}}{\sqrt{n_1+\ldots+n_{K-1}}} & \ldots & \frac{\sqrt{n_{K-1}}}{\sqrt{n_1+\ldots+n_{K-1}}} & & 0 \\ \frac{\sqrt{n_1}}{\sqrt{n_1+\ldots+n_K}} & & \ldots\ldots\ldots & & \frac{\sqrt{n_K}}{\sqrt{n_1+\ldots+n_K}} \end{pmatrix}$$

and

$$\boldsymbol{a} = \left(\delta\sqrt{n_1}, \delta\sqrt{n_1+n_2}, \ldots, \delta\sqrt{n_1+\ldots n_K}\right)^T,$$

respectively. One easily finds that, indeed, (1.7) yields the random vector $\boldsymbol{Z}^*$. Note that, using this matrix notation, one alternatively finds the covariance matrix $\boldsymbol{\Sigma}$ with elements given by (1.6) through $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^T$.

Consider a two-stage design (i.e., $K = 2$) with equal sample sizes $n_1 = n_2$ for the two stages. In this case, from (1.4) one finds that $\text{Cov}(Z_1^*, Z_2^*) = 1/\sqrt{2}$. Suppose one uses the unadjusted critical value $u = \Phi^{-1}(1 - \alpha/2)$ in the interim analysis as well as in the final analysis. Setting $\alpha = 0.05$, the hypothesis H_0 is rejected in the interim analysis if $|z_1^*| \geq 1.96$. If $|z_1^*| < 1.96$, the second stage data will be observed and the test statistic z_2^* will be calculated; H_0 is rejected if $|z_2^*| \geq 1.96$. The Type I error rate of this procedure is given by

$$\begin{aligned} &P_{H_0}(|Z_1^*| \geq 1.96 \text{ or } |Z_2^*| \geq 1.96) \\ &\quad = P_{H_0}(|Z_1^*| \geq 1.96) + P_{H_0}(|Z_1^*| < 1.96 \text{ and } |Z_2^*| \geq 1.96) \\ &\quad = 2\,P_{H_0}(Z_1^* \leq -1.96) + 2\left(P_{H_0}(|Z_1^*| < 1.96 \text{ and } Z_2^* \leq -1.96\right) \\ &\quad = 2\,\Phi(-1.96) + 2\left(F(1.96, -1.96) - F(-1.96, -1.96)\right) \\ &\quad = 0.05 + 0.0331 = 0.0831 \; > \; 0.05 \,. \end{aligned} \tag{1.8}$$

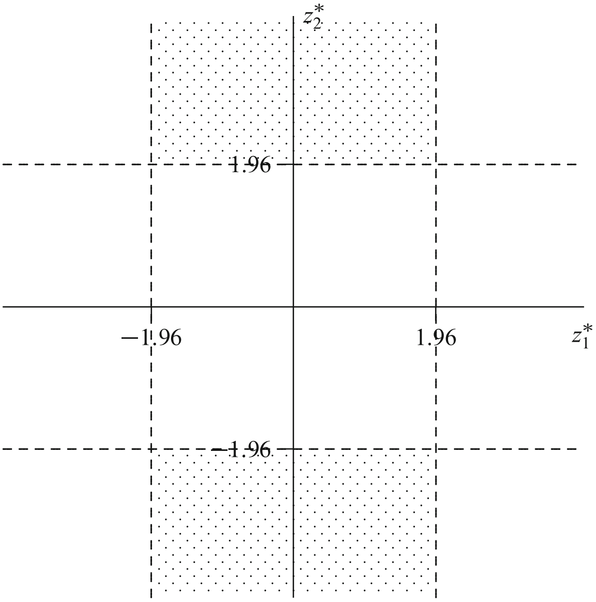

Fig. 1.1 Illustration for the calculation of the probability of a Type I error at the second stage in a two-stage group sequential test design with critical value 1.96 for both stages. The *shaded areas* indicate the region for the integration, whereby the integral is the same for both shaded regions

where F denotes the bivariate standard normal cumulative distribution function with correlation $1/\sqrt{2}$. The calculation of the probability for the Type I error at the second stage in (1.8) is geometrically illustrated in Fig. 1.1.

The probability of a false rejection of H_0 is substantially above the significance level $\alpha = 0.05$. Hence, using a nominal level of $\alpha = 0.05$ at both stages of the test procedure increases the Type I error rate by a large amount such that a suitable adjustment is necessary. One could use the Bonferroni correction to adjust for the multiplicity that arises from the multiple looks at the data. Using this method, in the interim as well as in the final analysis the p-value is compared with $\alpha/2$ or, equivalently, the critical value $u' = \Phi^{-1}(1 - \alpha/4)$ is used for the test decision. For $\alpha = 0.05$, the critical value is $u' = 2.241$. One obtains

$$P_{H_0}(|Z_1^*| \geq 2.241 \text{ or } |Z_2^*| \geq 2.241) = 0.0428 < 0.05 \; .$$

Using the Bonferroni correction is thus a valid though conservative solution to this kind of multiplicity since the Type I error rate does not fully exhaust α. A better solution is provided by using the critical value for which the Type I error rate is exactly equal to α. Using a numerical search, one finds that

$$P_{H_0}(|Z_1^*| \geq 2.178 \text{ or } |Z_2^*| \geq 2.178) = 0.05\,, \tag{1.9}$$

and hence the adjusted critical value is given by $u' = 2.178$. The adjusted nominal level to be applied at each stage of the procedure is given by $\alpha' = 0.0294$. That is, H_0 can be rejected in the interim or in the final analysis if the p-value is smaller than 0.0294 in one of the two stages of the test procedure.

Using identical critical values for both stages, however, is only one possible solution. For example, since

$$P_{H_0}(|Z_1^*| \geq 2.797 \text{ or } |Z_2^*| \geq 1.977) = 0.05\,, \tag{1.10}$$

one can also use the critical value $u_1 = 2.797\,(= \sqrt{2}\;\,1.977)$ for the interim stage, and the critical value $u_2 = 1.977$ for the final stage to obtain a valid level α test. Of course, an infinite number of critical levels exist reflecting the different choices of "spreading" the significance level α over the stages. The use of constant values at each stage of the test procedure is due to Pocock (1977) whereas the use of monotonically decreasing critical values according to (1.10) was proposed in O'Brien and Fleming (1979). We will thoroughly discuss these designs in Chap. 2.

Extending this to more than two stages is not straightforward since the calculation of the general multivariate normal integral is a difficult task. The calculation of the Type I error rate is possible, though, through the use of the *recursive integration formula* which is due to Armitage et al. (1969). With this formula, the complex determination of the multivariate normal integral is avoided and replaced by a successive calculation of univariate integrals. It is then possible to calculate the multivariate normal integral in the group sequential setting for an arbitrary number of stages, and hence to calculate, for example, the Type I error rate. In Table 1.1 the Type I error rate when using the unadjusted critical value $u = \Phi^{-1}(1 - \alpha/2)$ for different values of α and for K up to 50 is presented. It is assumed that the stage sizes are equal, i.e., $n_1 = \cdots = n_K$. The figures in this table were already presented in Armitage et al. (1969) who also checked up on their theoretical results (obtained with the recursive integration formula) by simulation. The recursive integration formula will be presented later on in Sect. 1.4.

Armitage et al. (1969) pioneered the idea of controlling the Type I error rate at a specified level by adjusting the critical values accordingly. They found critical values analogously to (1.9) by inverse interpolation. We will discuss the different choices of critical values applicable to the group sequential setting in Chaps. 2 and 3.

A group sequential test design consists of specifying the continuation regions $\mathscr{C}_k^*$ at the analysis stages $k = 1, \ldots, K-1$, and the rejection region $\mathscr{R}_K^*$ for the final analysis. The study is continued if $Z_k^* \in \mathscr{C}_k^*$, $k = 1, \ldots, K-1$, and H_0 is rejected in the final stage if $Z_K^* \in \mathscr{R}_K^*$. In the simplest case, the null hypothesis can be rejected

Table 1.1 Type I error rate when using the unadjusted critical value $u = \Phi^{-1}(1-\alpha/2)$ at each stage of the test procedure in a two-sided test design

	α	0.001	0.01	0.05
K	u	3.291	2.576	1.960
1		0.00100	0.01000	0.05000
2		0.00186	0.01766	0.08312
3		0.00257	0.02366	0.10726
4		0.00319	0.02858	0.12617
5		0.00372	0.03274	0.14169
10		0.00569	0.04738	0.19336
15		0.00705	0.05692	0.22509
20		0.00808	0.06403	0.24791
25		0.00892	0.06971	0.26567
30		0.00963	0.07444	0.28016
35		0.01025	0.07849	0.29238
40		0.01079	0.08204	0.30293
45		0.01128	0.08519	0.31220
50		0.01172	0.08803	0.32045

The stage sizes are assumed to be equal

at stage k if $Z_k^* \in \overline{\mathscr{C}}_k^*$, $k = 1, \ldots, K$, where $\overline{\mathscr{C}}_k^*$ denotes the complement of $\mathscr{C}_k^*$ and $\overline{\mathscr{C}}_K^* = \mathscr{R}_K^*$. In this case, any design fulfilling

$$\begin{aligned} &P_{H_0}(Z_1^* \in \overline{\mathscr{C}}_1^* \text{ or } Z_2^* \in \overline{\mathscr{C}}_2^* \text{ or } \ldots \text{ or } Z_K^* \in \overline{\mathscr{C}}_K^*) \\ &= P_{H_0}\left(\bigcup_{k=1}^{K}\{Z_k^* \in \overline{\mathscr{C}}_k^*\}\right) = 1 - P_{H_0}\left(\bigcap_{k=1}^{K}\{Z_k^* \in \mathscr{C}_k^*\}\right) = \alpha \end{aligned} \tag{1.11}$$

is a valid level-α test procedure. If $f(z_1^*, \ldots, z_K^*)$ denotes the multivariate normal density with zero mean vector and correlation matrix given by (1.6), the Type I error rate can be written as

$$1 - \int_{\mathscr{C}_K^*} \cdots \int_{\mathscr{C}_1^*} f(z_1^*, \ldots, z_K^*)\, dz_1^* \ldots dz_K^*.$$

It becomes clear again that the Type I error rate must be computed from the multivariate normal distribution but the recursive integration, described in Sect. 1.4, will do this task.

Note that an alternative expression for the left-hand side of (1.11) is

$$\begin{aligned} &P_{H_0}(Z_1^* \in \overline{\mathscr{C}}_1^*) + P_{H_0}(Z_1^* \in \mathscr{C}_1^*, Z_2^* \in \overline{\mathscr{C}}_2^*) + \cdots \\ &+ P_{H_0}(Z_1^* \in \mathscr{C}_1^*, \ldots, Z_{K-1}^* \in \mathscr{C}_{K-1}^*, Z_K^* \in \overline{\mathscr{C}}_K^*)\,. \end{aligned} \tag{1.12}$$

This presentation enables a more general calculation of the Type I error rate of a specified test procedure described as follows. Generally, the rejection region is not the complement of the continuation region. Denoting the rejection regions at stage k by $\mathscr{R}_k^*$, $k = 1, \ldots, K$, the Type I error rate must be computed as indicated in (1.12). It is given by

$$\begin{aligned} &P_{H_0}(Z_1^* \in \mathscr{R}_1^*) + P_{H_0}(Z_1^* \in \mathscr{C}_1^*, Z_2^* \in \mathscr{R}_2^*) + \cdots \\ &\quad + P_{H_0}(Z_1^* \in \mathscr{C}_1^*, \ldots, Z_{K-1}^* \in \mathscr{C}_{K-1}^*, Z_K^* \in \mathscr{R}_K^*) \ . \end{aligned} \tag{1.13}$$

Using (1.13), the Type I error rate of any group sequential test design can be calculated.

For example, consider a two-stage group sequential test design with $n_1 = n_2$. Let the continuation region be given by

$$\mathscr{C}_1^* = (-2.178;\ -1.0) \cup (1.0;\ 2.178) \ ,$$

and the rejection regions be defined by

$$\mathscr{R}_1^* = \mathscr{R}_2^* = (-\infty;\ -2.178] \cup [2.178;\ \infty) \ .$$

That is, in the first interim analysis the study will be continued if, and only if, a more or less substantial effect was observed that did not yet reach significance in the first stage. Note that $Z_1^* \in \mathscr{C}_1^*$ is equivalent to the condition that the two-sided p-value is smaller than $2(1 - \Phi(1)) = 31.7\,\%$ (and larger than $2(1 - \Phi(2.178)) = 2.94\,\%$). If the p-value is larger than 31.7 %, the study will be stopped for futility. We will discuss this issue that is related to a futility stop in more detail in Chap. 2.

The Type I error rate of the test procedure is calculated as

$$\begin{aligned} &P_{H_0}(Z_1^* \in \mathscr{R}_1^*) + P_{H_0}(Z_1^* \in \mathscr{C}_1^*, Z_2^* \in \mathscr{R}_2^*) \\ &\quad = \int_{\mathscr{R}_1^*} f(z_1^*)\ dz_1^* + \int_{\mathscr{R}_2^*} \int_{\mathscr{C}_1^*} f(z_1^*, z_2^*)\ dz_1^* dz_2^* \\ &\quad = P_{H_0}(|Z_1^*| \geq 2.178) + P_{H_0}(|Z_1^*| \in (1.0;\ 2.178), |Z_2^*| \geq 2.178) \\ &\quad = 2\,\Phi(-2.178) + 2\,\big(F(2.178, -2.178) - F(-2.178, -2.178) \\ &\qquad - F(1.0, -2.178) + F(-1.0, -2.178)\big) = 0.0458 \ , \end{aligned}$$

where the calculation of the bivariate normal integral is illustrated in Fig. 1.2. Therefore, using the critical values 2.178 for both stages of the procedure yields a conservative test. In order to fully exhaust the significance level α of the test, the critical values can be made somewhat smaller. Setting $u_1 = u_2 = 2.140$ yields a test with constant critical values for both stages whose Type I error rate is equal to 0.05. This value is found by a numerical search.

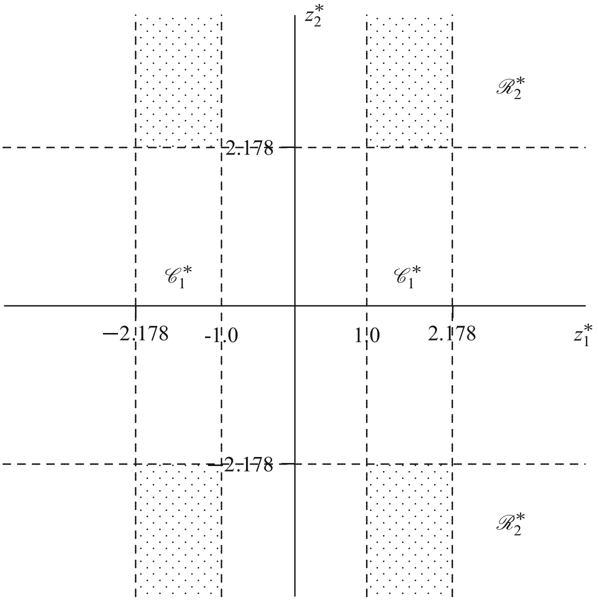

Fig. 1.2 Illustration for the calculation of the probability of a Type I error at the second stage in a two-stage group sequential test design with critical value 2.178 for both stages and continuation region $\mathscr{C}_1^* = (-2.178, -1.0) \cup (1.0, 2.178)$. The *shaded areas* indicate the region for the integration where the integral is the same for the lower shaded and the upper shaded regions

1.3 Power and Average Sample Size

After having defined the continuation and rejection regions of a group sequential test procedure, it is of interest to assess the test procedure in terms of its power, i.e., the probability to reject H_0 in one of the stages of the trial if the alternative is true. For a specified alternative hypothesis $H_1 : \mu = \mu_1$ and specified stage sample sizes $n_1, \ldots, n_K$ the power is given by

$$\begin{aligned}&P_{H_1}(Z_1^* \in \mathscr{R}_1^*) + P_{H_1}(Z_1^* \in \mathscr{C}_1^*, Z_2^* \in \mathscr{R}_2^*) + \cdots \\ &\quad + P_{H_1}(Z_1^* \in \mathscr{C}_1^*, \ldots, Z_{K-1}^* \in \mathscr{C}_{K-1}^*, Z_K^* \in \mathscr{R}_K^*) \,.\end{aligned}$$

Under H_1, the vector $(Z_1^*, \ldots, Z_K^*)$ is multivariate normal with mean vector $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_K)$ and correlation matrix (1.6), where ϑ_k is given by (1.5) setting $\mu = \mu_1$. Hence, $(Z_1^* - \vartheta_1, \ldots, Z_K^* - \vartheta_K)$ is multivariate normal with zero mean vector, and the calculation of the power is analogous to the calculation of the Type I error rate with a shift of the continuation and rejection regions. That is, the power is given by

$$\begin{aligned}&P(Z_1^* \in \mathscr{R}_1^* - \vartheta_1) + P(Z_1^* \in \mathscr{C}_1^* - \vartheta_1, Z_2^* \in \mathscr{R}_2^* - \vartheta_2) + \cdots \\ &\quad + P(Z_1^* \in \mathscr{C}_1^* - \vartheta_1, \ldots, Z_{K-1}^* \in \mathscr{C}_{K-1}^* - \vartheta_{K-1}, Z_K^* \in \mathscr{R}_K^* - \vartheta_K) \,,\end{aligned} \tag{1.14}$$

where the random vectors are multivariate normal with zero mean vector, and the subtraction of the intervals with ϑ_k, $k = 1, \ldots, K$, is understood element-wise.

The sample size necessary to achieve a test decision is not fixed but random. The group sequential test procedure is therefore not only assessed by its power but also by its expected or average sample size. The average sample size under H_1, ASN_{H_1}, is given by

$$\mathrm{ASN}_{H_1} = n_1 + \sum_{k=2}^{K} n_k P_{H_1}\left(\bigcap_{\tilde{k}=1}^{k-1} \{Z^*_{\tilde{k}} \in \mathscr{C}^*_{\tilde{k}}\}\right) .$$

It only depends on the continuation regions of the test design since these regions control how many stages will be actually performed. Since, as above, $(Z^*_1 - \vartheta_1, \ldots, Z^*_K - \vartheta_K)$ is multivariate normal with zero mean vector, ASN_{H_1} can be calculated through

$$\mathrm{ASN}_{H_1} = n_1 + \sum_{k=2}^{K} n_k P\left(\bigcap_{\tilde{k}=1}^{k-1} \{Z^*_{\tilde{k}} \in \mathscr{C}^*_{\tilde{k}} - \vartheta_{\tilde{k}}\}\right) ,$$

where again the probability distribution is the multivariate normal distribution with zero mean vector. Note that it might also be of interest to calculate the average sample size if H_0 is true. This provides information about the sample size if there is no effect.

The calculation of the power and the average sample size can be performed, for any K, with the recursive integration formula (see Sect. 1.4). Nevertheless, we will first illustrate the calculation of the power and the average sample size in a two-stage design for testing H_0 assuming equal stage sample sizes. Let the continuation region be given by

$$\mathscr{C}^*_1 = (-2.140; -1.0) \cup (1.0; 2.140) ,$$

and the rejection regions be defined by

$$\mathscr{R}^*_1 = \mathscr{R}^*_2 = (-\infty; -2.140] \cup [2.140; \infty) .$$

This procedure has Type I error rate equal to 0.05, which was shown in the last section. Suppose it is planned to make 20 observations per stage, i.e., $n_1 = n_2 = 20$. The calculation of the power is performed with the bivariate normal integral and suitably shifted continuation and rejection regions. The average sample size under H_1 is computed analogously. Using the common bivariate cdf $F(\cdot, \cdot)$ with correlation

$1/\sqrt{2}$ the expressions for the power and ASN_{H_1} are

$$\begin{aligned}
&P_{H_1}(Z_1^* \in \mathscr{R}_1^*) + P_{H_1}(Z_1^* \in \mathscr{C}_1^*, Z_2^* \in \mathscr{R}_2^*)\\
&\quad = \Phi(-2.140 - \sqrt{20}\delta) + 1 - \Phi(2.140 - \sqrt{20}\delta)\\
&\qquad + \Phi(2.140 - \sqrt{20}\delta) - \Phi(-2.140 - \sqrt{20}\delta) - \Phi(1.0 - \sqrt{20}\delta)\\
&\qquad + \Phi(-1.0 - \sqrt{20}\delta) - \big(F(2.140 - \sqrt{20}\delta, 2.140 - \sqrt{40}\delta)\\
&\qquad - F(-2.140 - \sqrt{20}\delta, 2.140 - \sqrt{40}\delta) - F(1.0 - \sqrt{20}\delta, 2.140 - \sqrt{40}\delta)\\
&\qquad + F(-1.0 - \sqrt{20}\delta, 2.140 - \sqrt{40}\delta)\big) + F(2.140 - \sqrt{20}\delta, -2.140 - \sqrt{40}\delta)\\
&\qquad - F(-2.140 - \sqrt{20}\delta, -2.140 - \sqrt{40}\delta)\\
&\qquad + F(-1.0 - \sqrt{20}\delta, -2.140 - \sqrt{40}\delta) \qquad (1.15)
\end{aligned}$$

and

$$\begin{aligned}
n_1 + n_2 P_{H_1}\big(Z_1^* \in \mathscr{C}_1^*) = 20 + 20\,(&\Phi(2.140 - \sqrt{20}\delta) - \Phi(-2.140 - \sqrt{20}\delta)\\
&- \Phi(1.0 - \sqrt{20}\delta) + \Phi(-1.0 - \sqrt{20}\delta)\big)\ ,
\end{aligned}$$

respectively. Although the expression for the power is rather lengthy, we supply it in order to clarify the necessary calculation for arbitrary δ. The second line in (1.15) is the power of the first stage of the test which is calculated as for a fixed sample size design. Compare Fig. 1.2 to understand the calculation of the integral for the specified regions with the help of the bivariate standard normal cdf.

Figure 1.3 displays the power and the average sample size of this test procedure for the standardized effect size δ within the range $[-1; 1]$. For comparison, Fig. 1.4 displays the power and the average sample size of the test procedure where the study is stopped in the interim analysis only if the null hypothesis can be rejected. A sequence of critical values which fully exhaust the 5 % level is $u_1 = u_2 = 2.178$, as was shown in the last section. That is, the continuation region is given by

$$\mathscr{C}_1^* = (-2.178;\ 2.178)\ ,$$

and the rejection regions are defined by

$$\mathscr{R}_1^* = \mathscr{R}_2^* = (-\infty;\ -2.178] \cup [2.178;\ \infty)\ .$$

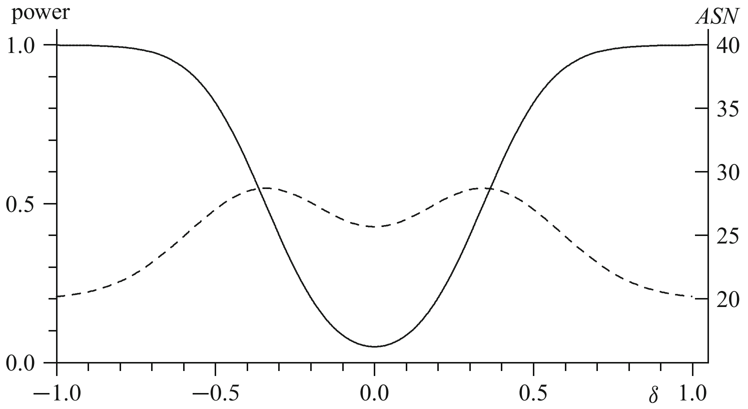

Fig. 1.3 Power (*solid line*) and average sample size (*dashed line*) of a two-stage group sequential test design for testing $H_0 : \mu = \mu_0$ with critical value 2.140 for both stages and continuation region $\mathscr{C}_1^* = (-2.140, -1.0) \cup (1.0, 2.140)$. The stage sample sizes are $n_1 = n_2 = 20$

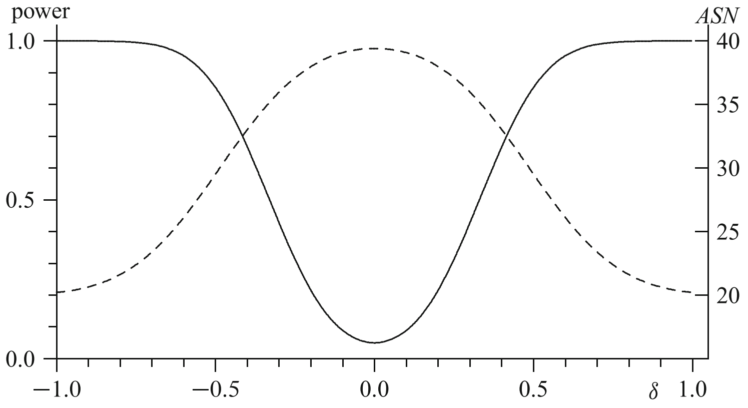

Fig. 1.4 Power (*solid line*) and average sample size (*dashed line*) of a two-stage group sequential test design for testing $H_0 : \mu = \mu_0$ with critical value 2.178 for both stages and continuation region $\mathscr{C}_1^* = (-2.178, 2.178)$. The stage sample sizes are $n_1 = n_2 = 20$

The expressions for the power and the average sample size simplify to

$$\begin{aligned}
&P_{H_1}(Z_1^* \in \mathscr{R}_1^*) + P_{H_1}(Z_1^* \in \mathscr{C}_1^*, Z_2^* \in \mathscr{R}_2^*)\\
&\quad = \Phi(-2.178 - \sqrt{20}\delta) + 1 - \Phi(2.178 - \sqrt{20}\delta)\\
&\qquad + \Phi(2.178 - \sqrt{20}\delta) - \Phi(-2.178 - \sqrt{20}\delta)\\
&\qquad - \big(F(2.178 - \sqrt{20}\delta, 2.178 - \sqrt{40}\delta) - F(-2.178 - \sqrt{20}\delta, 2.178 - \sqrt{40}\delta)\big)\\
&\qquad + F(2.178 - \sqrt{20}\delta, -2.178 - \sqrt{40}\delta) - F(-2.178 - \sqrt{20}\delta, -2.178 - \sqrt{40}\delta)
\end{aligned}$$

and

$$n_1 + n_2 P_{H_1}(Z_1^* \in \mathscr{C}_1^*) = 20 + 20\big(\Phi(2.178 - \sqrt{20}\delta) - \Phi(-2.178 - \sqrt{20}\delta)\big) ,$$

respectively.

We recognize that the power of the two procedures virtually coincide, but the average sample size of the second procedure is much higher for effect sizes close to H_0. This is clear since for effect sizes close to H_0 the probability to stop the trial and not reject H_0 is high for the first procedure. The second procedure does not stop the trial in favor of H_0 and hence the power is somewhat larger, which is true although the critical values are slightly larger. For example, for $\delta = 0.50$ the power values are 0.818 and 0.853, respectively. The (small) gain in power refers to the probability that the test rejects H_0 in the second stage although the observed effect was small in the first stage. For both procedures, if the effect size in absolute terms is large, the probability to reject H_0 after observation of the first stage data comes close to 1. In this case, the average sample size then is close to the sample size of the first stage.

The calculation of the power and the average sample size was performed for given stage sample sizes $n_1 = n_2$ and standardized effect size δ. Conversely, for given δ, one finds the sample sizes $n_1 = n_2$ necessary to achieve a pre-specified power $1 - \beta$. For example, consider the second procedure with critical values $u_1 = u_2 = 2.178$ and continuation region $\mathscr{C}_1^* = (-2.178;\ 2.178)$. Through the use of (1.15) by implementing different values of $n_1 = n_2$ one finds that, for $\delta = 0.40$, the power is 79.7 % if $n_1 = n_2 = 27$. If $n_1 = n_2 = 28$, the power is 81.1 %. Hence, the sample sizes necessary to achieve power $1 - \beta = 80\,\%$ if $\delta = 0.4$ are $n_1 = n_2 = 28$. One can compare the maximum sample size $n_1 + n_2 = 56$ with the sample size necessary in a fixed sample size design to achieve power $1 - \beta = 80\,\%$ if $\delta = 0.40$. It is given by the smallest value of n_f for which

$$P_{H_1}(|Z| \geq 1.96) = \Phi(-1.960 - \sqrt{n_f}\,0.40) + 1 - \Phi(1.960 - \sqrt{n_f}\,0.40) \geq 0.80\ . \tag{1.16}$$

The smallest integer value for which (1.16) is fulfilled is $n_f = 50$. The maximum sample size of the group sequential test design is thus larger by 12 %. The average sample size, on the other hand, is smaller. For $n_1 = n_2 = 28$ the average sample size is 42.7. It is therefore expected that the two-stage test design needs less patients than the fixed sample size design. Still, the maximum sample size is larger and one must therefore balance the reasons for implementing a group sequential test design prior to the start of the trial. In Chaps. 2 and 3 this will be discussed more thoroughly.

Note that one could also calculate the sample sizes $n_1 = n_2$ that yield a specified average sample size, for example, under H_1. Usually, however, when planning a trial, one wants to find the necessary maximum sample size rather the average sample size. Nevertheless, one might wish to find a group sequential plan with an average sample size which is as small as possible. The way of how to define such *optimum* plans will also be described in Chaps. 2 and 3.

1.4 The Recursive Integration Formula

In this section, we will introduce the recursive integration formula for the calculation of the multivariate normal integral involved in group sequential designs. This is a bit technical description and on a first reading this section can be skipped.

Each stage of a group sequential test procedure adds one dimension to the multivariate normal distribution of the random vector $\boldsymbol{Z}^* = (Z_1^*, \ldots, Z_K^*)^T$. Consequently, the calculations needed to determine the Type I error rate, the power and the average sample size of a test design as detailed in the preceding sections can become quite unwieldy. Armitage et al. (1969) expressed the multivariate normal density for $\boldsymbol{Z}^*$ by a recursive formula, sequentially dealing with one of the dimensions at a time. So the calculation of the test characteristics for arbitrarily chosen K becomes computationally feasible.

For technical reasons, we will express the decision regions in terms of the variables

$$S_k = \sum_{\tilde{k}=1}^{k} \sqrt{\frac{n_{\tilde{k}}}{n_1}} Z_{\tilde{k}} \ , \ k = 1, \ldots, K. \tag{1.17}$$

Note that for equal sample sizes S_k is just the sum of the stage-wise test statistics Z_k and is therefore often called the *score statistic* S_k. The variances of the variables S_k are given by $\mathrm{Var}(S_1) = 1$ and

$$\mathrm{Var}(S_k) = \mathrm{Var}(S_{k-1}) + \tau_k \ ,$$

where $\tau_k := \frac{n_k}{n_1}$, $k = 2, \ldots, K$, denotes the standardized time interval between the $(k-1)$th and kth stage relative to n_1. With this notation, the first interim analysis is performed at time $\tau_1 \equiv 1$.

$$\tilde{\vartheta}_k = E(S_k) = \frac{\mu - \mu_0}{\sigma} \frac{1}{\sqrt{n_1}} \sum_{\tilde{k}=1}^{k} n_{\tilde{k}} = \delta \sqrt{n_1} \sum_{\tilde{k}=1}^{k} \tau_{\tilde{k}} \ , \ k = 1, \ldots, K,$$

defines the non-centrality parameter of the distribution for S_k under a specified alternative H_1 [see (1.5)]. The covariance between S_k and $S_{k'}$ is

$$\mathrm{Cov}(S_k, S_{k'}) = \sum_{\tilde{k}=1}^{k^*} \tau_{\tilde{k}}, \text{ where } k^* = \min\{k, k'\} \ .$$

The relationship between the regions $\mathscr{C}_k^*$ and $\mathscr{R}_k^*$ for Z_k^* and the corresponding regions $\mathscr{C}_k$ and $\mathscr{R}_k$ for S_k is as follows:

$$\mathscr{C}_k = \sqrt{\frac{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}}{n_1}}\,\mathscr{C}_k^* = \sqrt{\sum_{\tilde{k}=1}^{k} \tau_{\tilde{k}}}\,\mathscr{C}_k^*\,,$$

$$\mathscr{R}_k = \sqrt{\sum_{\tilde{k}=1}^{k} \tau_{\tilde{k}}}\,\mathscr{R}_k^*, \quad k = 1, \ldots, K,$$

with an element-wise multiplication. The reason for using S_k instead of Z_k^* is that it demands less technical details to derive a density for a (non-standardized) sum of independent statistics. Evidently, using S_k with the continuation regions $\mathscr{C}_k$ will lead to the same test procedure as using Z_k^* with the continuation regions $\mathscr{C}_k^*$ and is just an alternative way to monitor the trial (Proschan et al. 2006).

Describing the density of S_k, note that the outcome of a group sequential design is described by the realizations of two random variables: the test statistic at the end of the trial, S_k, and the number of stages performed until reaching the end of the trial, M. These are clearly highly dependent as the realization of S_k determines at which stage the study stops. The outcome of a group sequential design is therefore described by a bivariate random vector. One of its dimensions—describing the behavior of S_k—is continuous, the other one—describing the number of stages performed—is discrete. In the following, the two-dimensional density of this bivariate random vector is derived iteratively, starting at the first stage and continuing up to the Kth stage.

Consider S_1, the test statistic for the first stage. S_1 is normally distributed with mean $\tilde{\vartheta}_1$ and unit variance. It contributes solely to the outcome of the group sequential design if $s_1 \notin \mathscr{C}_1$. The first part of the recursive integration formula is therefore given by

$$f_{\delta}(s_1, 1) = \begin{cases} \varphi(s_1 - \tilde{\vartheta}_1) & s_1 \notin \mathscr{C}_1 \\ 0 & \text{else} \end{cases}, \tag{1.18}$$

where $\varphi(\cdot)$ denotes the standard normal density. Note that (1.18) is not a density since it accounts only for one of the possible realizations of the number of stages performed, namely the case that the study is stopped after the first stage. In order to make (1.18) a density, a standardization would be needed; we would then obtain the density of a truncated normal distribution (not truncated at the tails, as usual, but rather truncated in the middle). Yet with such a standardization, the formula could only describe the behavior of S_1 and could not cover subsequent stages. The aim of describing the behavior of the test statistic at the end of the trial would not be met. We will nevertheless deal with such standardizations later on as they are of interest in the context of point estimation.

If $s_1 \in \mathscr{C}_1$, (1.18) does not apply. The study is then continued and the weighted sum $S_2 = S_1 + \sqrt{\tau_2} Z_2$ is calculated, yielding the test statistic for the second stage. This sum consists of two addends, which are the test statistic S_1 from the first stage and the increment $S_2 - S_1 = \sqrt{\tau_2} Z_2$. These addends are independent due to the preplanned sample sizes. Note that this independency assumption is true only for S_1 and the increment $S_2 - S_1$ rather than for S_1 and S_2. The density of S_2 is therefore obtained by convoluting S_1 and $S_2 - S_1$, where $S_2 - S_1$ is normally distributed with mean $\tilde{\vartheta}_2 - \tilde{\vartheta}_1$ and variance τ_2. S_2 in turn contributes solely to the outcome of the group sequential design if $s_2 \notin \mathscr{C}_2$. The second part of the density is therefore given by

$$f_\delta(s_2, 2) = \begin{cases} \int_{\mathscr{C}_1} \frac{1}{\sqrt{\tau_2}} \varphi(\frac{(s_2 - s_1) - (\tilde{\vartheta}_2 - \tilde{\vartheta}_1)}{\sqrt{\tau_2}}) \, \varphi(s_1 - \tilde{\vartheta}_1) \, ds_1 & s_2 \notin \mathscr{C}_2 \\ 0 & \text{else} \; . \end{cases} \tag{1.19}$$

Again, for the same reasons as (1.18), (1.19) is not a density.

This approach to determine the partial densities continues for all stages $k = 1, \ldots, K$: The test statistic for the kth stage, S_k, can always be expressed as the sum of the test statistic S_{k-1} from the previous stage and the increment $S_k - S_{k-1} = \sqrt{\tau_k} Z_k$, the increment being independent of S_{k-1} and normally distributed with mean $\tilde{\vartheta}_k - \tilde{\vartheta}_{k-1}$ and variance τ_k. The group sequential density is therefore given by

$$f_\delta(s_k, k) = \begin{cases} p_\delta(s_k, k) & s_k \notin \mathscr{C}_k \vee k = K \\ 0 & \text{else} \; , \end{cases} \tag{1.20}$$

where

$$p_\delta(s_k, k) = \int_{\mathscr{C}_{k-1}} p_\delta(s_{k-1}, k-1) \, \frac{1}{\sqrt{\tau_k}} \, \varphi\left(\frac{(s_k - s_{k-1}) - (\tilde{\vartheta}_k - \tilde{\vartheta}_{k-1})}{\sqrt{\tau_k}} \right) \, ds_{k-1} \; , \tag{1.21}$$

$k = 2, \ldots, K$, and $p_\delta(s_1, 1)$ is given by

$$p_\delta(s_1, 1) = \varphi(s_1 - \tilde{\vartheta}_1) \; .$$

The presentation (1.21) is due to Armitage et al. (1969). It involves a repeated numerical integration technique which can be solved, for example, with the Newton–Cotes method.

As a density, (1.20) fulfills the condition

$$\sum_{k=1}^{K} \int_{\mathbb{R}\setminus\mathscr{C}_k} f_\delta(s_k, k)\, ds_k = 1 \ .$$

That is, integrating every partial density over its domain, the complement of the respective continuation region, and then summing up the results of those integrations for all stages yields the total probability mass.

Chapter 2
Procedures with Equally Sized Stages

In this chapter, we describe group sequential test procedures that are designed for equal sample sizes per stage of the group sequential trial, i.e., $n_1 = \cdots = n_K = n$. The procedures that were originally developed in the literature (which we refer to as *classical group sequential designs*) make this assumption. In practice, the situation with equally sized stages often occurs. Namely, in all cases when there is no specific reason for assuming different stage sizes the sample sizes per stage should be the same. Nevertheless, it might be questionable if equal sample sizes are practically feasible. Particularly, regardless of whether equal stage sizes are planned, the concrete implementation of an interim analysis will rarely be based on equally sized groups of observations. A more or less slight departure will occur in nearly all cases since usually observations at hand are used for the interim stage. Especially from an organizational perspective, the theoretical requirement is therefore hard to fulfill. We will see, however, that the assumption of equally sized stages can be weakened, or other designs are perhaps better suited. These more general designs will be studied in Chap. 3.

As in the last chapter, we assume independent and normally distributed observations with a known variance. The hypothesis to be tested is $H_0 : \mu = \mu_0$. Assuming n observations per stage and a standardized effect size $\delta = (\mu - \mu_0)/\sigma$, the expressions (1.3)–(1.5) simplify to

$$Z_k^* = \frac{1}{\sqrt{k}} \sum_{\tilde{k}=1}^{k} Z_{\tilde{k}} \,,$$

$$\mathrm{Cov}(Z_k^*, Z_{k'}^*) = \sqrt{\frac{k}{k'}} \,, \text{ for } k \le k',$$

$$\vartheta_k = E(Z_k^*) = \delta \sqrt{k\,n} \,, \; k = 1, \ldots, K.$$

G. Wassmer, W. Brannath, *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*, Springer Series in Pharmaceutical Statistics,
DOI 10.1007/978-3-319-32562-0_2

In the following, the procedures will be defined in terms of the continuation regions $\mathscr{C}_k^*$ for the standardized test statistics Z_k^* rather than in terms of the continuation regions $\mathscr{C}_k$ for the test statistic S_k as introduced in Sect. 1.4. Obviously, the two modes of presentation are completely equivalent. In the literature, however, both modes of presentation were used.

2.1 Classical Designs

2.1.1 Definition

Pocock (1977) and O'Brien and Fleming (1979) proposed group sequential plans for the two-sided testing problem. That is, H_0 is tested against the two-sided alternative $H_1 : \mu \neq \mu_0$. These designs have found widespread acceptance in the scientific community. In clinical research, they are often used and can hence be regarded as standard techniques. In defining a group sequential plan, Pocock (1977) assumed constant critical values $u_1 = \cdots = u_K =: u'$ for Z_k^*, $k = 1, \ldots, K$, over the stages of the trial. That is, instead of using the unadjusted critical value $u = \Phi(1 - \alpha/2)$ at each stage of the testing procedure, an adjusted critical value $u' > u$ has to be used in order to obtain a level α testing procedure. The latter constant critical value u' of Pocock's design is chosen to give overall Type I error rate α and hence it is defined through

$$P_{H_0}(|Z_1^*| \geq u' \text{ or} \ldots \text{or } |Z_K^*| \geq u') = \alpha \ .$$

u' depends on K and α and is denoted by $c_{\mathrm{P}}(K, \alpha)$. The procedure of O'Brien and Fleming (1979), on the other hand, is characterized by monotonically decreasing critical values defined by $u_k = c_{\mathrm{OBF}}(K, \alpha)/\sqrt{k}$, where $c_{\mathrm{OBF}}(K, \alpha)$ is a constant that ensures that

$$P_{H_0}(|Z_1^*| \geq u_1 \text{ or} \ldots \text{or } |Z_K^*| \geq u_K) = \alpha$$

is fulfilled. Formally, the continuation regions $\mathscr{C}_k^*$ for Z_k^* are given by

- $\mathscr{C}_k^* = (-u'; u')$ with $u' = c_{\mathrm{P}}(K, \alpha)$, $k = 1, \ldots, K$, for Pocock's design,
- $\mathscr{C}_k^* = (-u_k; u_k)$ with $u_k = c_{\mathrm{OBF}}(K, \alpha)/\sqrt{k}$, $k = 1, \ldots, K$, for O'Brien and Fleming's design.

The study is stopped with the rejection of H_0 if $Z_k^* \in \overline{\mathscr{C}}_k^*$ for some $k = 1, \ldots, K-1$, otherwise the trial is continued. At the Kth stage, H_0 is rejected if $Z_K^* \in \overline{\mathscr{C}}_K^*$, otherwise it is not rejected. In other words, in this simple case the rejection regions $\mathscr{R}_k^*$ are given by the complement of $\mathscr{C}_k^*$, $k = 1, \ldots, K$.

The group sequential tests can also be defined by means of the *adjusted nominal significance levels* $\alpha' = 2\,(1 - \Phi(u'))$ and $\alpha_k = 2\,(1 - \Phi(u_k))$, respectively. The study is stopped with the rejection of H_0 if the (two-sided) p-value of the test statistic calculated at the kth stage, which is given by $2\,(1 - \Phi(|z_k^*|))$ is below the level α' and α_k, respectively. This once again illustrates the idea of "repeated significance tests": Repeatedly, at each stage k, a significance test at the corresponding adjusted significance level is performed.

The constants $c_{\mathrm{P}}(K, \alpha)$ and $c_{\mathrm{OBF}}(K, \alpha)$ can be computed numerically using the recursive formula described in the last section. Table 2.1 supplies the constants $c_{\mathrm{P}}(K, \alpha)$ and $c_{\mathrm{OBF}}(K, \alpha)$ for a number of values for K and α. For Pocock's design, at each stage k the standardized test statistic Z_k^* is compared with $u' = c_{\mathrm{P}}(K, \alpha)$, whereas, for O'Brien and Fleming's design, the sequence of critical values is calculated through $u_k = c_{\mathrm{OBF}}(K, \alpha)/\sqrt{k}$, $k = 1, \ldots, K$. Additionally, the last stage critical value for O'Brien and Fleming's design is provided in the table. The validity of the constants $c_{\mathrm{P}}(K, \alpha)$ and $c_{\mathrm{OBF}}(K, \alpha)$ have been extensively proofed. They are correct up to the stated decimal places.

The table shows that, as K increases, the necessary adjustment becomes stronger. In O'Brien and Fleming's design, for suitably large K, it is extremely unlikely to terminate the trial at a very early stage but it is easier to reject H_0 later on. As a consequence, the last stage critical value is near to the critical value of the two-sided fixed sample size design. For example, for $\alpha = 0.05$, the latter is given by 1.960 whereas in a five-group design the critical value is given by 2.040. In other words, the price to pay for interim looks in terms of having to use a more conservative level of the Type I error rate is low when using O'Brien and Fleming's test design. The situation where the final test statistic comes up with a p-value lower than α but the null hypothesis cannot be rejected is therefore unlikely to arise in O'Brien and Fleming's design. Although this is clear from the statistical point of view, it might be bothering to practitioners. Because it is more likely in Pocock's design, O'Brien and Fleming's test design might be preferred.

For illustrating the difference between the two designs, Table 2.2 presents the critical values u_k, $k = 1, \ldots, K$, and u', together with the adjusted nominal significance levels α_k and α' according to O'Brien & Fleming's and Pocock's design, respectively, for $\alpha = 0.05$ and $K = 2, 3, 4, 5$. The decision regions of the two designs for $K = 5$ and $\alpha = 0.05$ are illustrated in Fig. 2.1.

Table 2.1 Constants $c_{\text{OBF}}(K, \alpha)$ and $c_{\text{P}}(K, \alpha)$ to determine the sequence of critical values according to O'Brien & Fleming's and Pocock's design, respectively

	K	$\alpha = 0.001$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
OBrien and Fleming	1	3.2905	2.5758	1.9600	1.6449
	2	4.6541 (3.2909)	3.6481 (2.5796)	2.7965 (1.9774)	2.3730 (1.6780)
	3	5.7096 (3.2964)	4.4945 (2.5949)	3.4711 (2.0040)	2.9611 (1.7096)
	4	6.6093 (3.3047)	5.2182 (2.6091)	4.0486 (2.0243)	3.4662 (1.7331)
	5	7.4076 (3.3128)	5.8611 (2.6212)	4.5617 (2.0401)	3.9151 (1.7509)
	6	8.1328 (3.3202)	6.4455 (2.6314)	5.0283 (2.0528)	4.3231 (1.7649)
	7	8.8020 (3.3268)	6.9849 (2.6401)	5.4590 (2.0633)	4.6998 (1.7763)
	8	9.4265 (3.3328)	7.4884 (2.6476)	5.8611 (2.0722)	5.0514 (1.7859)
	9	10.014 (3.3381)	7.9623 (2.6541)	6.2395 (2.0798)	5.3824 (1.7941)
	10	10.571 (3.3428)	8.4113 (2.6599)	6.5981 (2.0865)	5.6959 (1.8012)
	11	11.101 (3.3472)	8.8390 (2.6651)	6.9396 (2.0924)	5.9946 (1.8074)
	12	11.609 (3.3511)	9.2481 (2.6697)	7.2663 (2.0976)	6.2803 (1.8130)
	13	12.096 (3.3547)	9.6408 (2.6739)	7.5799 (2.1023)	6.5546 (1.8179)
	14	12.565 (3.3580)	10.019 (2.6777)	7.8820 (2.1065)	6.8187 (1.8224)
	15	13.017 (3.3611)	10.384 (2.6812)	8.1736 (2.1104)	7.0737 (1.8264)
	20	15.087 (3.3735)	12.053 (2.6951)	9.5062 (2.1257)	8.2391 (1.8423)
Pocock	1	3.2905	2.5758	1.9600	1.6449
	2	3.4634	2.7718	2.1783	1.8754
	3	3.5542	2.8730	2.2895	1.9922
	4	3.6136	2.9387	2.3613	2.0674
	5	3.6570	2.9863	2.4132	2.1217
	6	3.6905	3.0231	2.4532	2.1635
	7	3.7177	3.0528	2.4855	2.1973
	8	3.7403	3.0775	2.5123	2.2253
	9	3.7597	3.0986	2.5352	2.2492
	10	3.7764	3.1169	2.5550	2.2699
	11	3.7912	3.1329	2.5724	2.2881
	12	3.8043	3.1472	2.5880	2.3043
	13	3.8161	3.1601	2.6019	2.3189
	14	3.8268	3.1718	2.6146	2.3321
	15	3.8366	3.1824	2.6261	2.3441
	20	3.8754	3.2247	2.6720	2.3921

In parentheses: last stage critical value for O'Brien and Fleming's design

Pocock's design requires less conservative levels for early stages and hence it is more likely to terminate the trial early. Stopping early, however, depends on the sample size chosen, and the probability of reaching a rejection of H_0 at some stage given a sequence of sample sizes (i.e., the power of the test) or at specific stages will be the crucial point. The pros and cons for choosing a specific design will thus

Table 2.2 Critical values u_k and u' for Z_k^* in the designs of O'Brien & Fleming and Pocock, respectively; $\alpha = 0.05$

	K	u_1	u_2	u_3	u_4	u_5
O'Brien and Fleming	2	2.797	1.977			
		(0.0052)	(0.0480)			
	3	3.471	2.454	2.004		
		(0.0005)	(0.0141)	(0.0451)		
	4	4.049	2.863	2.337	2.024	
		(0.00005)	(0.0042)	(0.0194)	(0.0429)	
	5	4.562	3.226	2.634	2.281	2.040
		(0.000005)	(0.0013)	(0.0084)	(0.0226)	(0.0413)
Pocock	2	2.178	2.178			
		(0.0294)	(0.0294)			
	3	2.289	2.289	2.289		
		(0.0221)	(0.0221)	(0.0221)		
	4	2.361	2.361	2.361	2.361	
		(0.0182)	(0.0182)	(0.0182)	(0.0182)	
	5	2.413	2.413	2.413	2.413	2.413
		(0.0158)	(0.0158)	(0.0158)	(0.0158)	(0.0158)

In parentheses: corresponding adjusted nominal significance levels α_k and α', respectively

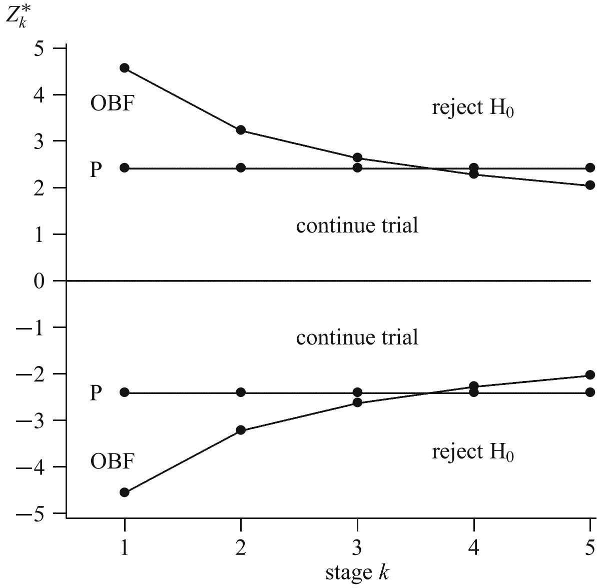

Fig. 2.1 Decision regions for O'Brien and Fleming's (OBF) and Pocock's (P) design for equal stage sizes; $K = 5$, $\alpha = 0.05$

be discussed more extensively after we have presented the power and the average sample size characteristics of O'Brien & Fleming's and Pocock's design.

An interesting feature of O'Brien and Fleming's test is the following. Since $\mathscr{C}_k = \sqrt{k}\mathscr{C}_k^*$, the critical values for the non-standardized test statistic S_k are constant in k. That is, H_0 can be rejected until the sum of the stage-wise test statistics exceeds a specified constant $c_{\mathrm{OBF}}(K, \alpha)$. This corresponds to stop and reject H_0 if the likelihood ratio exceeds a constant boundary. Hence, it appears to be a reasonable technique for defining a group sequential test design: As long as summing up the test statistics does not exceed a constant threshold, H_0 cannot be rejected and further observations are needed for possibly rejecting H_0.

2.1.2 Power and Average Sample Size

In a fixed sample size design with sample size n_f, the power of the level-α test for testing H_0 is given by

$$\begin{aligned} P_{H_1}(|Z| \geq \Phi^{-1}(1-\alpha/2)) = \Phi(-\Phi^{-1}(1-\alpha/2) - \sqrt{n_f}\delta) + 1 \\ - \Phi(\Phi^{-1}(1-\alpha/2) - \sqrt{n_f}\delta) \;. \end{aligned}$$

If one wants to find the necessary sample size to achieve a prespecified power $1-\beta$, δ denotes the standardized effect which is worthwhile to detect and the interest lies in either positive or negative values of δ. As a consequence, in practical cases, either $\Phi(-\Phi^{-1}(1-\alpha/2) - \sqrt{n_f}\delta)$ or $1 - \Phi(\Phi^{-1}(1-\alpha/2) - \sqrt{n_f}\delta)$ is extremely small, and the power requirement reads as

$$P_{H_1}(|Z| \geq \Phi^{-1}(1-\alpha/2)) \approx 1 - \Phi(\Phi^{-1}(1-\alpha/2) - \sqrt{n_f}|\delta|) = 1 - \beta \;. \tag{2.1}$$

From (2.1) one finds the (approximate) sample size formula

$$n_f = \frac{\left(\Phi^{-1}(1-\alpha/2) + \Phi^{-1}(1-\beta)\right)^2}{\delta^2} \tag{2.2}$$

for the fixed sample size design. In practical applications, the actual sample size is the next integer greater than n_f obtained from (2.2). For the following calculations, however, the exact version of the sample size calculation will be used. That is, n_f is found through

$$\Phi(-\Phi^{-1}(1-\alpha/2) - \sqrt{n_f}\delta) + 1 - \Phi(\Phi^{-1}(1-\alpha/2) - \sqrt{n_f}\delta) = 1 - \beta \;,$$

and the decimal number n_f will be applied for further calculations.

The power of a group sequential test considered in this section is given by

$$1 - P(Z_1^* \in \mathscr{C}_1^* - \vartheta_1, \dots, Z_K^* \in \mathscr{C}_K^* - \vartheta_K) , \tag{2.3}$$

where $\vartheta_k = \delta\sqrt{k\,n}$, $k = 1, \dots, K$, and the substraction is understood element-wise (see (1.14) in Sect. 1.3). The random vector $(Z_1^*, \dots, Z_K^*)$ in (2.3) is multivariate normal with zero mean vector. Hence, at given α, $1-\beta$, and K, solving (2.3) through the use of the recursive integration formula (with $\delta = 0$), one finds the shift value $\vartheta^* = \vartheta^*(K, \alpha, \beta)$ such that the power (2.3) with $\vartheta_k = \vartheta^*\sqrt{k}$, $k = 1, \dots, K$, equals $1-\beta$. The sample size per stage of the group sequential test design, n, is then given by

$$n = n(K, \alpha, \beta) = \frac{\vartheta^{*2}}{\delta^2} ,$$

and the maximum sample size is

$$N = K\,n . \tag{2.4}$$

The sample size formulas (2.2) and (2.4) are both inversely proportional in δ^2. This motivates the following definition: The *inflation factor* $I = I(K, \alpha, \beta)$ is the ratio

$$\frac{N}{n_f} , \tag{2.5}$$

relating the sample size of a group sequential test to its corresponding fixed sample size test. It is independent of the standardized effect size δ, and will serve as a basis for sample size calculations in group sequential test designs under very different testing situations (see Chap. 4). The average sample size under H_1, ASN_{H_1}, given K, α, and $1 - \beta$, is inversely proportional to δ^2, too. This easily follows from the representation

$$\mathrm{ASN}_{H_1} = \frac{\vartheta^{*2}}{\delta^2} + \sum_{k=2}^{K} \frac{\vartheta^{*2}}{\delta^2} P\left(\bigcap_{\tilde{k}=1}^{k-1} \{Z_{\tilde{k}}^* \in \mathscr{C}_{\tilde{k}}^* - \vartheta^*\sqrt{\tilde{k}}\} \right) .$$

That is, it suffices to calculate the average sample size for a specific value of δ, for example, $\delta = 1$. For $\delta \neq 1$, ASN_{H_1} is obtained by dividing the average sample size calculated for $\delta = 1$ by δ^2. Furthermore, the ratio

$$\frac{\mathrm{ASN}_{H_1}}{n_f} \tag{2.6}$$

is also independent of δ and characterizes the average sample size reduction when considering a group sequential test design as compared to the fixed sample size

Table 2.3 Inflation factor $I = I(K, \alpha, \beta)$ and expected reduction in sample size under H_1 relative to n_f, ASN_{H_1}/n_f, (in parentheses) for the designs of O'Brien & Fleming and Pocock, respectively, for different values of K, significance level α, and power $1 - \beta$

		$1-\beta=0.80$		$1-\beta=0.90$	
	K	$\alpha=0.01$	$\alpha=0.05$	$\alpha=0.01$	$\alpha=0.05$
O'Brien and Fleming	1	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
	2	1.001 (0.947)	1.008 (0.902)	1.001 (0.912)	1.007 (0.851)
	3	1.007 (0.886)	1.017 (0.856)	1.006 (0.837)	1.016 (0.799)
	4	1.011 (0.862)	1.024 (0.831)	1.010 (0.806)	1.022 (0.767)
	5	1.015 (0.847)	1.028 (0.818)	1.014 (0.789)	1.026 (0.750)
	6	1.017 (0.838)	1.032 (0.809)	1.016 (0.777)	1.030 (0.739)
	7	1.019 (0.831)	1.035 (0.802)	1.018 (0.769)	1.032 (0.732)
	8	1.021 (0.826)	1.037 (0.798)	1.020 (0.763)	1.034 (0.726)
	9	1.022 (0.822)	1.038 (0.794)	1.021 (0.758)	1.036 (0.721)
	10	1.024 (0.819)	1.040 (0.791)	1.022 (0.754)	1.037 (0.718)
	11	1.025 (0.817)	1.041 (0.789)	1.023 (0.752)	1.039 (0.715)
	12	1.026 (0.815)	1.042 (0.787)	1.024 (0.749)	1.040 (0.713)
	13	1.026 (0.813)	1.043 (0.786)	1.025 (0.747)	1.041 (0.711)
	14	1.027 (0.812)	1.044 (0.784)	1.026 (0.745)	1.041 (0.709)
	15	1.028 (0.810)	1.045 (0.783)	1.026 (0.744)	1.042 (0.708)
	20	1.030 (0.806)	1.047 (0.779)	1.029 (0.739)	1.045 (0.703)
Pocock	1	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
	2	1.092 (0.872)	1.110 (0.853)	1.083 (0.798)	1.100 (0.776)
	3	1.137 (0.841)	1.166 (0.818)	1.125 (0.750)	1.151 (0.721)
	4	1.166 (0.828)	1.202 (0.805)	1.152 (0.728)	1.183 (0.697)
	5	1.187 (0.822)	1.228 (0.799)	1.171 (0.717)	1.206 (0.685)
	6	1.203 (0.818)	1.249 (0.796)	1.185 (0.709)	1.225 (0.677)
	7	1.216 (0.816)	1.265 (0.795)	1.197 (0.705)	1.239 (0.673)
	8	1.226 (0.816)	1.278 (0.794)	1.206 (0.701)	1.251 (0.669)
	9	1.235 (0.815)	1.290 (0.794)	1.214 (0.699)	1.262 (0.667)
	10	1.243 (0.816)	1.300 (0.795)	1.222 (0.698)	1.271 (0.666)
	11	1.250 (0.816)	1.309 (0.796)	1.228 (0.697)	1.279 (0.665)
	12	1.257 (0.816)	1.317 (0.797)	1.234 (0.696)	1.286 (0.664)
	13	1.262 (0.817)	1.325 (0.797)	1.239 (0.695)	1.292 (0.664)
	14	1.267 (0.817)	1.331 (0.798)	1.243 (0.695)	1.298 (0.664)
	15	1.272 (0.818)	1.337 (0.799)	1.247 (0.695)	1.304 (0.663)
	20	1.291 (0.822)	1.362 (0.805)	1.264 (0.695)	1.325 (0.664)

design under the assumption that H_1 is true. In Table 2.3 the quantities (2.5) and (2.6) are provided for O'Brien & Fleming's and Pocock's design for $K = 1, \ldots, 15, 20$, $\alpha = 0.01,\ 0.05$, and $1 - \beta = 0.80,\ 0.90$. The case $K = 1$ refers to the fixed sample size design and is added for the sake of completeness.

For both designs, Table 2.3 shows that the maximum sample size is increasing in K. For O'Brien and Fleming's design, the average sample size under H_1 is decreasing in K, whereas for Pocock's design it is only decreasing in K for small and moderate K whereas for large values of K it slightly increases again. The expected reduction in the sample size increases with the postulated power and is more distinct for Pocock's design than for O'Brien and Fleming's design. For Pocock's design, on the other hand, the maximum necessary sample is higher than for O'Brien and Fleming's design.

By an example, we illustrate the use of Table 2.3 for practical sample size calculations. Suppose, in a four-stage two-sided group sequential design at significance level $\alpha = 0.05$, the necessary maximum and average sample size are to be calculated for power $1 - \beta = 0.80$ and a standardized effect $|\delta| = 0.50$. That is, the alternative to be detected is $H_1 : |\delta| = 0.50$. From (2.2), the sample size in a fixed sample size design is

$$n_f = \frac{(1.96 + 0.842)^2}{0.50^2} = 31.4\ .$$

If one wants to use O'Brien and Fleming's design, the maximum sample size is

$$N = 1.024 \times 31.4 = 32.2\ ,$$

and hence one needs $32.2/4 = 8.05$ observations per stage. Using Pocock's design, the maximum sample size is

$$N = 1.202 \times 31.4 = 37.7\ ,$$

and hence $37.7/4 = 9.4$ observations per stage are necessary to achieve 80 % power. If the alternative is true, the average sample size, under H_1, for O'Brien and Fleming's design is

$$\text{ASN}_{H_1} = 0.831 \times 31.4 = 26.1\ ,$$

and for Pocock's design it is

$$\text{ASN}_{H_1} = 0.805 \times 31.4 = 25.3\ .$$

To ensure the Type I error rate and the postulated power of the design, the sample size per stage to choose in O'Brien and Fleming's design is $n = 9$, and in Pocock's design it is $n = 10$. Clearly, the attained power is then slightly larger than the postulated power, and the actual maximum sample size, especially in O'Brien and Fleming's design, exceeds N to a quite large amount. It might thus be desirable to use, say, a maximum of 33 observations and to use slightly differing sample sizes per stage, for example, $n_1 = 9$, $n_2 = 8$, $n_3 = 8$, $n_4 = 8$. Although the assumption of equal group sizes is then (slightly) violated, we will see that this is of no practical concern. In this example, however, we proceed to use the exact values to provide some insight into the theoretical performance of the procedures. For larger sample size, of course, this problem is only of minor concern. We will also reconsider this example when treating the unknown variance case (see Sect. 5.1).

Both designs are quite effective in reducing the expected sample size. To understand why, under H_1, Pocock's design is expected to need less observations, consider the probabilities to stop at one of the interim stages. For O'Brien and Fleming's design, under H_1, the probabilities to stop at the first, second, and third stage are given by 0.4 %, 19.1 %, and 35.7 %, respectively. The probability to reach the maximum sample size is 44.8 % (and the probability to reach a significant result at this stage is 22.8 %). For Pocock's design, the corresponding probabilities are 20.5 %, 25.2 %, 20.3 %, and 34.0 % (14.0 %), respectively. Hence, it is much more likely to stop at the first or the second stage for the latter design, resulting in a smaller average sample size under H_1 for Pocock's design although the maximum sample size is higher. We further note that the standard deviation of the sample size is greater for Pocock's design as compared to O'Brien and Fleming's design. With the above probabilities the standard deviations are calculated as

$$\left(0.004\,(8.05-26.1)^2+\cdots+0.448\,(32.2-26.1)^2\right)^{1/2}=6.24$$

and

$$\left(0.205\,(9.4-25.3)^2+\cdots+0.340\,(37.7-25.3)^2\right)^{1/2}=10.75\ ,$$

respectively. Hence, in addition to the larger maximum sample size, the greater variability of the sample size in Pocock's design might be regarded as an unfavorable feature of this test procedure but note that this is an intrinsic feature of stopping the trial early.

It is also of interest to consider the average sample size under the assumption that H_1 is not true, i.e., for $\delta \neq |0.50|$. The calculation is straightforward with the recursive integration formula using $\vartheta_k = \delta\sqrt{kn}$, $k = 1,\ldots,K$, where n is the sample size per stage necessary to achieve power at $|\delta| = 0.50$. That is, the average sample is given by

$$\mathrm{ASN}=\frac{\vartheta^{*2}}{\delta^2}+\sum_{k=2}^{K}\left(\frac{\vartheta^{*2}}{\delta^2}\right)P\left(\bigcap_{\tilde{k}=1}^{k-1}\{Z^*_{\tilde{k}}\in\mathscr{C}^*_{\tilde{k}}-\tilde{\delta}\sqrt{\tilde{k}\,n}\}\right)\ ,$$

which can be calculated for different values of $\tilde{\delta}$. For example, $\tilde{\delta} = 0$ refers to the average sample size if H_0 is true. Figure 2.2 illustrates the average sample size for values of $\tilde{\delta}$ ranging from -1 to 1, where the sample size of the fixed sample size design is added as a reference line.

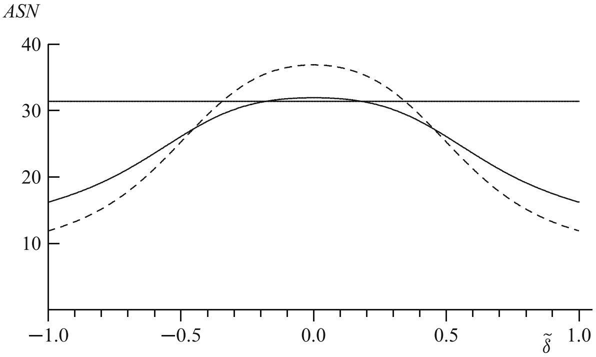

Fig. 2.2 Average sample size of O'Brien and Fleming's (*solid line*) and Pocock's (*dashed line*) test. The sample size was calculated to achieve power $1 - \beta = 0.80$ at $|\delta| = 0.50$ ($\alpha = 0.05$, four-stage design). *Horizontal line*: sample size for fixed sample design

Pocock's test has a smaller average sample size for values of $|\tilde{\delta}|$ near and larger than the assumed one. Due to the lower maximum sample size, however, O'Brien and Fleming's test has smaller average sample size if $|\delta|$ is considerably smaller than anticipated. The gain in average sample size of Pocock's test increases for $|\delta| > 0.50$, but, for very large values, both tests reject H_0 at the first stage and hence O'Brien and Fleming's test is (slightly) better in terms of saving sample size due to the smaller first stage sample size. It was already mentioned in Sect. 1.3 that the average sample size can be reduced if the study is stopped for futility (i.e., stopping with the non-rejection of H_0) when only a small effect is observed. These designs will be thoroughly discussed later on. For the moment, we recapitulate that Pocock's design turns out to be favorable in terms of the expected saving in sample size for suitably chosen significance level and power. Despite this, users might be deterred by the larger maximum sample size and perhaps are not even willing to stop the trial in very early stages, for example, because of safety issues (see Sect. 2.3). That is, the average sample size is not always the key issue, and O'Brien and Fleming's design is therefore often preferable.

Another important issue is the choice of the maximum number of stages, K, to be performed. The most relevant expected reduction in the sample size is already reached for a moderate number of stages, especially for Pocock's design. This is illustrated in Fig. 2.3 for power $1 - \beta = 0.80$ and significance level $\alpha = 0.05$. It shows that for Pocock's design it suffices to consider at most a maximum of, say, $K = 5$ stages (Pocock 1982; McPherson 1982), whereas for O'Brien and Fleming's design it might be reasonable to consider up to, say, $K = 8$ stages in order to receive a substantial reduction in the average sample size. Nevertheless, it is worth mentioning that for both designs even the implementation of one or two interim analyses is quite effective in reducing the average sample size.

Although an important issue in clinical trials is saving sample size, logistic reasoning about the feasibility of interim analyses should influence the decision

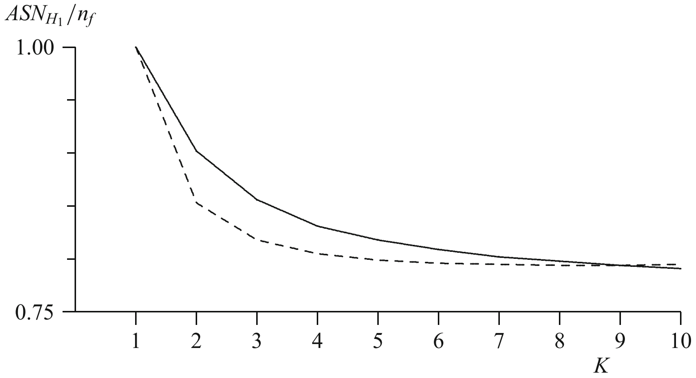

Fig. 2.3 Expected reduction in sample size under H_1 relative to the sample size in a fixed sample size design, ASN_{H_1}/n_f, of O'Brien and Fleming's (*solid line*) and Pocock's (*dashed line*) test for different K; power $1-\beta = 0.80$, significance level $\alpha = 0.05$

concerning the number of stages, too. The preparation of interim reports and the organization of the meetings for the independent Data Monitoring Committee (iDMC) might be time consuming and costly. One should also bear in mind that an interim analysis potentially involves unblinding the data and therefore other than permitted people could become aware of the results. As a consequence, the result of the whole trial could then be biased. Sound practical reasons might therefore bring about a decision of choosing, say, a four-stage design, or even a design with only one or two interim analyses.

2.1.3 Wang and Tsiatis Power Family

Wang and Tsiatis (1987) proposed a class of boundaries indexed by a power parameter $\varDelta$. The continuation regions $\mathscr{C}_k^*$ for Z_k^* are given by

$$\mathscr{C}_k^* = (-u_k;\, u_k), \text{ where } u_k = c_{\mathrm{WT}}(K,\alpha,\varDelta)\, k^{\varDelta-0.5},\; k = 1,\ldots,K. \tag{2.7}$$

As above, the rejection regions are the complement of $\mathscr{C}_k^*$, i.e., $\mathscr{R}_k^* = \overline{\mathscr{C}}_k^*$, $k = 1,\ldots,K$. The $\varDelta$-class of boundaries as given in (2.7) yields O'Brien and Fleming's test for $\varDelta = 0$ and Pocock's test for $\varDelta = 0.50$. The constants $c_{\mathrm{WT}}(K,\alpha,\varDelta)$ are tabulated in Table 2.4 for a number of values $\varDelta$, α, and K. Values for $K > 10$ were omitted since from the above discussion it follows that even $K > 5$ will hardly occur in practice. We mention that the constants provided in Table 2.4 slightly differ from the constants provided in Table 1 of Wang and Tsiatis (1987). Nevertheless, we feel confident that our figures are correct up to the stated decimal places (Wassmer and Bock 1999).

Table 2.4 Constants $c_{\mathrm{WT}}(K, \alpha, \Delta)$ for the Wang and Tsiatis design

	K	$\alpha = 0.001$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
$\Delta = 0.10$	1	3.2905	2.5758	1.9600	1.6449
	2	4.3447	3.4136	2.6314	2.2425
	3	5.1276	4.0496	3.1442	2.6943
	4	5.7743	4.5752	3.5692	3.0690
	5	6.3341	5.0304	3.9371	3.3936
	6	6.8327	5.4356	4.2645	3.6823
	7	7.2851	5.8034	4.5614	3.9442
	8	7.7012	6.1415	4.8344	4.1848
	9	8.0878	6.4557	5.0879	4.4082
	10	8.4500	6.7500	5.3253	4.6174
$\Delta = 0.25$	1	3.2905	2.5758	1.9600	1.6449
	2	3.9331	3.1131	2.4239	2.0777
	3	4.3860	3.4906	2.7411	2.3674
	4	4.7420	3.7873	2.9887	2.5915
	5	5.0385	4.0341	3.1941	2.7767
	6	5.2943	4.2468	3.3708	2.9357
	7	5.5204	4.4344	3.5265	3.0756
	8	5.7237	4.6030	3.6662	3.2011
	9	5.9089	4.7564	3.7932	3.3151
	10	6.0792	4.8975	3.9099	3.4198
$\Delta = 0.40$	1	3.2905	2.5758	1.9600	1.6449
	2	3.6115	2.8837	2.2625	1.9465
	3	3.8146	3.0709	2.4395	2.1197
	4	3.9642	3.2062	2.5651	2.2412
	5	4.0829	3.3124	2.6624	2.3349
	6	4.1813	3.4000	2.7420	2.4110
	7	4.2655	3.4745	2.8093	2.4752
	8	4.3391	3.5393	2.8676	2.5306
	9	4.4045	3.5968	2.9191	2.5794
	10	4.4634	3.6484	2.9651	2.6230
$\Delta = 0.70$	1	3.2905	2.5758	1.9600	1.6449
	2	3.3203	2.6364	2.0590	1.7676
	3	3.3260	2.6529	2.0917	1.8113
	4	3.3277	2.6592	2.1068	1.8327
	5	3.3282	2.6622	2.1149	1.8449
	6	3.3285	2.6637	2.1198	1.8526
	7	3.3286	2.6645	2.1229	1.8578
	8	3.3286	2.6650	2.1250	1.8615
	9	3.3286	2.6653	2.1265	1.8641
	10	3.3286	2.6655	2.1275	1.8661

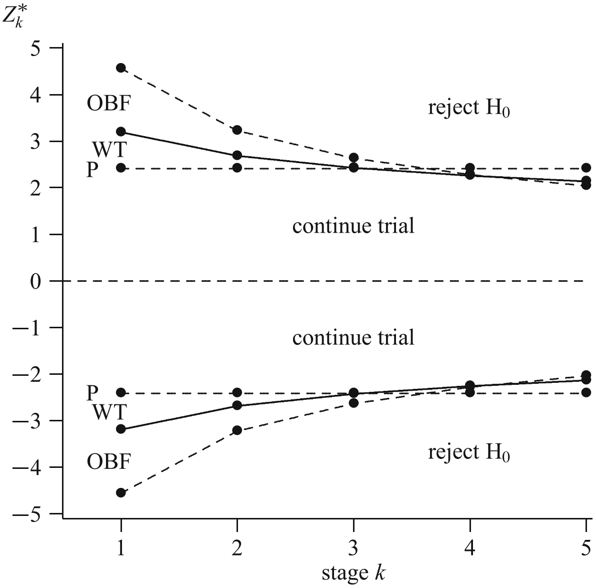

Fig. 2.4 Decision regions of Wang and Tsiatis test (WT) for $\Delta = 0.25$ (*solid line*) as compared to O'Brien and Fleming's (OBF) and Pocock's (P) design (*dashed lines*); $K = 5, \alpha = 0.05$

The sequence of critical values is increasing if $\Delta > 0.50$ whereas it is decreasing for $\Delta < 0.50$. $\Delta < 0$ refers to still stricter criteria for stopping the trial early than the stopping criteria resulting from O'Brien and Fleming's design. For $0 < \Delta < 0.50$, the boundaries have an intermediate shape between O'Brien & Fleming and Pocock type boundaries. For example, if $K = 5$, $\alpha = 0.05$, and $\Delta = 0.25$, $u_k = 3.1941\,k^{-0.25}$, $k = 1, \ldots, K$, and the sequence of critical values is 3.1941, 2.6859, 2.4270, 2.2586, 2.1360. The corresponding decision regions for Z_k^* are illustrated in Fig. 2.4.

The calculation of the maximum sample size and the average sample size under H_1 is straightforward. Table 2.5 supplies the inflation factor $I = I(K, \alpha, \beta, \Delta)$ and the expected reduction in sample size under H_1 relative to n_f, ASN_{H_1}/n_f, of the Wang and Tsiatis design for $\Delta = 0.10, 0.25, 0.40$. This covers practically relevant situations. The table can be used for the sample size calculation as described for the O'Brien & Fleming and Pocock tests (see Table 2.3). It is interesting that the choice of $\Delta = 0.40$ in the cases considered here yields better values for the inflation factor *and* the expected reduction in the sample size. Hence, a Wang and Tsiatis design with $\Delta = 0.40$ is preferable.

Wang and Tsiatis (1987) found that the Δ-class of boundaries is approximately optimal in terms of minimizing the average sample size under H_1. That is, when searching for a design which minimizes ASN_{H_1}, ASN_{H_1} is only slightly higher within the Δ-class as compared to ASN_{H_1} when searching under all possible critical values that define a K-stage group sequential test at level α. The latter

Table 2.5 Inflation factor $I = I(K, \alpha, \beta, \Delta)$ and expected reduction in sample size under H_1, relative to n_f (in parentheses) for the Wang and Tsiatis family of tests, for different values of Δ, K, significance level α, and power $1-\beta$

		$1-\beta = 0.80$		$1-\beta = 0.90$	
	K	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$
$\Delta = 0.10$	1	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
	2	1.005 (0.924)	1.016 (0.882)	1.004 (0.880)	1.014 (0.825)
	3	1.013 (0.872)	1.027 (0.841)	1.012 (0.818)	1.025 (0.777)
	4	1.018 (0.847)	1.035 (0.818)	1.017 (0.786)	1.032 (0.749)
	5	1.022 (0.833)	1.040 (0.804)	1.021 (0.769)	1.037 (0.732)
	6	1.025 (0.824)	1.044 (0.796)	1.024 (0.758)	1.041 (0.721)
	7	1.028 (0.817)	1.047 (0.789)	1.026 (0.750)	1.044 (0.713)
	8	1.030 (0.812)	1.050 (0.785)	1.028 (0.744)	1.046 (0.707)
	9	1.032 (0.809)	1.052 (0.781)	1.030 (0.739)	1.048 (0.703)
	10	1.033 (0.806)	1.054 (0.779)	1.031 (0.736)	1.050 (0.700)
$\Delta = 0.25$	1	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
	2	1.019 (0.891)	1.038 (0.860)	1.017 (0.835)	1.034 (0.795)
	3	1.030 (0.848)	1.054 (0.820)	1.028 (0.784)	1.050 (0.745)
	4	1.039 (0.825)	1.065 (0.799)	1.036 (0.755)	1.059 (0.719)
	5	1.044 (0.812)	1.072 (0.787)	1.041 (0.738)	1.066 (0.704)
	6	1.049 (0.803)	1.077 (0.778)	1.045 (0.727)	1.071 (0.693)
	7	1.052 (0.797)	1.081 (0.772)	1.049 (0.719)	1.075 (0.685)
	8	1.055 (0.792)	1.084 (0.768)	1.051 (0.714)	1.078 (0.680)
	9	1.057 (0.789)	1.087 (0.765)	1.053 (0.709)	1.081 (0.675)
	10	1.059 (0.786)	1.089 (0.762)	1.055 (0.706)	1.083 (0.672)
$\Delta = 0.40$	1	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
	2	1.052 (0.872)	1.075 (0.851)	1.048 (0.805)	1.068 (0.778)
	3	1.076 (0.834)	1.108 (0.812)	1.070 (0.755)	1.099 (0.724)
	4	1.091 (0.815)	1.128 (0.793)	1.084 (0.730)	1.117 (0.699)
	5	1.101 (0.803)	1.142 (0.783)	1.093 (0.714)	1.129 (0.684)
	6	1.108 (0.795)	1.152 (0.776)	1.099 (0.704)	1.138 (0.674)
	7	1.113 (0.790)	1.159 (0.771)	1.105 (0.697)	1.145 (0.667)
	8	1.118 (0.786)	1.165 (0.767)	1.109 (0.691)	1.151 (0.662)
	9	1.122 (0.783)	1.170 (0.765)	1.112 (0.687)	1.155 (0.658)
	10	1.125 (0.780)	1.174 (0.763)	1.115 (0.684)	1.159 (0.655)

designs were already presented by Pocock (1982). He found optimum decision regions with minimum ASN_{H_1} subject to given K, α, and power $1-\beta$ using a (multidimensional) grid search over all possible decision regions. For example, for $\alpha = 0.05$, $1-\beta = 0.90$, and $K = 5$, the sequence of optimum critical values is given by

$$(u_1, u_2, u_3, u_4, u_5) = (2.597, 2.390, 2.390, 2.390, 2.310)$$

with minimizing $\text{ASN}_{H_1}/n_f = 0.681$. For $\alpha = 0.05$, $1 - \beta = 0.50$, and $K = 5$,

$$(u_1, u_2, u_3, u_4, u_5) = (3.671, 2.884, 2.573, 2.375, 2.037)$$

minimizes the average sample size with $\text{ASN}_{H_1}/n_f = 0.929$.

This means, for $1 - \beta = 0.90$, a design with roughly constant boundaries yields the minimum ASN_{H_1} (see Table 2.3), whereas, for $1 - \beta = 0.50$, a design with decreasing critical values is optimum. In other words, for small postulated power an O'Brien and Fleming type design is optimum whereas for higher power a Pocock type design is preferable.

On the other hand, searching for $\Delta = \Delta^*$ which minimizes ASN_{H_1} yields, for $1 - \beta = 0.90$,

$$\Delta^* = 0.445 \text{ and } c_{\text{WT}}(5, 0.05, 0.445) = 2.542,$$
$$(u_1, u_2, u_3, u_4, u_5) = (2.542, 2.446, 2.392, 2.354, 2.325)$$

with minimizing $\text{ASN}_{H_1}/n_f = 0.682$, and, for $1 - \beta = 0.50$,

$$\Delta^* = 0.077 \text{ and } c_{\text{WT}}(5, 0.05, 0.077) = 4.071,$$
$$(u_1, u_2, u_3, u_4, u_5) = (4.071, 3.036, 2.558, 2.265, 2.061)$$

with minimizing $\text{ASN}_{H_1}/n_f = 0.932$. That is, the optimum average sample is virtually the same when considering the one-parameter minimization as compared to the multidimensional search. Hence, the critical values found within the Δ-class of boundaries approximately minimize ASN_{H_1}. This is true for all practically relevant situations.

Table 2.6 contains the optimum Δ^*, the constants $c(K, \alpha, \Delta^*)$, the optimum average sample size (expressed relative to n_f), and the inflation factor $I = I(K, \alpha, \beta, \Delta^*)$ for $\alpha = 0.05$, 0.01, $1 - \beta = 0.50$, 0.80, 0.90. We mention that, again, the optimum Δ^* slightly differ from the figures presented in Table 4 of Wang and Tsiatis (1987). This is partly due to the fact that the average sample size to minimize is fairly flat over the range of values of Δ and, hence, ASN_{H_1} for slightly differing Δ is roughly the same.

As it can be seen from Table 2.6, ASN_{H_1} is minimized for Δ near 0 only if a moderate power ($1 - \beta = 0.50$) is guaranteed. For practically more relevant cases (i.e., $1 - \beta = 0.80$ or $1 - \beta = 0.90$), an optimum design is far from assuming strongly decreasing critical values (as is the case for O'Brien and Fleming's test). We illustrate the use of Table 2.6 for planning a trial by an example. Suppose it is desired to perform a sample size calculation for a four-stage test design at level $\alpha = 0.05$ assuming a standardized effect $|\delta| = 0.30$. To guarantee a power of 80 %, from Table 2.6 one finds that the design with $\Delta^* = 0.366$ minimizes the average sample size under $H_1 : |\delta| = 0.30$. Since $c_{\text{WT}}(4, 0.05, 0.366) = 2.648$, the sequence of critical values for this design is 2.648, 2.413, 2.286, 2.199. The sample size in a

Table 2.6 Optimum Δ^* and constants $c = c(K, \alpha, \Delta^*)$ in the Δ-class of boundaries

K	Δ^*	c	I	Δ^*	c	I	Δ^*	c	I
$\alpha = 0.01$									
	$1 - \beta = 0.50$			$1 - \beta = 0.80$			$1 - \beta = 0.90$		
2	0.258	3.099	1.024 (0.970)	0.450	2.824	1.070 (0.871)	0.511	2.761	1.088 (0.798)
3	0.208	3.634	1.028 (0.952)	0.408	3.053	1.080 (0.834)	0.482	2.904	1.113 (0.750)
4	0.183	4.114	1.032 (0.943)	0.385	3.255	1.083 (0.814)	0.457	3.042	1.117 (0.727)
5	0.167	4.551	1.035 (0.938)	0.374	3.418	1.086 (0.802)	0.441	3.164	1.119 (0.712)
6	0.156	4.949	1.037 (0.934)	0.368	3.552	1.090 (0.794)	0.431	3.267	1.120 (0.703)
7	0.146	5.336	1.038 (0.932)	0.364	3.669	1.092 (0.788)	0.425	3.353	1.122 (0.696)
8	0.140	5.679	1.040 (0.930)	0.361	3.773	1.094 (0.784)	0.420	3.431	1.123 (0.691)
9	0.134	6.017	1.041 (0.929)	0.359	3.865	1.096 (0.781)	0.417	3.497	1.124 (0.687)
10	0.130	6.324	1.042 (0.928)	0.358	3.943	1.098 (0.778)	0.414	3.560	1.125 (0.684)
$\alpha = 0.05$									
	$1 - \beta = 0.50$			$1 - \beta = 0.80$			$1 - \beta = 0.90$		
2	0.176	2.520	1.031 (0.962)	0.417	2.247	1.080 (0.850)	0.485	2.190	1.095 (0.776)
3	0.127	3.064	1.036 (0.946)	0.388	2.460	1.103 (0.812)	0.480	2.316	1.139 (0.721)
4	0.094	3.596	1.040 (0.937)	0.366	2.648	1.110 (0.793)	0.460	2.436	1.153 (0.696)
5	0.076	4.077	1.043 (0.932)	0.353	2.808	1.113 (0.781)	0.444	2.543	1.159 (0.682)
6	0.065	4.515	1.045 (0.929)	0.344	2.948	1.116 (0.773)	0.431	2.642	1.160 (0.673)
7	0.055	4.943	1.047 (0.926)	0.339	3.065	1.119 (0.768)	0.422	2.728	1.162 (0.667)
8	0.048	5.340	1.049 (0.925)	0.335	3.172	1.121 (0.764)	0.416	2.802	1.163 (0.662)
9	0.043	5.712	1.050 (0.924)	0.332	3.268	1.123 (0.761)	0.410	2.874	1.163 (0.658)
10	0.038	6.079	1.051 (0.923)	0.330	3.353	1.124 (0.758)	0.406	2.936	1.164 (0.655)

The inflation factor $I = I(K, \alpha, \beta, \Delta^*)$ can be used for sample size calculations. In parentheses: minimized expected reduction in sample size under H_1, relative to n_f

fixed sample size design is given by

$$n_f = \frac{(1.96 + 0.842)^2}{0.30^2} = 87.2 \; ,$$

and the maximum sample size is

$$N = 1.110 \times 87.2 = 96.8 \; .$$

Hence, a design with $n = 96.8/4 = 24.2$ observations per stage fulfills the requirements. The average sample size under H_1 is

$$\mathrm{ASN}_{H_1} = 0.793 \times 87.2 = 69.1 \; .$$

We note that ASN_{H_1} of Pocock's test is only slightly above this value. From Table 2.3 one finds that the latter is $0.805 \times 87.2 = 70.2$, but the maximum sample size of Pocock's test given by $1.202 \times 87.2 = 104.8$ is considerably larger. That is, the Wang and Tsiatis family can be used to reduce both, the maximum sample size and the average sample size under H_1. Nevertheless, it is not guaranteed that the average sample size is still smallest if some other value in the alternative is true.

Table 2.6 can be used for planning a trial with minimum ASN_{H_1} if a standardized effect δ can be defined which explicitly refers to the minimum clinically relevant different *and* the assumed true state of nature. In other cases, for example, if the sample size calculation was based on the minimum clinically relevant different but the true state of nature is assumed to be larger (in term of δ) the use of Table 2.6 is questionable.

It is also possible to use an alternative optimality criterion. For example, one might find a Δ^* that minimizes

$$N + \mathrm{ASN}_{H_1}$$

or a Δ^* that minimizes

$$\mathrm{ASN}_{H_0} + \mathrm{ASN}_{H_{01}} + \mathrm{ASN}_{H_1} \; ,$$

where $\mathrm{ASN}_{H_{01}}$ denotes the expected sample size calculated midway between H_0 and H_1, i.e., for $\delta/2$. It is clear that these criteria yield designs which are some kind of compromises between the Pocock and the O'Brien and Fleming case.

2.1.4 Other Designs

Even stricter criteria for interim looks were earlier independently proposed by Haybittle (1971) and Peto et al. (1976). They suggested using $u_1 = \cdots = u_{K-1} = 3$

and $u_K = z_{\alpha/2}$. The actual Type I error rate of this procedure exceeds the nominal level α, but the excess is small, and one can simply adopt the approach to adjust the critical value u_K such that the Type I error rate of the procedure is maintained. For example, for $K = 5$ and $\alpha = 0.05$, the adjusted critical level for the Kth stage is given by 1.990 which only slightly exceeds 1.960 and is considerably smaller than the values 2.413 and 2.040 required by the Pocock and the O'Brien and Fleming design, respectively.

In Köpcke (1984, 1989) a "mixed strategy" was proposed which is a trade-off between the designs of O'Brien & Fleming and Pocock. For the first stages of the trial, critical values of the O'Brien and Fleming type and for the latter stages constant critical values were used, and vice versa. These designs also balance the pros and cons of the two competing claims to minimize the expected and the maximum sample sizes, respectively.

2.2 Symmetric Designs

In this section we focus on the two-sided version of a test design described by Pampallona and Tsiatis (1994). They proposed a group sequential test procedure that allows the early rejection of H_0 as well as the early rejection of H_1. The procedure involves a *controlled* acceptance of H_0, which is an essential difference to the procedures described so far. With these designs a symmetric consideration of both error rates will be possible. Early stopping in favor of the alternative (i.e., rejection of H_0) is under control with a specified Type I error probability α, early stopping in favor of the null hypothesis (i.e., controlled acceptance of H_0) is under control with a specified Type II error probability β or a power $1-\beta$ at some specified alternative H_1. Pampallona and Tsiatis (1994) generalized the approach proposed by Emerson and Fleming (1989) who considered the case where the Type I and Type II error probabilities are equal (see also Gould and Pecore 1982).

In the two-sided case, the continuation regions consist of two intervals at each stage k of the trial, $k = 1, \ldots, K-1$. The study is continued if

$$Z_k^* \in (u_k^0;\, u_k^1) \quad \text{or} \quad Z_k^* \in (-u_k^1;\, -u_k^0)\ ,$$

where $0 < u_k^0 < u_k^1$, $k = 1, \ldots, K-1$, are pairs of critical values that define the regions for the two test decisions. One further assumes that by the end of the trial a decision for either H_0 or H_1 is ensured. This is fulfilled by the condition

$$u_K^0 = u_K^1\ . \tag{2.8}$$

With this notation, H_0 is rejected at stage k if $|Z_k^*| \geq u_k^1$, and H_1 is rejected at stage k if $|Z_k^*| \leq u_k^0$, $k = 1, \ldots, K$. Pampallona and Tsiatis (1994) considered critical values u_k^0 and u_k^1 within the Δ-class of critical values according to Wang and Tsiatis (1987). The constants to determine will depend on K, α, Δ, and additionally on the

Type II error probability β. u_k^0 and u_k^1, $k = 1, \ldots, K$, are given by

$$u_k^0 = \vartheta_k - c^0(K, \alpha, \beta, \Delta)\, k^{\Delta - 0.5} \tag{2.9}$$

and

$$u_k^1 = c^1(K, \alpha, \beta, \Delta)\, k^{\Delta - 0.5}\ , \tag{2.10}$$

respectively. Recall that $\vartheta_k = E(Z_k^*) = \delta\sqrt{k\,n}$ and therefore the critical values u_k^0 specify to what extent, under H_1, the test statistic must fall below its expectation in order to reach a decision for H_0. Equivalently, the test statistic $Z_k^* - \vartheta_k$ is standard normal under H_1 and hence (2.9) defines Δ-class critical values for $Z_k^* - \vartheta_k$, $k = 1, \ldots, K$.

From (2.8)–(2.10) one finds that

$$\vartheta_K = (c^0 + c^1)\, K^{\Delta - 0.5}\ ,$$

where $c^0 = c^0(K, \alpha, \beta, \Delta)$ and $c^1 = c^1(K, \alpha, \beta, \Delta)$, and the sample size n per stage, from $\vartheta_K = \delta\sqrt{K\,n}$, is given by

$$n = \frac{(c^0 + c^1)^2\, K^{2(\Delta - 1)}}{\delta^2}\ . \tag{2.11}$$

Since, using this sample size,

$$\vartheta_k = \sqrt{k}(c^0 + c^1)\, K^{\Delta - 1}\ , \tag{2.12}$$

the constants c^0 and c^1 can be calculated, independently of δ, at given K, α, β, and Δ from

$$\sum_{k=1}^{K} P_{H_0}\left(|Z_k^*| \ \geq u_k^1 \cap \bigcap_{\tilde{k}=1}^{k-1} \{|Z_{\tilde{k}}^*| \in (u_{\tilde{k}}^0;\ u_{\tilde{k}}^1)\} \right) = \alpha \quad \text{and}$$

$$\sum_{k=1}^{K} P_{H_1}\left(|Z_k^*| \geq u_k^1 \cap \bigcap_{\tilde{k}=1}^{k-1} \{|Z_{\tilde{k}}^*| \in (u_{\tilde{k}}^0;\ u_{\tilde{k}}^1)\} \right) = 1 - \beta$$

using a bi-dimensional linear search algorithm. It is an important characteristic of the procedure that there might be no u_k^0 such that (2.9) is fulfilled for the first few k. That is, it might happen that it is not possible at all to reach a decision for the acceptance of H_0 at an early stage of the test procedure. This is a reasonable feature of the test procedure since at an early stage the sample size might be too small to reach evidence for H_0.

From (2.11) it becomes clear that the sample size per stage can be calculated, without loss of generality, for $\delta = 1$. The sample size per stage for $\delta \neq 1$ is obtained from dividing n by δ^2. This is along the lines described in the last section. The same is true for the average sample size under H_1. It is obtained from dividing the average sample size calculated for $\delta = 1$ by δ^2. Alternatively, it is possible to supply the maximum and the average sample size under H_1 relative to the sample size, n_f, in a fixed sample size design.

In Pampallona and Tsiatis (1994), c^0 and c^1 were tabulated for $\alpha = 0.01$, 0.05, $\beta = 0.05$, 0.10, 0.20, $\Delta = 0.0, 0.1, \ldots, 0.5$, and $K = 2, \ldots, 5, 10$. The average sample size was calculated under the assumption that H_0 is true, under the assumption that H_1 is true and under the assumption that the intermediate value between H_0 and H_1 is true. In Tables 2.7 and 2.8 we provide the constants c^0 and c^1 together with the value k^* which indicates the first interim analysis where H_0 can be accepted. In order to characterize the procedures in terms of the maximum and the average sample size we tabulate these values relative to n_f. Following §5.2 in Jennison and Turnbull (2000), we also consider a negative value of Δ ($\Delta = -0.25$) and provide the constants and the test characteristics for $\alpha = 0.01$, 0.05, $\Delta = -0.25, 0, 0.25, 0.50$, and $K = 2, \ldots, 5$. Tables 2.7 and 2.8 refer to power $1 - \beta = 0.80$ and $1 - \beta = 0.90$, respectively. Note that the entries for c^0 and c^1 in the corresponding tables in Jennison and Turnbull (2000) are different since they used a different parametrization of the testing problem which is defined in terms of *information levels*. This is not the approach considered here.

We illustrate the use of the tables by an example. Suppose it is desired to use a four-stage design at significance level $\alpha = 0.05$, $1 - \beta = 0.80$, and boundary shape parameter $\Delta = 0$. From Table 2.7, the constants c^0 and c^1 are given by

$$c^0 = 1.9892 \quad \text{and} \quad c^1 = 3.9055\,.$$

Since, from (2.12),

$$\vartheta_k = \sqrt{k}\,(1.9892 + 3.9055)/4 = \sqrt{k}\,1.4738\,,$$

the values u_k^0 are given by

$$\begin{aligned}
u_1^0 &= 1.4738 - 1.9892 < 0\,,\\
u_2^0 &= \sqrt{2}\,1.4738 - 1.9892/\sqrt{2} = 0.678\,,\\
u_3^0 &= \sqrt{3}\,1.4738 - 1.9892/\sqrt{3} = 1.404\,,\\
u_4^0 &= \sqrt{4}\,1.4738 - 1.9892/\sqrt{4} = 1.953\,.
\end{aligned}$$

Table 2.7 Constants $c^0 = c^0(K, \alpha, \beta, \Delta)$, $c^1 = c^1(K, \alpha, \beta, \Delta)$, inflation factor $I = I(K, \alpha, \beta, \Delta)$, and expected reduction in sample size, relative to n_f, for the two-sided Pampallona and Tsiatis family of tests, for different values of Δ, K, significance level α, and power $1 - \beta = 0.80$

		K	c^0	c^1	k^*	I	
$\alpha = 0.01$							
	$\Delta = -0.25$	2	1.5203	4.2926	1	1.023	(0.693, 0.835, 0.974)
		3	2.1201	5.7896	2	1.031	(0.727, 0.831, 0.902)
		4	2.7182	7.1608	2	1.045	(0.641, 0.777, 0.869)
		5	3.2690	8.4578	2	1.053	(0.635, 0.766, 0.850)
	$\Delta = 0.00$	2	1.3473	3.5918	1	1.044	(0.659, 0.807, 0.930)
		3	1.7191	4.3940	1	1.067	(0.654, 0.781, 0.861)
		4	2.0150	5.0819	2	1.078	(0.627, 0.759, 0.831)
		5	2.3113	5.6861	2	1.095	(0.591, 0.733, 0.811)
	$\Delta = 0.25$	2	1.2035	3.0505	1	1.096	(0.648, 0.791, 0.868)
		3	1.4232	3.3971	1	1.149	(0.593, 0.741, 0.813)
		4	1.5752	3.6738	1	1.180	(0.585, 0.728, 0.785)
		5	1.6596	3.9091	1	1.187	(0.578, 0.718, 0.769)
	$\Delta = 0.50$	2	1.0755	2.7242	1	1.236	(0.680, 0.811, 0.849)
		3	1.1974	2.8165	1	1.380	(0.589, 0.750, 0.795)
		4	1.2777	2.8793	1	1.480	(0.555, 0.725, 0.767)
		5	1.3343	2.9258	1	1.554	(0.540, 0.714, 0.751)
$\alpha = 0.05$							
	$\Delta = -0.25$	2	1.5169	3.2561	1	1.026	(0.833, 0.899, 0.942)
		3	2.1087	4.4048	2	1.040	(0.788, 0.858, 0.869)
		4	2.6946	5.4592	2	1.059	(0.747, 0.824, 0.838)
		5	3.2046	6.4679	3	1.066	(0.748, 0.820, 0.822)
	$\Delta = 0.00$	2	1.3402	2.7360	1	1.058	(0.783, 0.860, 0.884)
		3	1.6539	3.3768	1	1.075	(0.785, 0.847, 0.836)
		4	1.9892	3.9055	2	1.107	(0.722, 0.802, 0.802)
		5	2.2730	4.3792	2	1.128	(0.710, 0.788, 0.784)
	$\Delta = 0.25$	2	1.1891	2.3572	1	1.133	(0.759, 0.839, 0.840)
		3	1.3937	2.6439	1	1.199	(0.732, 0.809, 0.790)
		4	1.4980	2.8770	1	1.219	(0.718, 0.793, 0.765)
		5	1.6265	3.0620	2	1.252	(0.688, 0.769, 0.746)
	$\Delta = 0.50$	2	1.0578	2.1190	1	1.286	(0.772, 0.851, 0.839)
		3	1.1748	2.2164	1	1.465	(0.715, 0.807, 0.781)
		4	1.2488	2.2837	1	1.590	(0.695, 0.791, 0.753)
		5	1.2930	2.3348	1	1.677	(0.684, 0.784, 0.737)

k^* denotes the first stage where H_0 can be accepted. In parentheses: expected reduction in sample size under H_0, the value midway between H_0 and H_1, and H_1, respectively

Table 2.8 Constants $c^0 = c^0(K, \alpha, \beta, \Delta)$, $c^1 = c^1(K, \alpha, \beta, \Delta)$, inflation factor $I = I(K, \alpha, \beta, \Delta)$, and expected reduction in sample size, relative to n_f, for the two-sided Pampallona and Tsiatis family of tests, for different values of Δ, K, significance level α, and power $1 - \beta = 0.90$

		K	c^0	c^1	k^*	I
$\alpha = 0.01$						
	$\Delta = -0.25$	2	2.1900	4.3202	1	1.007 (0.798, 0.917, 0.972)
		3	3.0313	5.8345	2	1.017 (0.735, 0.866, 0.874)
		4	3.8158	7.2306	2	1.025 (0.693, 0.835, 0.844)
		5	4.5487	8.5477	3	1.031 (0.691, 0.826, 0.823)
	$\Delta = 0.00$	2	1.8979	3.6211	1	1.024 (0.711, 0.869, 0.909)
		3	2.3595	4.4403	2	1.036 (0.733, 0.851, 0.833)
		4	2.7876	5.1313	2	1.054 (0.650, 0.804, 0.798)
		5	3.1555	5.7471	2	1.065 (0.642, 0.791, 0.779)
	$\Delta = 0.25$	2	1.6750	3.0731	1	1.071 (0.669, 0.836, 0.831)
		3	1.9220	3.4277	1	1.110 (0.647, 0.801, 0.777)
		4	2.0656	3.7126	2	1.122 (0.641, 0.787, 0.746)
		5	2.2377	3.9432	2	1.148 (0.598, 0.760, 0.726)
	$\Delta = 0.50$	2	1.5021	2.7348	1	1.206 (0.681, 0.841, 0.803)
		3	1.6178	2.8294	1	1.329 (0.606, 0.794, 0.743)
		4	1.6924	2.8932	1	1.413 (0.584, 0.777, 0.714)
		5	1.7384	2.9404	1	1.471 (0.576, 0.771, 0.697)
$\alpha = 0.05$						
	$\Delta = -0.25$	2	2.1761	3.2914	1	1.006 (0.956, 0.972, 0.922)
		3	3.0207	4.4532	2	1.023 (0.808, 0.885, 0.835)
		4	3.7815	5.5374	2	1.033 (0.813, 0.874, 0.804)
		5	4.5158	6.5550	3	1.043 (0.772, 0.848, 0.785)
	$\Delta = 0.00$	2	1.8894	2.7686	1	1.032 (0.854, 0.912, 0.848)
		3	2.3430	3.4142	2	1.051 (0.797, 0.867, 0.794)
		4	2.7626	3.9583	2	1.075 (0.758, 0.836, 0.760)
		5	3.0962	4.4490	3	1.084 (0.758, 0.829, 0.741)
	$\Delta = 0.25$	2	1.6621	2.3815	1	1.100 (0.793, 0.871, 0.794)
		3	1.8718	2.6804	1	1.139 (0.785, 0.850, 0.741)
		4	2.0413	2.9108	2	1.167 (0.735, 0.814, 0.711)
		5	2.2089	3.0992	2	1.199 (0.714, 0.797, 0.693)
	$\Delta = 0.50$	2	1.4880	2.1325	1	1.247 (0.780, 0.870, 0.790)
		3	1.5987	2.2338	1	1.398 (0.740, 0.836, 0.723)
		4	1.6581	2.3035	1	1.494 (0.727, 0.826, 0.691)
		5	1.6730	2.3564	1	1.545 (0.714, 0.815, 0.672)

k^* denotes the first stage where H_0 can be accepted. In parentheses: expected reduction in sample size under H_0, the value midway between H_0 and H_1, and H_1, respectively

Therefore, the first facility to reject H_1 is at the second interim analysis (i.e., $k^* = 2$). The values u_k^1 are given by

$$\begin{aligned} u_1^1 &= 3.9055\ , \\ u_2^1 &= 3.9055/\sqrt{2} = 2.762\ , \\ u_3^1 &= 3.9055/\sqrt{3} = 2.255\ , \\ u_4^1 &= 3.9055/\sqrt{4} = 1.953\ . \end{aligned}$$

Analogously, for $K = 4$, $\alpha = 0.05$, $1 - \beta = 0.80$, and boundary shape parameter $\Delta = 0.50$, the critical values are given by

$$(u_1^0, u_2^0, u_3^0, u_4^0) = (0.517, 1.249, 1.810, 2.284) \quad \text{and}$$

$$(u_1^1, u_2^1, u_3^1, u_4^1) = (2.284, 2.284, 2.284, 2.284)\ .$$

The corresponding sample size characteristics are as follows. Suppose the sample size should be calculated for a standardized effect $|\delta| = 0.50$. The sample size in a fixed sample size design is

$$n_f = \frac{(1.96 + 0.842)^2}{0.50^2} = 31.4\ .$$

If one wants to use a Pampallona and Tsiatis design with $\Delta = 0$, the maximum sample size is (see Table 2.7)

$$N = 1.107 \times 31.4 = 34.8\ ,$$

and hence one needs $34.8/4 = 8.7$ observations per stage. Note that this value can also directly be achieved by (2.11). Using the design with $\Delta = 0.5$, the maximum sample size is

$$N = 1.590 \times 31.4 = 49.9\ ,$$

and hence $49.9/4 = 12.5$ observations per stage are necessary to achieve 80 % power. The decision regions at the calculated cumulative sample sizes for these designs are illustrated in Fig. 2.5.

If the alternative is true, the average sample size for the design with $\Delta = 0$ is

$$\text{ASN}_{H_1} = 0.802 \times 31.4 = 25.2\ ,$$

for $\Delta = 0.5$ it is

$$\text{ASN}_{H_1} = 0.753 \times 31.4 = 23.6\ .$$

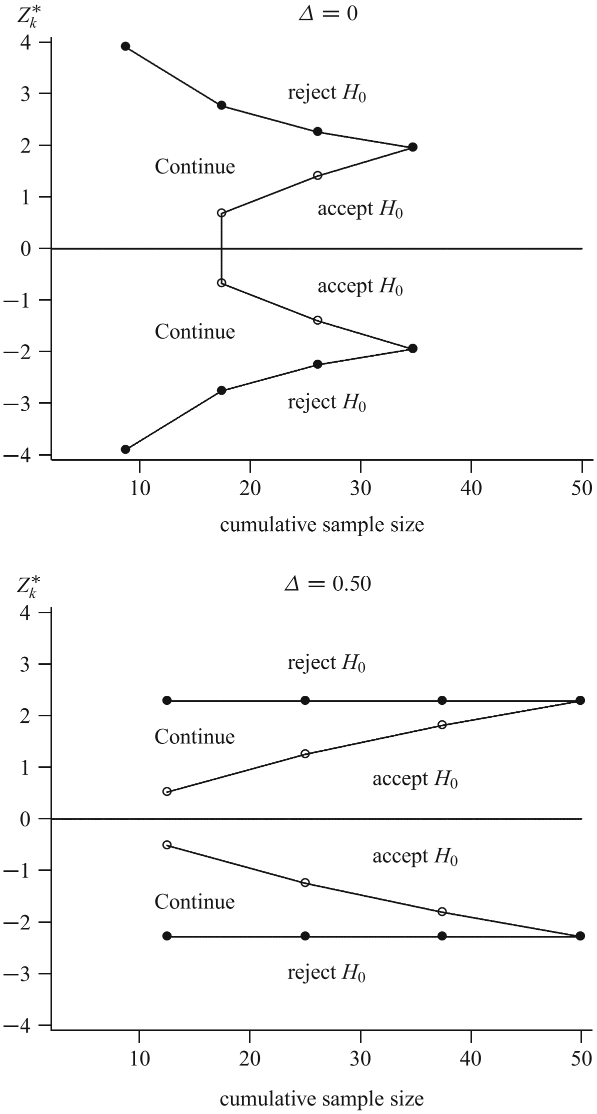

Fig. 2.5 Continuation and decision regions for the two-sided design of Pampallona and Tsiatis; $K = 4$, $\alpha = 0.05$, $1 - \beta = 0.80$, $\delta = 0.5$, $\Delta = 0$ (*upper graph*) and $\Delta = 0.50$ (*lower graph)*

Due to the early stopping in favor of H_0, the average sample size under H_0 is considerably reduced, too. The reduction is 72.2 % and 69.5 % of n_f for $\Delta = 0$ and $\Delta = 0.5$, respectively. For $\Delta = 0.5$, however, the maximum sample is much higher than the sample size in a fixed sample size design, which might prevent someone from using a design with $\Delta = 0.5$.

Compared to the critical values for the designs of O'Brien and Fleming (1979) and Pocock (1977) the critical values for rejecting H_0 of the Pampallona and Tsiatis design are somewhat smaller (see Table 2.2). Note that, for $\varDelta = 0$, the last stage critical value can be even smaller than the critical value in a fixed sample size design. The reason is that a rejection at this stage requires evidence against H_0 also for the earlier stages in the sense that the absolute value of the test statistic may not fall short of the lower boundary. This is a general feature of sequential testing and will be discussed further on in Sect. 2.3. For increasing power, a smaller critical value is required for the acceptance of H_0. As a consequence, roughly speaking, the Wang and Tsiatis critical values for rejecting H_0 result from considering the case where the power approaches 100 %.

Some extensions of the procedure are obvious. First, it is possible to find an optimal $\varDelta^*$ that minimizes the average sample size under, say, H_1. Wassmer (1999c) provides the minimizing $\varDelta^*$ together with the maximum sample size. This is in accordance to the optimum regions found in Wang and Tsiatis (1987). Under reasonable assumptions, a design that minimizes ASN_{H_1} is for values of $\varDelta$ between 0.4 and 0.5. These optimum designs with respect to the average sample size under H_1 implicate a large increase in the necessary maximum sample size. It is therefore questionable if these designs are appealing for practical use.

It is also possible to consider different shapes for rejecting H_0 and accepting H_0. In the simplest case, two values, $\varDelta_0$ and $\varDelta_1$, are specified which determine the rejection regions for H_0 and H_1, respectively. In this case, the critical values are given by

$$u_k^0 = \vartheta_k - c^0(K, \alpha, \beta, \varDelta_0, \varDelta_1)\, k^{\varDelta_0 - 0.5}$$

and

$$u_k^1 = c^1(K, \alpha, \beta, \varDelta_0, \varDelta_1)\, k^{\varDelta_1 - 0.5} \ ,$$

respectively, and it is straightforward to calculate the two constants such that the resulting test procedure preserves the significance level and the required power.

2.3 One-Sided Designs

The issues involved in the question of whether a test should be two-sided or one-sided are getting more complex when considering group sequential designs (O'Brien 1998). Nevertheless, it is conceptually straightforward to define one-sided tests for this case. Thereby, one-sided testing means, without loss of generality, testing $H_0 : \mu \le \mu_0$ against the one-sided alternative

$$H_1 : \mu > \mu_0 \ .$$

We note that in all cases considered here this is equivalent to testing the null hypothesis $H_0 : \mu = \mu_0$ which is tested against the one-sided alternative H_1, i.e., all relevant calculations are performed on the boundary of $H_0 : \mu \leq \mu_0$.

In the simplest case, the continuation (and acceptance) regions $\mathscr{C}_k^*$ for Z_k^* are given by

$$\mathscr{C}_k^* = (-\infty; u_k), \ k = 1, \ldots, K,$$

and the rejection regions $\mathscr{R}_k^*$ are given by the complement of $\mathscr{C}_k^*$, i.e.,

$$\mathscr{R}_k^* = [u_k; \infty), \ k = 1, \ldots, K.$$

In this case, a set of critical values $u_1, \ldots, u_K$ satisfying

$$P_{H_0}(Z_1^* \geq u_1 \text{ or } \ldots \text{ or } Z_K^* \geq u_K) = \alpha \tag{2.13}$$

defines a one-sided group sequential test design. The critical values can be found, as for the two-sided case, within the $\varDelta$-class of critical values due to Wang and Tsiatis (1987) with critical values given by

$$u_k = c_{\mathrm{WT}}(K, \alpha, \varDelta)\, k^{\varDelta - 0.5}, \ k = 1, \ldots, K.$$

The test characteristics in terms of the inflation factor and the average sample size can be computed along the lines described in Sect. 2.1. The (exact) sample size formula for the fixed sample size design is

$$n_f = \frac{\left(\varPhi^{-1}(1-\alpha) + \varPhi^{-1}(1-\beta)\right)^2}{\delta^2}\,.$$

The sample size and the expected sample size for the group sequential design are found numerically.

It turns out that the critical values and the test characteristics of the one-sided test at level $2\,\alpha$ are the same as those already displayed in Tables 2.1, 2.3, 2.4, and 2.5 for the two-sided test at level α. This is in analogy to the fixed sample size case. Note, however, that it is a only numerical coincidence which is accurate at the given decimal places for $K \leq 10$. To understand this, consider the case $K = 2$. If the critical values were exactly the same, the two probabilities

$$P_1 = 1 - P_{H_0}(|Z_1^*| < u, \ |Z_2^*| < u) \quad \text{and}$$
$$P_2 = 2(1 - P_{H_0}(Z_1^* < u, \ Z_2^* < u))$$

must coincide at given u. Numerically, this is indeed the case, but only if u is sufficiently large. Below are the values for P_1 and P_2 for some u which were calculated with the use of bivariate standard normal cdf with correlation $1/\sqrt{2}$:

u	0.4	0.8	1.2	1.6	2.0	2.4
P_1	0.8699	0.5978	0.34704	0.174531	0.0759732	0.0285025575
P_2	0.9189	0.6021	0.34720	0.174533	0.0759732	0.0285025575

The numerical identity for $u > 1.6$ is due to the fact that under the bivariate normal distribution with positive correlation the probability of "significantly opposed effects" is extremely small if the critical value u is sufficiently large. That is, for sufficiently large u, the probability

$$P_{H_0}(Z_1^* < -u,\, Z_2^* > u)\ ,$$

which is taken into account in the two-sided case, but not for the one-sided case, is negligibly small. The difference in terms of the corresponding power (and the average sample size) is even smaller and, for commonly used values of K, α, and β, negligible too. Our calculations showed that for up to $K = 10$ the maximum difference in the critical values is smaller than 0.00005 and therefore of no practical concern (see Proschan 1999; Wassmer 1999c). Hence the figures in Tables 2.1, 2.3, 2.4, and 2.5 with $2\,\alpha$ can be used for a one-sided test design at significance level α. Nevertheless, particularly if it is desired to use a plan with many stages, the decision regions and the test characteristics should be specifically calculated for the one-sided case in order to achieve precise results.

DeMets and Ware (1980, 1982) considered the issue of stopping the trial for futility, which becomes evident in the one-sided setting. If, for example, the one-sided p-value at some stage k is greater than 0.50, the effect is directed opposed to the alternative. Hence, there might be no reasonable chance to obtain a significant result at the end of the trial. In the planning phase of the trial it can be decided that in this case the trial will be stopped. DeMets and Ware considered various choices of stopping for futility options, including the "asymmetric method" which involves a constant boundary for stopping for futility. Using this method, the continuation regions $\mathscr{C}_k^*$ are given by

$$\mathscr{C}_k^* = (u^L;\, u_k),\ k = 1, \ldots, K-1,$$

i.e., the trial is stopped for futility if Z_k^* falls short of a constant u^L. The rejection regions $\mathscr{R}_k^*$ are, as above, given by

$$\mathscr{R}_k^* = [u_k;\, \infty),\ k = 1, \ldots, K.$$

Taking into account this stopping for futility option, the critical values are different as compared to the original design defined through (2.13). Indeed, the critical values are somewhat smaller and, most importantly, if the null hypothesis is true, the average sample size reduces considerably. This is shown in Table 2.9 which extends Tables 1 and 2 in DeMets and Ware (1980). In Table 2.9, the critical values are

Table 2.9 Constants $c = c_{\mathrm{WT}}(K, \alpha, \Delta)$ in the one-sided test design due to DeMets and Ware with stopping in favor of H_0 if $Z_k^* < u^L$

	K	$u^L = 0.5$	$u^L = 0$	$u^L = -0.5$	$u^L = -\infty$
$\alpha = 0.005$					
O'Brien/ Fleming	2	3.6397 (1.31)	3.6469 (1.50)	3.6480 (1.69)	3.6481 (2.00)
	3	4.4478 (1.51)	4.4823 (1.87)	4.4921 (2.27)	4.4945 (3.00)
	4	5.1177 (1.66)	5.1867 (2.19)	5.2103 (2.79)	5.2182 (4.00)
	5	5.6974 (1.78)	5.8046 (2.46)	5.8451 (3.27)	5.8611 (5.00)
	6	6.2125 (1.89)	6.3601 (2.70)	6.4194 (3.72)	6.4455 (6.00)
	7	6.6785 (1.98)	6.8679 (2.93)	6.9471 (4.14)	6.9849 (7.00)
	8	7.1057 (2.06)	7.3375 (3.14)	7.4376 (4.54)	7.4884 (8.00)
	9	7.5012 (2.14)	7.7760 (3.33)	7.8976 (4.93)	7.9623 (8.99)
	10	7.8702 (2.21)	8.1884 (3.52)	8.3318 (5.30)	8.4113 (9.99)
Pocock	2	2.7698 (1.31)	2.7715 (1.50)	2.7718 (1.69)	2.7718 (2.00)
	3	2.8646 (1.50)	2.8710 (1.87)	2.8726 (2.27)	2.8730 (2.99)
	4	2.9226 (1.65)	2.9341 (2.18)	2.9377 (2.78)	2.9387 (3.99)
	5	2.9625 (1.77)	2.9787 (2.45)	2.9844 (3.26)	2.9863 (4.99)
	6	2.9919 (1.87)	3.0124 (2.69)	3.0201 (3.70)	3.0231 (5.99)
	7	3.0146 (1.97)	3.0391 (2.91)	3.0487 (4.12)	3.0528 (6.98)
	8	3.0327 (2.05)	3.0608 (3.12)	3.0724 (4.53)	3.0775 (7.98)
	9	3.0476 (2.12)	3.0789 (3.31)	3.0923 (4.91)	3.0986 (8.98)
	10	3.0600 (2.19)	3.0943 (3.50)	3.1094 (5.28)	3.1169 (9.97)
$\alpha = 0.025$					
O'Brien/ Fleming	2	2.7615 (1.31)	2.7897 (1.50)	2.7956 (1.69)	2.7965 (2.00)
	3	3.3566 (1.50)	3.4370 (1.87)	3.4631 (2.27)	3.4711 (2.99)
	4	3.8345 (1.64)	3.9763 (2.17)	4.0283 (2.78)	4.0486 (3.99)
	5	4.2365 (1.76)	4.4442 (2.44)	4.5256 (3.25)	4.5617 (4.98)
	6	4.5845 (1.86)	4.8609 (2.68)	4.9736 (3.69)	5.0283 (5.98)
	7	4.8914 (1.94)	5.2384 (2.90)	5.3838 (4.11)	5.4590 (6.97)
	8	5.1660 (2.02)	5.5849 (3.10)	5.7638 (4.51)	5.8611 (7.97)
	9	5.4142 (2.08)	5.9061 (3.29)	6.1190 (4.89)	6.2395 (8.96)
	10	5.6404 (2.14)	6.2061 (3.47)	6.4535 (5.26)	6.5981 (9.95)
Pocock	2	2.1683 (1.29)	2.1765 (1.49)	2.1781 (1.68)	2.1783 (1.99)
	3	2.2639 (1.48)	2.2826 (1.84)	2.2880 (2.24)	2.2895 (2.97)
	4	2.3203 (1.61)	2.3484 (2.14)	2.3580 (2.75)	2.3613 (3.95)
	5	2.3580 (1.72)	2.3942 (2.40)	2.4078 (3.21)	2.4132 (4.94)
	6	2.3851 (1.81)	2.4285 (2.63)	2.4457 (3.64)	2.4532 (5.92)
	7	2.4056 (1.89)	2.4552 (2.84)	2.4759 (4.05)	2.4855 (6.91)
	8	2.4217 (1.95)	2.4769 (3.03)	2.5006 (4.43)	2.5123 (7.89)
	9	2.4347 (2.01)	2.4948 (3.21)	2.5215 (4.81)	2.5352 (8.87)
	10	2.4455 (2.07)	2.5100 (3.38)	2.5393 (5.16)	2.5550 (9.85)

O'Brien and Fleming: $\Delta = 0$; Pocock: $\Delta = 0.5$. In parentheses: expected number of performed stages under the assumption that H_0 is true

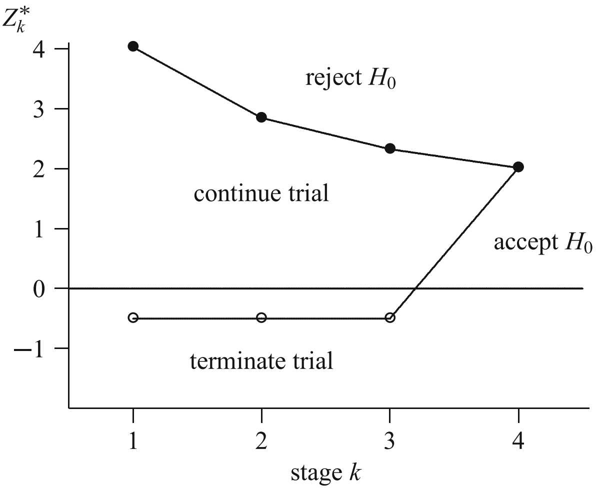

Fig. 2.6 Continuation and decision regions for the one-sided design of DeMets and Ware with O'Brien and Fleming type boundaries; $K = 4$, $\alpha = 0.025$, $u^L = -0.5$

supplied for $\alpha = 0.005, 0.025$, and stopping for futility bounds $u^L = 0.5, 0, -0.5, -\infty$. Note that $u^L = 0.5$, 0, and -0.5 are related to the cases where the trial is terminated if the p-value at some stage exceeds 0.3085, 0.50, and 0.6915, respectively. $u^L = -\infty$ refers to the case where no early stopping in favor of H_0 is taken into account, and is added for the sake of completeness although the constants c are the same (at the given decimal places) as those provided in Table 2.1. Table 2.9 also contains the expected number of stages under the assumption that H_0 is true. Note that this quantity does not depend on the sample size n per stage and hence on the power of the test. In Fig. 2.6 the decision regions are illustrated for an O'Brien and Fleming type design when considering early stopping for futility if $Z_k^* < -0.5$.

Table 2.9 shows that the critical values are quite insensitive to changes in the stopping for futility bound u^L. Only for $u^L = 0.5$ the upper critical values are considerably smaller than for $u^L = -\infty$, and hence it is more likely to reject H_0 at some stage k. Therefore, the "profit" using critical values which are adjusted for the stopping for futility option is quite small. Furthermore, positive values of u^L are not recommended for practical applications since it is likely to terminate the trial for futility also if the alternative is true. On the other hand, the expected number of performed stages is sensitive to the choice of u^L. This effect even increases for increasing K.

The considered strategy *enforces* to stop the trial if the test statistic falls short of u^L. In this sense, the futility bounds are *binding*. One has to bear in mind, however, that the termination of the trial (i.e., uncontrolled acceptance of H_0) is always possible under control of the significance level α. Hence one can use the boundaries that do not take into account stopping for futility and stop the trial if it is unlikely to reject H_0 at the subsequent stages of the trial. If one performs futility stops in this way the corresponding futility bounds are *non-binding*. From a practical perspective, the non-binding bounds have advantages because, in practice, often a weak study result should not force the study to be stopped. It is even true that an early rejection should not force the study to be stopped. For example, although a formal rejection is already reached, in a clinical trial more safety issues may be of interest and the trial should then proceed. This is possible with the rejection boundaries that do not account for the futility stopping because the Type I error rate is not influenced by this. Consequently, in the non-binding futility case both rejection and futility bounds are treated the same. Note that clearly the Type II error is not under control with such a flexible futility rule.

A further disadvantage of the asymmetric method is that it generally needs a higher maximum number of observations under H_1 at given power $1 - \beta$ when considering various choices of u^L. Although there is no such strong effect for $u^L \leq -0.5$, it is more pronounced for larger u^L. This is illustrated in Table 2.10. It supplies the inflation factor $I = I(K, \alpha, \beta)$ and the average sample size under H_1 for O'Brien & Fleming's and Pocock's design relative to the sample size in a fixed sample size design for power $1 - \beta = 0.80$. Again, the case $u^L = -\infty$ is added although these figures were already provided in Table 2.3. The table shows that there is a distinct increase in the maximum sample size which is also increasing in K. Although there are cases where the average sample size under H_1 is even smaller than for the one-sided test design without provision for early stopping in favor of H_0 (particularly, for Pocock type boundaries) the necessary increase in the maximum sample size can be large. Note that this of course is true for both the bonding and the non-binding case because also for the non-binding case the power will be calculated for the considered rejection and futility bounds and the former are somewhat larger under the non-binding case.

DeMets and Ware also proposed the "constant likelihood group sequential method" which uses stopping for futility bounds which were motivated by the work of Wald (1947). These bounds explicitly depend on the power $1-\beta$ of the trial. They compared the test characteristics of the two methods which were applied for Pocock type test designs (DeMets and Ware 1980) as well as for O'Brien and Fleming type test designs (DeMets and Ware 1982). This method, however, also does not allow for the acceptance of H_0 that controls the Type II error rate.

On the other hand, Pampallona and Tsiatis (1994) proposed the symmetric approach also for the one-sided test design. As already described in Sect. 2.2, this procedure controls the acceptance of H_0 by a prespecified Type II error rate β. The lower and upper critical values are defined within the Δ-class of critical values, i.e., as for the two-sided case, they are given by

$$u_k^0 = \vartheta_k - c^0(K, \alpha, \beta, \Delta)\, k^{\Delta - 0.5}$$

and

$$u_k^1 = c^1(K, \alpha, \beta, \Delta)\, k^{\Delta - 0.5} \ ,$$

Table 2.10 Inflation factor $I = I(K, \alpha, \beta)$ and expected reduction in sample size under H_1, relative to n_f, for the one-sided test design due to DeMets and Ware with stopping in favor of H_0 if $Z_k^* < u^L$

	K	$u^L = 0.5$	$u^L = 0$	$u^L = -0.5$	$u^L = -\infty$
$\alpha = 0.005$					
O’Brien/ Fleming	2	1.007 (0.936)	1.002 (0.943)	1.002 (0.946)	1.001 (0.947)
	3	1.049 (0.865)	1.016 (0.876)	1.008 (0.883)	1.007 (0.886)
	4	1.115 (0.837)	1.039 (0.844)	1.017 (0.855)	1.011 (0.862)
	5	1.197 (0.826)	1.068 (0.825)	1.028 (0.837)	1.015 (0.847)
	6	1.292 (0.825)	1.101 (0.813)	1.040 (0.824)	1.017 (0.838)
	7	1.400 (0.833)	1.136 (0.804)	1.052 (0.815)	1.019 (0.831)
	8	1.519 (0.847)	1.174 (0.799)	1.065 (0.808)	1.021 (0.826)
	9	1.652 (0.866)	1.214 (0.797)	1.079 (0.802)	1.022 (0.822)
	10	1.795 (0.889)	1.255 (0.797)	1.093 (0.797)	1.024 (0.819)
Pocock	2	1.096 (0.862)	1.092 (0.869)	1.092 (0.871)	1.092 (0.872)
	3	1.173 (0.815)	1.145 (0.831)	1.138 (0.838)	1.137 (0.841)
	4	1.255 (0.785)	1.190 (0.808)	1.171 (0.821)	1.166 (0.828)
	5	1.344 (0.764)	1.233 (0.792)	1.198 (0.811)	1.187 (0.822)
	6	1.439 (0.749)	1.276 (0.779)	1.222 (0.803)	1.203 (0.818)
	7	1.541 (0.738)	1.318 (0.769)	1.244 (0.796)	1.216 (0.817)
	8	1.650 (0.729)	1.361 (0.760)	1.265 (0.791)	1.226 (0.816)
	9	1.766 (0.723)	1.405 (0.752)	1.286 (0.786)	1.236 (0.815)
	10	1.892 (0.719)	1.449 (0.745)	1.305 (0.782)	1.243 (0.816)
$\alpha = 0.025$					
O’Brien/ Fleming	2	1.035 (0.884)	1.012 (0.893)	1.008 (0.899)	1.008 (0.902)
	3	1.143 (0.838)	1.050 (0.838)	1.024 (0.848)	1.017 (0.856)
	4	1.286 (0.825)	1.099 (0.809)	1.043 (0.819)	1.024 (0.831)
	5	1.457 (0.832)	1.157 (0.794)	1.063 (0.801)	1.028 (0.818)
	6	1.653 (0.852)	1.220 (0.787)	1.085 (0.790)	1.032 (0.809)
	7	1.869 (0.879)	1.287 (0.785)	1.108 (0.781)	1.035 (0.802)
	8	2.103 (0.909)	1.359 (0.787)	1.131 (0.775)	1.037 (0.798)
	9	2.347 (0.940)	1.435 (0.793)	1.154 (0.771)	1.038 (0.794)
	10	2.599 (0.971)	1.516 (0.800)	1.178 (0.768)	1.040 (0.791)
Pocock	2	1.133 (0.833)	1.114 (0.844)	1.111 (0.850)	1.110 (0.853)
	3	1.273 (0.780)	1.193 (0.797)	1.172 (0.811)	1.166 (0.819)
	4	1.427 (0.752)	1.267 (0.771)	1.218 (0.791)	1.202 (0.805)
	5	1.594 (0.735)	1.340 (0.754)	1.258 (0.779)	1.229 (0.799)
	6	1.775 (0.726)	1.413 (0.741)	1.295 (0.770)	1.249 (0.796)
	7	1.970 (0.721)	1.487 (0.731)	1.330 (0.762)	1.265 (0.795)
	8	2.179 (0.719)	1.562 (0.723)	1.363 (0.757)	1.279 (0.795)
	9	2.401 (0.719)	1.638 (0.717)	1.395 (0.752)	1.291 (0.795)
	10	2.634 (0.720)	1.716 (0.712)	1.427 (0.747)	1.301 (0.795)

Power $1 - \beta = 0.80$. In parentheses: expected reduction in sample size under H_1

respectively. At stage k, the trial is continued if

$$Z_k^* \in (u_k^0;\ u_k^1)\ ,$$

where $u_k^0 < u_k^1$, $k = 1, \ldots, K-1$. Note that the condition $u_k^0 > 0$ that was postulated for the two-sided case is not met. Consequently, requiring

$$u_K^0 = u_K^1$$

as for the two-sided case, at each stage k of the test procedure a decision in favor of H_0 is possible, and one yields a decision for either H_0 or H_1, at the latest, at stage K. In analogy to the two-sided case, the constants c^0 and c^1 can be calculated, independently of the specified standardized effect size δ, at given α, β, K, and $\varDelta$ from (see Sect. 2.2)

$$\sum_{k=1}^{K} P_{H_0}\left(Z_k^* \ \geq u_k^1 \cap \bigcap_{\tilde{k}=1}^{k-1} \{Z_{\tilde{k}}^* \in (u_{\tilde{k}}^0;\ u_{\tilde{k}}^1)\} \right) = \alpha \quad \text{and}$$

$$\sum_{k=1}^{K} P_{H_1}\left(Z_k^* \geq u_k^1 \cap \bigcap_{\tilde{k}=1}^{k-1} \{Z_{\tilde{k}}^* \in (u_{\tilde{k}}^0;\ u_{\tilde{k}}^1)\} \right) = 1 - \beta\ .$$

In Pampallona and Tsiatis (1994), c^0 and c^1 were tabulated for one-sided $\alpha = 0.01$, 0.05, $\beta = 0.05$, 0.10, 0.20, $\varDelta = 0.0, 0.1, \ldots, 0.5$, and $K = 2, \ldots, 5, 10$, with the average sample size calculated under the assumption that H_0 is true, under the assumption that H_1 is true and under the assumption that the intermediate value between H_0 and H_1 is true. In Tables 2.11 and 2.12 we provide the constants c^0 and c^1 for the one-sided significance levels $\alpha = 0.005$ and $\alpha = 0.025$ since these are the one-sided levels which are commonly used in practice. Following §4.2 in Jennison and Turnbull (2000) we also consider a negative value of $\varDelta$ ($\varDelta = -0.25$) and provide the constants and the test characteristics for $\varDelta = -0.25$, 0, 0.25, 0.50, and $K = 2, \ldots, 5$. The constants c^0 and c^1 differ from the constants provided in the tables in Jennison and Turnbull (2000) since they use a different parametrization. As for the two-sided case, Tables 2.11 and 2.12 refer to power $1 - \beta = 0.80$ and $1 - \beta = 0.90$, respectively.

We illustrate the use of the tables by an example. Suppose it is desired to use a one-sided four-stage design at significance level $\alpha = 0.025$, $1 - \beta = 0.80$, and boundary shape parameter $\varDelta = 0$. From Table 2.11, the constants c^0 and c^1 are given by

$$c^0 = 2.0191 \quad \text{and} \quad c^1 = 3.8989\ .$$

Table 2.11 Constants $c^0 = c^0(K, \alpha, \beta, \Delta)$, $c^1 = c^1(K, \alpha, \beta, \Delta)$, inflation factor $I = I(K, \alpha, \beta, \Delta)$, and expected reduction in sample size, relative to n_f, for the one-sided Pampallona and Tsiatis family of tests, for different values of Δ, K, significance level α, and power $1 - \beta = 0.80$

		K	c^0	c^1	I	
$\alpha = 0.005$						
	$\Delta = -0.25$	2	1.5203	4.2925	1.023	(0.602, 0.827, 0.974)
		3	2.1396	5.7843	1.035	(0.555, 0.789, 0.898)
		4	2.7214	7.1601	1.045	(0.539, 0.763, 0.868)
		5	3.2718	8.4571	1.054	(0.519, 0.746, 0.848)
	$\Delta = 0.00$	2	1.3473	3.5918	1.044	(0.590, 0.802, 0.930)
		3	1.7249	4.3924	1.068	(0.506, 0.750, 0.858)
		4	2.0431	5.0765	1.085	(0.478, 0.725, 0.827)
		5	2.3248	5.6838	1.098	(0.464, 0.710, 0.809)
	$\Delta = 0.25$	2	1.2035	3.0505	1.096	(0.598, 0.788, 0.868)
		3	1.4238	3.3970	1.149	(0.488, 0.725, 0.812)
		4	1.5879	3.6716	1.184	(0.442, 0.696, 0.782)
		5	1.7208	3.9007	1.210	(0.418, 0.678, 0.763)
	$\Delta = 0.50$	2	1.0755	2.7242	1.236	(0.649, 0.810, 0.849)
		3	1.1975	2.8165	1.380	(0.525, 0.745, 0.795)
		4	1.2781	2.8792	1.480	(0.463, 0.713, 0.767)
		5	1.3373	2.9255	1.556	(0.426, 0.694, 0.749)
$\alpha = 0.025$						
	$\Delta = -0.25$	2	1.5191	3.2548	1.027	(0.674, 0.865, 0.939)
		3	2.1287	4.3986	1.045	(0.633, 0.822, 0.865)
		4	2.6993	5.4578	1.060	(0.611, 0.798, 0.836)
		5	3.2388	6.4579	1.071	(0.592, 0.781, 0.818)
	$\Delta = 0.00$	2	1.3409	2.7356	1.059	(0.656, 0.838, 0.882)
		3	1.7086	3.3626	1.092	(0.586, 0.789, 0.828)
		4	2.0191	3.8989	1.116	(0.560, 0.765, 0.797)
		5	2.2944	4.3744	1.133	(0.544, 0.750, 0.780)
	$\Delta = 0.25$	2	1.1893	2.3571	1.133	(0.663, 0.826, 0.839)
		3	1.4038	2.6416	1.204	(0.569, 0.769, 0.786)
		4	1.5644	2.8652	1.250	(0.528, 0.741, 0.757)
		5	1.6945	3.0513	1.283	(0.506, 0.724, 0.739)
	$\Delta = 0.50$	2	1.0578	2.1190	1.286	(0.707, 0.845, 0.839)
		3	1.1756	2.2162	1.466	(0.602, 0.788, 0.780)
		4	1.2548	2.2830	1.595	(0.548, 0.759, 0.750)
		5	1.3133	2.3326	1.694	(0.515, 0.743, 0.731)

In parentheses: expected reduction in sample size under H_0, the value midway between H_0 and H_1, and H_1, respectively

Table 2.12 Constants $c^0 = c^0(K, \alpha, \beta, \Delta)$, $c^1 = c^1(K, \alpha, \beta, \Delta)$, inflation factor $I = I(K, \alpha, \beta, \Delta)$, and expected reduction in sample size, relative to n_f, for the one-sided Pampallona and Tsiatis family of tests, for different values of Δ, K, significance level α, and power $1 - \beta = 0.90$

		K	c^0	c^1	I	
$\alpha = 0.005$						
	$\Delta = -0.25$	2	2.1903	4.3201	1.007	(0.651, 0.903, 0.971)
		3	3.0334	5.8340	1.017	(0.634, 0.856, 0.874)
		4	3.8160	7.2306	1.025	(0.595, 0.827, 0.844)
		5	4.5549	8.5462	1.032	(0.570, 0.810, 0.822)
	$\Delta = 0.00$	2	1.8980	3.6211	1.024	(0.611, 0.863, 0.909)
		3	2.3813	4.4360	1.041	(0.557, 0.817, 0.832)
		4	2.7938	5.1303	1.055	(0.534, 0.791, 0.797)
		5	3.1593	5.7464	1.066	(0.516, 0.775, 0.778)
	$\Delta = 0.25$	2	1.6750	3.0731	1.071	(0.602, 0.833, 0.831)
		3	1.9251	3.4272	1.112	(0.509, 0.781, 0.776)
		4	2.1153	3.7069	1.139	(0.473, 0.755, 0.745)
		5	2.2710	3.9399	1.159	(0.454, 0.739, 0.726)
	$\Delta = 0.50$	2	1.5021	2.7348	1.206	(0.642, 0.841, 0.803)
		3	1.6179	2.8294	1.329	(0.525, 0.788, 0.743)
		4	1.6942	2.8930	1.414	(0.469, 0.763, 0.713)
		5	1.7500	2.9398	1.478	(0.437, 0.749, 0.695)
$\alpha = 0.025$						
	$\Delta = -0.25$	2	2.1892	3.2859	1.009	(0.734, 0.925, 0.919)
		3	3.0228	4.4526	1.024	(0.701, 0.874, 0.834)
		4	3.7944	5.5333	1.035	(0.662, 0.846, 0.802)
		5	4.5225	6.5533	1.044	(0.641, 0.830, 0.784)
	$\Delta = 0.00$	2	1.8920	2.7674	1.033	(0.686, 0.884, 0.847)
		3	2.3657	3.4094	1.058	(0.637, 0.839, 0.793)
		4	2.7704	3.9568	1.077	(0.611, 0.814, 0.759)
		5	3.1294	4.4422	1.091	(0.593, 0.798, 0.740)
	$\Delta = 0.25$	2	1.6626	2.3813	1.100	(0.672, 0.857, 0.794)
		3	1.9072	2.6744	1.153	(0.593, 0.809, 0.739)
		4	2.0939	2.9040	1.189	(0.559, 0.784, 0.710)
		5	2.2469	3.0947	1.214	(0.540, 0.768, 0.692)
	$\Delta = 0.50$	2	1.4880	2.1325	1.248	(0.702, 0.864, 0.790)
		3	1.6009	2.2336	1.399	(0.604, 0.817, 0.723)
		4	1.6763	2.3018	1.506	(0.556, 0.794, 0.689)
		5	1.7317	2.3521	1.587	(0.528, 0.781, 0.670)

In parentheses: expected reduction in sample size under H_0, the value midway between H_0 and H_1, and H_1, respectively

Since, from (2.12),

$$\vartheta_k = \sqrt{k}\,(2.0191 + 3.8989)/4 = \sqrt{k}\,1.4795\ ,$$

the values u_k^0 are given by

$$\begin{aligned} u_1^0 &= 1.4795 - 2.0191 = -0.540\ , \\ u_2^0 &= \sqrt{2}\,1.4795 - 2.0191/\sqrt{2} = 0.665\ , \\ u_3^0 &= \sqrt{3}\,1.4795 - 2.0191/\sqrt{3} = 1.397\ , \\ u_4^0 &= \sqrt{4}\,1.4795 - 2.0191/\sqrt{4} = 1.949\ . \end{aligned}$$

The values u_k^1 are given by

$$\begin{aligned} u_1^1 &= 3.899\ , \\ u_2^1 &= 3.8989/\sqrt{2} = 2.757\ , \\ u_3^1 &= 3.8989/\sqrt{3} = 2.251\ , \\ u_4^1 &= 3.8989/\sqrt{4} = 1.949\ . \end{aligned}$$

Analogously, for $K = 4$, $\alpha = 0.025$, $1 - \beta = 0.80$ and boundary shape parameter $\Delta = 0.50$, the critical values are given by

$$(u_1^0, u_2^0, u_3^0, u_4^0) = (0.514, 1.247, 1.809, 2.283) \quad \text{and}$$

$$(u_1^1, u_2^1, u_3^1, u_4^1) = (2.283, 2.283, 2.283, 2.283)\ .$$

We recognize that the critical values (except u_1^0 for $\Delta = 0$) are very similar to the two-sided case. Nevertheless, the difference can be more pronounced for other choices of the design parameters. Hence it is important to decide if the design should be one-sided or two-sided, and the calculation must take this into account.

The corresponding sample size characteristics are similar but not equal to the two-sided case, too. Suppose the sample size is to be calculated for a standardized effect $|\delta| = 0.50$. The sample size in a fixed sample size design at one-sided level $\alpha = 0.025$ is then same as that for two-sided $\alpha = 0.05$, namely (see Sect. 2.2)

$$n_f = 31.4\ .$$

If one wants to use the one-sided Pampallona and Tsiatis design with $\Delta = 0$, the maximum sample size is

$$N = 1.116 \times 31.4 = 35.0 \ ,$$

which is slightly above the corresponding value for the two-sided case. Similarly, using the design with $\Delta = 0.5$, the maximum sample size is

$$N = 1.595 \times 31.4 = 50.1 \ .$$

The decision regions for these one-sided designs are illustrated in Fig. 2.7.

Due to the stopping criterion in favor of the acceptance of H_0 the critical values are somewhat smaller than for the pure one-sided case. The "profit" of using such a method is generally larger as compared to the *asymmetric method* of DeMets and Ware (1980, 1982). This is due to the different decision region which dictates the acceptance of H_0 and the termination of the study (see Fig. 2.6). Pampallona and

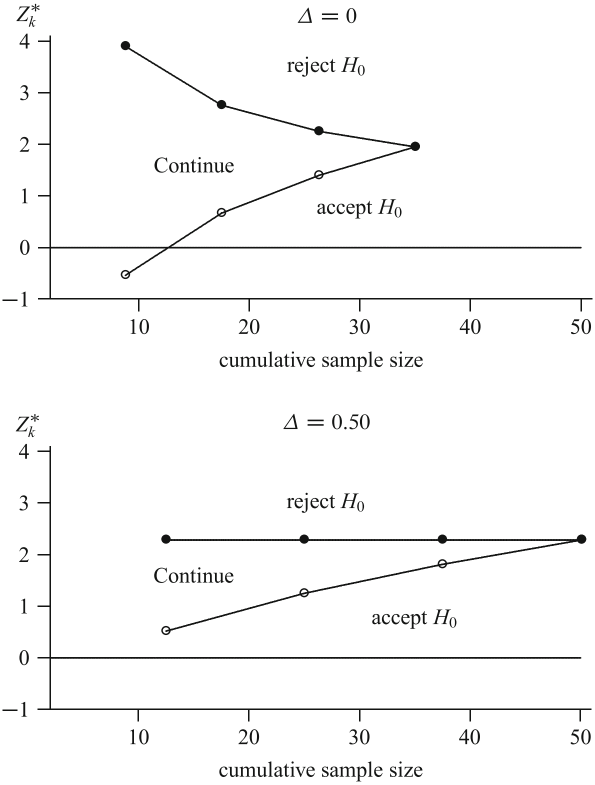

Fig. 2.7 Continuation and decision regions for the one-sided design of Pampallona and Tsiatis; $K = 4$, $\alpha = 0.025$, $1 - \beta = 0.80$, $\delta = 0.5$, $\Delta = 0$ (*upper graph*) and $\Delta = 0.50$ (*lower graph*)

Tsiatis also compared their approach with the *constant likelihood method* which also depends on the power of the test. Concerning the average sample size, however, the approach considered here is favorable and thus recommended.

A fully symmetric approach is obtained in the one-sided case for $\alpha = \beta$. In this case, $c^0 = c^1$, and the decision regions for rejecting H_0 when using the test statistic Z_k^* at stage k and the decision regions for rejecting H_1 when using the test statistic $Z_k^* - \vartheta_k$ at stage k are the same. Recall that $Z_k^* - \vartheta_k$ is standard normal under H_1 and hence H_1 is rejected if a standard normal variable falls short of $-c^0 k^{\Delta-0.5}$, $k = 1, \ldots, K$. This procedure exactly coincides with the one-sided test proposed by Emerson and Fleming (1989).

As for the two-sided case, it is straightforward to search for optimum designs within the Δ-class of critical values, or to specify two different shapes for rejecting and accepting H_0. Optimum designs for the one-sided case were also found by Jennison (1987) and Eales and Jennison (1992) who extended optimization to the average over several values of the parameter space and to a Bayesian optimality criterion (see also, Barber and Jennison 2002; Anderson 2007).

2.4 A Note on Two-Sided Designs

In single stage designs a two-sided test at level α can usually be understood as two one-sided tests at level $\alpha/2$. This is also possible for a group sequential design, however, it is not as straightforward as in the single stage case.

A two-sided group sequential design without a (binding) futility boundary, for instance, has to be understood as two one-sided designs with binding futility boundaries. This follows from the observation that rejection of $H_0 : \mu = \mu_0$ at stage k with $Z_k^* < -u_k$ cannot be understood as rejection of $H_0^{(+)} : \mu \leq \mu_0$, since this null hypothesis is completely in line with the data. It has to be understood as rejection of $H_0^{(-)} : \mu \geq \mu_0$ and acceptance of $H_0^{(+)}$. Similarly, $Z_k^* \geq u_k$ implies rejection of $H_0^{(+)}$ and a futility stop for $H_0^{(-)}$. Note that the Type I error rate condition (1.11) implies that the one-sided futility boundaries are binding. The two one-sided tests of a two-stage design with binding futility boundary are more obvious. In this case $\mathscr{C}_k^{*,+} = (u_k^0; u_k^1)$ is the continuation region of the one-sided test for $H_0^{(+)} : \mu \leq \mu_0$ and $\mathscr{C}_k^{*,-} = (-u_k^1; -u_k^0)$ the continuation region for $H_0^{(-)} : \mu \geq \mu_0$.

Note that understanding a two-sided test as two one-sided tests implies that the approximate power formula (2.1), which is the power of a one-sided test, is the more correct formula than the traditional two-sided one. The latter counts rejections in the "wrong direction" as success even though they are obviously erroneous, since the actually true one-sided null hypothesis has been rejected. The same is true for group sequential designs. As mentioned, the numerical differences between the two- and one-sided power formula are usually negligible, however, the one-sided version is not only more correct but also more simple and hence preferable. Finally, only the two one-sided tests permit decisions on the direction of the treatment effect and hence are essential for all clinical trials.

Chapter 3
Procedures with Unequally Sized Stages

The assumption of equally sized stages is quite restrictive. For example, although it might be planned to perform the interim analyses after stages of equal size, the schedule of interim analyses can often at best ensure that the stage sizes are roughly the same. Furthermore, the accomplishment of an interim analysis requires the definition of the population to be analyzed and the actual size of this population will be rarely exactly equal to the planned one. Thus, due to practical constraints the sample size will slightly vary between the stages. Pocock (1977, 1982) showed that the critical values which are designed for equally sized stages can also be used for unequal stages sizes if the assumption of equal sample sizes is not grossly violated. The effect on the size and the power of the test procedure is small in this case and might be neglected for practical purposes. Nevertheless, there are cases where the influence is not negligible and it is imperative to find procedures especially designed to account for this effect. It might also be desired from the outset to plan the interim analyses at some prespecified number of observations without the restriction of equally sized stages. For example, one might wish to drop the first interim analysis in an O'Brien and Fleming design to increase the probability for a positive test result in the first stage. The critical values will then differ from the original O'Brien and Fleming critical values. Furthermore, it might be more attractive to perform the interim analyses at given time points of analyses rather than after the observation of a specified number of subjects. This will typically lead to unequally sized stages as well.

In the following we will present group sequential test procedures that are specifically designed for unequally sized stages. We adopt the basic notation

G. Wassmer, W. Brannath, *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*, Springer Series in Pharmaceutical Statistics,
DOI 10.1007/978-3-319-32562-0_3

introduced in Sect. 1.2. That is, at stage k of the group sequential test the statistic

$$Z_k^* = \frac{\sum_{\tilde{k}=1}^{k} \sqrt{n_{\tilde{k}}} Z_{\tilde{k}}}{\sqrt{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}}} \ , \ k = 1, \ldots, K,$$

summarizes the information obtained in the data up to stage k. The expectation and the elements of the correlation matrix are given by (1.5) and (1.6), respectively. Since

$$Z_k^* = \frac{\sum_{\tilde{k}=1}^{k} \sqrt{t_{\tilde{k}} - t_{\tilde{k}-1}}\, Z_{\tilde{k}}}{\sqrt{t_k}} \ , \ k = 1, \ldots, K, \qquad (3.1)$$

where $t_k = \sum_{\tilde{k}=1}^{k} n_{\tilde{k}}/N$, $k = 1, \ldots, K$, and $t_0 = 0$, the statistic Z_k^* depends on the sample sizes per stage only through t_k, $k = 1, \ldots, K$. Thus, it is possible to characterize a group sequential design in terms of the *information rates* t_k. These values express how much information, relative to the maximum sample size, N, is obtained up to stage k. In our prototype case, the information rates are expressed in terms of the sample sizes but this is not necessarily the case (see Chap. 5). Note that, alternatively, one might express the information rates in terms of the parameters

$$\tau_k = \frac{t_k - t_{k-1}}{t_1} \ , \ k = 2, \ldots, K,$$

which were introduced in Sect. 1.4 and denote the standardized time interval between the $(k-1)$th and kth stage relative to n_1. In terms of the sample sizes, this is simply n_k/n_1, $k = 2, \ldots, K$.

We first describe the effect of using the decision boundaries designed for equally sized stages to the more general case and briefly describe a worst case scenario adjustment procedure. We then sketch the use of designs with prefixed sample sizes that need not to be equal to each other. A more general approach is provided by the use of the α-*spending function* or *use function approach*. This more sophisticated approach can handle unpredictable sample sizes per stage and we will see that even the maximum number of stages, K, needs not be fixed in advance when using this approach. The extension to the β-*spending function* approach is also briefly discussed.

3.1 Effect of Using Boundaries for Equally Sized Stages

As already mentioned, Pocock (1977, 1982) suggested the use of the critical values calculated for equal sample sizes also for the more general case where departures from this assumption may occur. That is, the test statistic Z_k^* given by (1.2) is calculated from the observations recorded at each stage and the test is conducted

Table 3.1 Type I error rate (size) and power for using the two-sided O'Brien & Fleming's and Pocock's boundaries, respectively, for unequally sized stages

	n_1, n_2, n_3, n_4	Size	Power
O'Brien and Fleming	$20, 20, 20, 20$	0.050	0.800
	$18, 18, 18, 26$	0.052	0.801
	$16, 16, 16, 32$	0.053	0.801
	$22, 22, 22, 14$	0.048	0.799
	$24, 24, 22, 10$	0.047	0.797
	$10, 10, 10, 50$	0.057	0.801
	$40, 20, 10, 10$	0.046	0.796
Pocock	$20, 20, 20, 20$	0.050	0.800
	$18, 18, 18, 26$	0.051	0.798
	$16, 16, 16, 32$	0.052	0.795
	$22, 22, 22, 14$	0.049	0.802
	$24, 24, 22, 10$	0.047	0.803
	$10, 10, 10, 50$	0.055	0.787
	$40, 20, 10, 10$	0.041	0.804

The standardized effect δ is chosen such that the test has power $1 - \beta = 0.80$ for $N = 80$

with critical regions as if the stage sizes were equal. In fact, it turns out that this simple way has a negligible influence on the Type I error rate and the power of the test as long as there are only slight departures from the intended sample sizes.

In Table 3.1, the size and the power are shown for this strategy using the original two-sided O'Brien & Fleming's and Pocock's boundaries, respectively, for $K = 4$ and $\alpha = 0.05$. These are given by $(u_1, u_2, u_3, u_4) = (4.049, 2.863, 2, 337, 2.024)$ and $(u_1, u_2, u_3, u_4) = (2.361, 2.361, 2.361, 2.361)$, respectively (see Table 2.2). The standardized effect δ is chosen such that the test has power $1 - \beta = 0.80$ for $N = 80$ and equal sample sizes between the stages, i.e., $n_1 = n_2 = n_3 = n_4 = 20$. For O'Brien and Fleming's test, $\delta = 0.317$, and, for Pocock's test, $\delta = 0.344$. The values in the table are obtained by numerical integration where the recursive integration formula was applied to the general case of arbitrarily sized stages as shown in Sect. 1.4. Different sample size allocations are presented ranging from slight departures to halving and doubling the preplanned sample sizes, if possible. For all considered cases, the maximum sample size, N, is set equal to its design value.

By definition, for $n_1 = n_2 = n_3 = n_4 = 20$ the size is equal to 0.05 and the power is equal to 0.80. The effects on the Type I error rate and the power are small for the sequences of sample sizes considered in this example. The effects on the Type I error rate are similar for O'Brien and Fleming's and for Pocock's test, but for both tests they are very small. This is true for many practically relevant situations and one can conclude that the use of the methods that assume equally sized stage for the more general case of unequal spacings will be satisfactory. Clearly, from a theoretical point of view this approximate consideration is unsatisfactory.

From a more theoretical point of view, Proschan et al. (1992) examined the maximum possible increase in Type I error rate considering all possible sequences of information rates. In the one-sided case,

$$1 - P\left(\bigcap_{k=1}^{K}\{Z_k^* < u_k\}\right) \le 1 - \prod_{k=1}^{K} P(Z_k^* < u_k)\ , \tag{3.2}$$

which follows from Slepian's inequality (Slepian 1962) for multivariate normally distributed observations with mean vector **0** and pairwise positive correlation. Hence, the right-hand side in (3.2) is an upper bound for the Type I error rate when using the critical values $u_1, \ldots, u_K$ in a group sequential test design with arbitrary information rates. Similarly, from Šidák's inequality (Sidák 1967) it follows for the two-sided case that

$$1 - P\left(\bigcap_{k=1}^{K}\{|Z_k^*| < u_k\}\right) \le 1 - \prod_{k=1}^{K} P(|Z_k^*| < u_k)\ , \tag{3.3}$$

since Šidák's inequality is valid for multivariate normal random vectors with mean vector **0** and any variance–covariance matrix.

The upper bounds defined in (3.2) and (3.3) are least upper bounds, which can be shown as follows. In terms of the information rates the correlation matrix reads as

$$\mathrm{Cov}(Z_k^*, Z_{k'}^*) = \sqrt{\frac{t_k}{t_{k'}}}\ , k < k'.$$

Setting $t_k = \epsilon^{K-k}$ and letting $\epsilon \to 0$ (i.e., considering the border case $t_k = 0$ for $k < K$ and $t_K = 1$), it follows that

$$\lim_{\epsilon \to 0} \mathrm{Cov}(Z_k^*, Z_{k'}^*) = \lim_{\epsilon \to 0} \sqrt{\frac{\epsilon^{K-k}}{\epsilon^{K-k'}}} = \lim_{\epsilon \to 0} \sqrt{\epsilon^{k'-k}} = 0$$

for $k < k'$. Hence, the distribution of $(Z_1^*, \ldots, Z_K^*)$ converges to the distribution of K independent standard normal variates as $\epsilon \to 0$. Therefore,

$$\sup_{0<t_1<t_2<\ldots<t_K=1} 1 - P_{H_0}\left(\bigcap_{k=1}^{K}\{Z_k^* < u_k\}\right) = 1 - \prod_{k=1}^{K} \Phi(u_k)$$

and

$$\sup_{0<t_1<t_2<\ldots<t_K=1} 1 - P_{H_0}\left(\bigcap_{k=1}^{K}\{|Z_k^*| < u_k\}\right) = 1 - \prod_{k=1}^{K} (2\Phi(u_k) - 1)\ .$$

For example, if one uses the unadjusted bounds in the two-sided case for $\alpha = 0.05$, then

$$\lim_{K\to\infty} \sup_{0<t_1<\ldots<t_K=1} 1 - P_{H_0}\left(\bigcap_{k=1}^{K}\{|Z_k^*| < 1.96\}\right) = \lim_{K\to\infty} 1 - (1-\alpha)^K = 1\ .$$

That is, the actual Type I error rate is arbitrarily close to 1 if the number of stages increases. Proschan et al. (1992) showed that this result is also true when using the adjusted bounds of O'Brien and Fleming or Pocock. If there are many looks close to the information time 0, the Type I error rate can considerably exceed α. This effect can be reduced if one agrees not to perform an interim analysis before some minimum information time.

By fixing K, a simple method to achieve valid bounds for arbitrarily sized stages is given by the calculation of critical values u_k, $k = 1, \ldots, K$, that fulfill

$$1 - \prod_{k=1}^{K} \Phi(u_k) = \alpha \tag{3.4}$$

in the one-sided case and

$$1 - \prod_{k=1}^{K} (2\Phi(u_k) - 1) = \alpha \tag{3.5}$$

in the two-sided case. These critical values are termed *worst case scenario adjusted critical values* as they anticipate the maximum possible increase in the Type I error rate which can result from using different stage sample sizes. Wassmer (1999a) provided tables for different boundary shapes and examined the characteristics of the resulting test procedures. For example, for O'Brien & Fleming and Pocock type boundaries, the critical values are defined through

$$u_k = \tilde{c}_{\text{OBF}}(K,\alpha)/\sqrt{k}\ , k = 1, \ldots, K,$$

and

$$u_k = \tilde{c}_{\text{P}}(K,\alpha)\ , k = 1, \ldots, K,$$

respectively, where the constants $\tilde{c}_{\text{OBF}}(K,\alpha)$ and $\tilde{c}_{\text{P}}(K,\alpha)$ are determined such that (3.4) or (3.5) is fulfilled. Some constants $\tilde{c}_{\text{OBF}}(K,\alpha)$ and $\tilde{c}_{\text{P}}(K,\alpha)$ are supplied in Table 3.2 for different K and α. For comparison, the original constants designed for equal stage sizes are also provided in the table.

Table 3.2 Constants $\tilde{c}_{\mathrm{OBF}}(K,\alpha)$ and $\tilde{c}_{\mathrm{P}}(K,\alpha)$ for the one-sided and the two-sided testing problem assuming the independence case

Two-sided test					
	K	$\alpha = 0.001$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
O'Brien and Fleming	2	4.655 (4.654)	3.655 (3.648)	2.828 (2.797)	2.427 (2.373)
	3	5.725 (5.710)	4.545 (4.494)	3.580 (3.471)	3.111 (2.961)
	4	6.655 (6.609)	5.326 (5.218)	4.247 (4.049)	3.724 (3.466)
	5	7.493 (7.412)	6.035 (5.861)	4.856 (4.562)	4.286 (3.915)
	10	10.91 (10.57)	8.956 (8.411)	7.394 (6.598)	6.643 (5.696)
Pocock	2	3.481 (3.464)	2.806 (2.772)	2.236 (2.178)	1.949 (1.875)
	3	3.588 (3.555)	2.934 (2.873)	2.388 (2.289)	2.114 (1.992)
	4	3.662 (3.614)	3.022 (2.939)	2.491 (2.361)	2.226 (2.067)
	5	3.719 (3.657)	3.089 (2.987)	2.569 (2.413)	2.311 (2.122)
	10	3.890 (3.774)	3.289 (3.117)	2.800 (2.555)	2.560 (2.270)
One-sided test					
	K	$\alpha = 0.001$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
O'Brien and Fleming	2	4.373 (4.371)	3.315 (3.300)	2.431 (2.373)	1.997 (1.899)
	3	5.389 (5.367)	4.148 (4.077)	3.118 (2.961)	2.613 (2.391)
	4	6.276 (6.217)	4.882 (4.740)	3.735 (3.466)	3.172 (2.814)
	5	7.076 (6.971)	5.549 (5.330)	4.299 (3.915)	3.688 (3.191)
	10	10.35 (9.961)	8.312 (7.670)	6.665 (5.696)	5.865 (4.686)
Pocock	2	3.290 (3.269)	2.575 (2.531)	1.955 (1.875)	1.632 (1.527)
	3	3.403 (3.363)	2.712 (2.636)	2.121 (1.992)	1.818 (1.650)
	4	3.481 (3.424)	2.806 (2.704)	2.234 (2.067)	1.943 (1.730)
	5	3.540 (3.468)	2.877 (2.754)	2.319 (2.122)	2.036 (1.787)
	10	3.719 (3.591)	3.089 (2.889)	2.568 (2.270)	2.309 (1.943)

In parentheses: critical values assuming equally sized stages

The worst case adjusted critical values do not differ very much from the original bounds. Only for large K ($K \geq 5$) there is a substantial effect and hence the use of the worst case adjusted critical values yields considerably conservative test procedures when the test is applied for "usual" scenarios. This property is a little more pronounced for O'Brien and Fleming's test as compared to Pocock's test. To illustrate this, consider a five-stage design at two-sided significance level $\alpha = 0.05$. From Table 3.2 the adjusted nominal significance levels α_k and $\tilde{\alpha}_k$ from the original and the worst case adjusted test procedure, respectively, are given by

$$\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0.0158\,,$$
$$\tilde{\alpha}_1 = \tilde{\alpha}_2 = \tilde{\alpha}_3 = \tilde{\alpha}_4 = \tilde{\alpha}_5 = 0.0102$$

for Pocock's test and

$$(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5) = (0.000005, 0.0013, 0.0084, 0.0226, 0.0413)\,,$$
$$(\tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3, \tilde{\alpha}_4, \tilde{\alpha}_5) = (0.000001, 0.0006, 0.0051, 0.0152, 0.0299)$$

for O'Brien and Fleming's test. Hence the latter test is a bit more sensitive with respect to this kind of adjustment. However, in Wassmer (1999a) it was shown that the worst case adjusted O'Brien & Fleming and Pocock tests perform nearly identically with respect to the loss of power that is due to the adjustment of the critical values.

The method that is based on the independence case provides a conservative solution to the problem of unequally sized stages in group sequential testing. Nevertheless, it yields a valid level α test procedure. The most important application is given if the interim analyses are performed at specific but arbitrary time points rather than after a specified number of observations. For example, the study protocol might contain a schedule of inspection times for the interim analyses during the course of the trial. In this case, the group sizes are unpredictable and the test procedure that is based on the worst case adjusted critical values can be used. An alternative, more sophisticated procedure that is based on the specification of an α-spending function will be presented in Sect. 3.3.

3.2 Sample Sizes Fixed in Advance

Generally, group sequential plans can be derived for any sequence of sample sizes $n_1, \ldots, n_K$. As already mentioned the test statistic depends on the sample size only through the information rates $t_1, \ldots, t_K$. That is, by specifying a vector $\boldsymbol{V} = (t_1, \ldots, t_K)$ of preplanned information rates it is possible to define decision regions $\mathscr{C}_k^*$ and $\mathscr{R}_k^*$ for the statistic (3.1) such that the Type I error rate of the test is α.

As for equally sized stages, it is possible to define decision regions within the $\varDelta$-class of Wang and Tsiatis (1987). Two different modes of calculating the continuation regions for unequally planned stage sizes can be defined. The first mode is that the critical values within the $\varDelta$-class of boundaries be the same for unequally and equally sized stages. That is, the critical values are given by

$$u_k = c(K, \alpha, \varDelta, \boldsymbol{V}) k^{\varDelta - 0.5}\ ,\ k = 1, \ldots, K, \tag{3.6}$$

where $c = c(K, \alpha, \varDelta, \boldsymbol{V})$ is a constant resulting from the chosen design that depends on the maximum number K of stages to be performed, the significance level α, the shape parameter $\varDelta$, and on the vector $\boldsymbol{V}$. In the second mode, the critical values u_k are calculated through

$$u_k = c(K, \alpha, \varDelta, \boldsymbol{V}) \left(\frac{t_k}{t_1}\right)^{\varDelta - 0.5}\ ,\ k = 1, \ldots, K, \tag{3.7}$$

where the constant c is defined analogously. Note that the first version which is based on critical values defined by (3.6) yields critical values that are "stage-wisely" similar to the original ones but account for the specific choice of the information rate

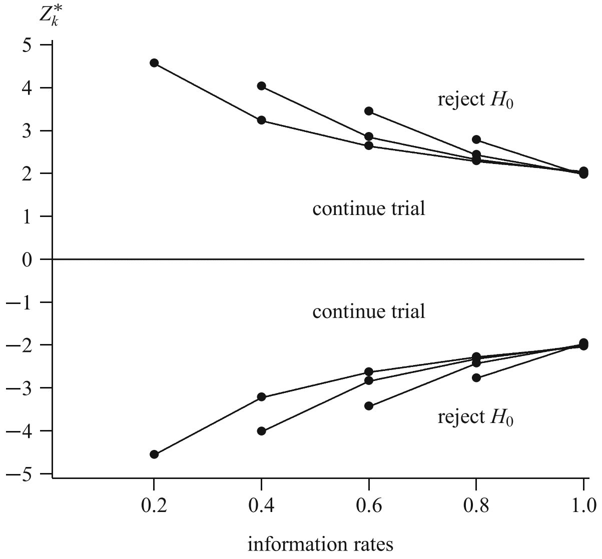

Fig. 3.1 Decision regions for O'Brien and Fleming's design ($\varDelta = 0$) for fixed, unequally spaced information rates using critical values defined by (3.6) and (3.7). Critical values defined by (3.7) are graphically indistinguishable from the five-stage design critical values (see text); $\alpha = 0.05$, two-sided

vector $\boldsymbol{V}$. The second version, on the other hand, provides critical values that are similar to the original ones at the given information rates. For $\varDelta = 0.5$, of course, both methods coincide. The difference of the two modes of specifying the critical values is illustrated in Fig. 3.1 for a two-sided design with critical values according to O'Brien and Fleming, i.e., $\varDelta = 0$. The decision regions are displayed for four different information rate vectors $\boldsymbol{V}$, namely $\boldsymbol{V} = (0.20, 0.40, 0.60, 0.80, 1)$ (equally sized stages), $\boldsymbol{V} = (0.40, 0.60, 0.80, 1)$, $\boldsymbol{V} = (0.60, 0.80, 1)$, and $\boldsymbol{V} = (0.80, 1)$.

An important characteristic of the O'Brien and Fleming design with critical values defined by (3.7) is that they are nearly identical to the respective stage-wise critical values of the five-stage design with equally sized stages. The critical values $(u_1, \ldots, u_K)$ are given by

$$\begin{gathered}(4.562, 3.226, 2.634, 2.281, 2.040) \text{ for } \boldsymbol{V} = (0.20, 0.40, 0.60, 0.80, 1)\ , \\ (3.226, 2.634, 2.281, 2.040) \text{ for } \boldsymbol{V} = (0.40, 0.60, 0.80, 1)\ , \\ (2.631, 2.278, 2.038) \text{ for } \boldsymbol{V} = (0.60, 0.80, 1)\ , \\ (2.260, 2.021) \text{ for } \boldsymbol{V} = (0.80, 1)\ ,\end{gathered}$$

and hence the critical values of the five-stage design are nearly the same for the different $\boldsymbol{V}$ (omitting the first interim analysis even yields the same critical values up to the given decimal places). The other way round: Inserting an interim analysis has nearly no effect if one uses the critical values according to (3.7). This is particularly true for conservative bounds for the first few stages, which is the case for O'Brien and Fleming's design. Note that this determination of critical values requires the stage sizes to be fixed in advance. We will see in the next section that the determination of critical values with unequally sized stages can be generalized obtaining valid bounds for arbitrarily sized stages and a flexible number of interim analyses.

Given $K, \alpha, \Delta, \boldsymbol{V}$, power $1-\beta$, and standardized effect $\delta = (\mu-\mu_0)/\sigma$, one finds, as for equally sized stages, the sample sizes $n_1, \ldots, n_K$ to achieve a prespecified power $1-\beta$. The calculation can be performed as follows. For given decision regions the power does only depend on the shift parameter $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_K)$ where $\vartheta_k = \delta\sqrt{t_k N}$, $k = 1, \ldots, K$ (see (1.14) in Sect. 1.3). Hence, it is possible to determine the shift value ϑ^* such that the power with $\vartheta_k = \vartheta^* \sqrt{t_k/t_1}$, $k = 1, \ldots, K$, equals $1-\beta$. Note that, with this parametrization, the shift ϑ^* in equally sized designs is obtained by setting $t_k = k/K$. The maximum sample size, N, is therefore given by

$$N = N(K, \alpha, \beta, \Delta, \boldsymbol{V}) = \frac{\vartheta^{*2}}{t_1 \delta^2} \ ,$$

the vector of accumulated sample sizes is

$$(n_1, n_1 + n_2, \ldots, N) = \boldsymbol{V} N \ ,$$

and the stage-wise sample sizes, n_k, $k = 1, \ldots, K$, can be easily obtained. Particularly, it suffices to provide the shift for a specific δ, for example, $\delta = 1$. The same is true for the average sample size under H_1, ASN_{H_1}, or under a different assumption for δ (for example, $\delta = 0$). This is completely analogous to the group sequential design with equally sized stages, and the fixed sample size design.

As for equally sized stages, N and ASN_{H_1} are inversely proportional in δ^2. Therefore, the inflation factor $I = I(K, \alpha, \beta, \Delta, \boldsymbol{V}) = N/n_f$ relates the maximum sample size to its corresponding fixed sample size. The expected reduction in sample size relative to n_f, ASN_{H_1}/n_f, is defined accordingly and can be calculated also under different values of δ.

Apparently, the information rates of a group sequential test can be optimized with respect to a chosen criterion. That is, one might search for the optimum vector $\boldsymbol{V}^* = (t_1^*, \ldots, t_K^*)$ where $t_K^* = 1$ such that the test has, say, minimum ASN_{H_1}. Alternatively, the test can be optimal with respect to another optimization criterion Q, for example, $Q = \text{ASN}_{H_1}/n_f + I$. Generally, this involves a multidimensional minimization routine. Brittain and Bailey (1993) and Müller and Schäfer (1999) found optimized information rates for different designs where even the critical values are optimized. In Table 3.3 we present optimized design parameters of two-sided tests for up to a maximum of $K = 4$ stages, where we found optimum information rates $t_1^*, \ldots, t_K^*$ and optimum Δ^* for decision regions within the Δ-class of boundaries defined by (3.7). This is reasonable since the Δ-class yields approximately optimum designs (see Sect. 2.1). The results of the table are obtained by a K-dimensional minimization routine that minimizes ASN_{H_1}, given K, α, and

Table 3.3 Optimum Δ^* and information rates $t_1^*, \ldots, t_K^*$ with constants $c = c(K, \alpha, \Delta^*, V^*)$ and minimum ASN_{H_1}, relative to n_f, in the Δ-class of boundaries for the two-sided testing problem

	K	$t_1^*, \ldots, t_K^*$	Δ^*	c	I	ASN_{H_1}/n_f
$\alpha = 0.01$						
$1-\beta = 0.50$						
	2	0.71, 1	0.12 (0.26)	2.983	1.018 (1.024)	0.955 (0.970)
	3	0.60, 0.78, 1	0.11 (0.21)	3.223	1.025 (1.028)	0.941 (0.952)
	4	0.54, 0.68, 0.82, 1	0.10 (0.18)	3.396	1.030 (1.032)	0.935 (0.943)
$1-\beta = 0.80$						
	2	0.59, 1	0.41 (0.45)	2.833	1.054 (1.070)	0.862 (0.871)
	3	0.47, 0.71, 1	0.39 (0.41)	2.988	1.072 (1.080)	0.822 (0.834)
	4	0.40, 0.58, 0.77, 1	0.38 (0.39)	3.102	1.082 (1.083)	0.804 (0.814)
$1-\beta = 0.90$						
	2	0.53, 1	0.49 (0.51)	2.777	1.077 (1.088)	0.796 (0.798)
	3	0.41, 0.66, 1	0.47 (0.48)	2.902	1.102 (1.113)	0.742 (0.750)
	4	0.35, 0.53, 0.73, 1	0.46 (0.46)	2.987	1.116 (1.117)	0.717 (0.727)
$\alpha = 0.05$						
$1-\beta = 0.50$						
	2	0.68, 1	0.02 (0.18)	2.417	1.022 (1.031)	0.951 (0.962)
	3	0.57, 0.76, 1	0.01 (0.13)	2.677	1.031 (1.036)	0.937 (0.946)
	4	0.51, 0.65, 0.80, 1	0.00 (0.09)	2.866	1.036 (1.040)	0.930 (0.937)
$1-\beta = 0.80$						
	2	0.55, 1	0.38 (0.42)	2.257	1.067 (1.080)	0.848 (0.850)
	3	0.42, 0.68, 1	0.36 (0.39)	2.434	1.090 (1.103)	0.806 (0.812)
	4	0.35, 0.54, 0.74, 1	0.35 (0.37)	2.562	1.103 (1.110)	0.786 (0.793)
$1-\beta = 0.90$						
	2	0.49, 1	0.49 (0.49)	2.188	1.098 (1.095)	0.776 (0.776)
	3	0.36, 0.63, 1	0.46 (0.48)	2.335	1.128 (1.139)	0.718 (0.721)
	4	0.30, 0.49, 0.70, 1	0.45 (0.46)	2.428	1.146 (1.153)	0.691 (0.696)

The inflation factor $I = I(K, \alpha, \beta, \Delta^*, V^*)$ of the optimum design can be used for sample size calculations. In parentheses: minimizing Δ^*, and its corresponding I and ASN_{H_1} for equally sized stages

power $1-\beta$ of the test. The optimum designs have minimum ASN_{H_1} within the Δ-class. In Table 3.3, we also supply the corresponding optimum Δ^* for equally sized stages from Table 2.6 together with the minimized ASN_{H_1} and its inflation factor I.

First of all, notice that the minimized average sample size is only slightly below the average sample size of the optimum design with equally spaced information rates. In most cases, hence, the gain of using an optimized design is small and of merely theoretical importance. This becomes even more important as, in practical applications, the sample sizes must be integers, and hence the reduction in ASN_{H_1} might actually completely vanish. Also, it is interesting that the maximum sample size is nearly unchanged in all considered cases and, in most cases, is even somewhat smaller than the maximum sample size of the design with equally sized stages.

Although practically less important, the optimum parameters provide some interesting insight into the "mechanism" of a group sequential test design. For example, for moderate power ($1-\beta = 0.50$), it is beneficial to look at the data at some later time points rather than at equally spaced time points, i.e., $t_k^* > k/K$, $k = 1, \ldots, K-1$. This is also true for the first information rate t_1^* when considering power $1-\beta = 0.80$ or 0.90. Moreover, the smaller Δ^*'s (as compared to the design with equally sized stages) indicate that ASN_{H_1} is reduced if the condition for rejecting H_0 at early steps becomes stronger. It can be observed, however, that the deviation of the design with optimized information rates from the optimum design with equally spaced looks gets smaller and is even negligibly small for $1-\beta = 0.90$.

The figures provided in Table 3.3 can be used to determine the critical values and the stage-wise sample sizes that minimize ASN_{H_1} subject to given K, α, and $1-\beta$. As an illustration, suppose one wants to determine the optimum two-sided four-stage design at significance level $\alpha = 0.05$ and power $1-\beta = 0.80$. From the table, one finds $t_1^* = 0.35$, $t_2^* = 0.54$, $t_3^* = 0.74$, $\Delta^* = 0.35$, and $c = 2.562$. This yields the optimum critical values

$$
\begin{aligned}
u_1 &= 2.562\;, \\
u_2 &= 2.562\,(0.54/0.35)^{0.35-0.5} = 2.401\;, \\
u_3 &= 2.562\,(0.74/0.35)^{0.35-0.5} = 2.290\;, \\
u_4 &= 2.562\,(1/0.35)^{0.35-0.5} = 2.189\;.
\end{aligned}
$$

The maximum sample size 1.103 times the sample size of a fixed sample size design. For example, if the fixed sample size is 76, the maximum sample size is $1.103 \times 76 = 83.8$, and the sample sizes per stage are given by $n_1 = 0.35 \times 83.8 = 29.3$, $n_2 = (0.54 - 0.35)\,83.8 = 15.9$, $n_3 = (0.74 - 0.54)\,83.8 = 16.8$, and $n_4 = (1 - 0.74)\,83.8 = 21.8$. The average sample size under H_1 of this test procedure is $0.786 \times 76 = 59.7$. Of course, as for the designs with equally sized stages the sample sizes must be suitably rounded to the next integer, and the test characteristics slightly differ from the theoretical ones.

Group sequential designs can be specified for arbitrarily fixed sample sizes per stage, and the performance of the tests can be studied. The calculation of decision regions for unequally sized stages can be performed for one-sided and two-sided testing situations. For one-sided testing, it is possible to consider a stopping for futility option, as was described in Sect. 2.3. It is even straightforward to derive symmetric designs, for example, those due to Pampallona and Tsiatis (1994), to the more general case of unequally sized stages. Note that all of these designs require the maximum number of stages, K, to be fixed at the designing phase of the trial.

Generally, the decision regions and the test characteristics derived for unequally sized stages differ from those of the standard methods. For all designs it is possible to search for designs that are optimum with respect to a chosen criterion. Interestingly, the gains are modest, which is true for all considered types of designs, providing justification for using designs with equally sizes stages. Nevertheless, from an organizational perspective if might be reasonable to plan with unequal information rates, and the choice of decision regions and sample sizes should account for this.

3.3 The α-Spending Function Approach

The *α-spending function* or *use function* approach was introduced by Lan and DeMets (1983). Similar but less general approaches were proposed by Slud and Wei (1982) and Fleming et al. (1984), see also Kim and DeMets (1987b). The conceptual design of the α-spending function approach is to provide a test procedure that enables interim analyses at arbitrary time points of analyses. Specifically, this approach can be used if the interim analyses are not scheduled at the observation of a specific number of observations but at fixed calendar times. It is accomplished by the use of a function $\alpha^*(t_k)$ that specifies the cumulative Type I error rate spent at the information rate t_k of the kth analysis. In our context, t_k is given by $t_k = \sum_{\tilde{k}=1}^{k} n_{\tilde{k}}/N$, $k = 1, \ldots, K$, and represents the information rate that arises from the specific course of the study (see Lan and DeMets 1989b; Lan et al. 1994; DeMets and Lan 1994). $\alpha^*(t_k)$ can be any non-decreasing function with $\alpha^*(0) = 0$ and $\alpha^*(1) = \alpha$. This function must be specified in advance and laid down in the study protocol. The information rates t_k need not to be prespecified but are observed during the actual course of the trial. Consequently, neither the number of observations at the kth analysis nor the maximum number of analyses, K, must be specified in advance. The maximum sample size, N, however, must be fixed when implementing this approach.

Given the maximum sample size, N, and the function $\alpha^*(t_k)$, the Type I error spent at the first stage is $\alpha^*(t_1)$, where $t_1 = n_1/N$ denotes the information rate of the first stage of the trial. In the two-sided case the critical value for the first analysis, u_1, is determined by

$$P_{H_0}(|Z_1^*| \geq u_1) = \alpha^*(t_1) \ ,$$

which yields $u_1 = \Phi^{-1}(1 - \alpha^*(t_1)/2)$ (the one-sided case is treated analogously). For the second stage, given the information rate $t_2 = (n_1 + n_2)/N$, the condition for u_2 reads as

$$P_{H_0}(|Z_1^*| < u_1, |Z_2^*| \geq u_2) = \alpha^*(t_2) - \alpha^*(t_1) \ ,$$

which can be solved numerically by the use of the recursive integration formula. Thus, at the second stage of the test procedure the Type I error spent up to this stage is $\alpha^*(t_2)$, and the increment $\alpha^*(t_2) - \alpha^*(t_1)$ represents the amount of the Type I error rate that is spent at the second stage. The calculation of the critical values is continued in this manner. That is, the critical values for the remaining stages are computed successively through

$$P_k = P_{H_0}\left(\bigcap_{\tilde{k}=1}^{k-1}\{|Z^*_{\tilde{k}}| < u_{\tilde{k}}\}, |Z^*_k| \geq u_k\right) = \alpha^*(t_k) - \alpha^*(t_{k-1}) \ ,$$

until the overall significance level α is completely spent, i.e., until $t_K = 1$. This corresponds to a partition of the overall significance level α by

$$P_1 + \ldots + P_K = \alpha \ ,$$

and hence the overall Type I error rate is exactly equal to α. The successive calculation of the critical values is possible since, under H_0, the distribution of the vector $(Z^*_1, \ldots, Z^*_k)^T$, conditional on the observed sequence of sample sizes $n_1, \ldots, n_k$ depends only on the information rates $t_1, \ldots, t_k$. Specifically, it does not depend on the yet unobserved information rates $t_{k+1}, \ldots, t_K$.

In this way, the overall significance level α is maintained if the study proceeds whenever $t_k < 1$. Specifically, the number of stages actually performed, K, results from the smallest k, for which $t_k \geq 1$. Since at the last stage the actually observed number of observations usually exceeds N, the α-spending function

$$\tilde{\alpha}^*(t_k) = \begin{cases} \alpha^*(t_k) & \text{if } 0 \leq t \leq 1 \\ \alpha & \text{if } t > 1 \end{cases}$$

should be used in place of $\alpha^*(t_k)$ to account for this kind of "random overrunning". If, on the other hand, the study stops with a smaller maximum sample size than anticipated, i.e., $t_K < 1$, then setting $\tilde{\alpha}^*(t_K) = \alpha$ forces the procedure to fully exhaust the level α up to the last stage (see Kim et al. 1995). We will later provide an example that describes in some more detail how this calculation is performed.

A number of proposals were made in the literature for the form of the function $\alpha^*(t_k)$. The α-spending functions

$$\alpha^*_1(t_k) = \begin{cases} 2\,(1 - \Phi(\Phi^{-1}(1 - \alpha/2)/\sqrt{t_k})) & \text{(one-sided case)} \\ 4\,(1 - \Phi(\Phi^{-1}(1 - \alpha/4)/\sqrt{t_k})) & \text{(two-sided case)} \end{cases}$$

and

$$\alpha^*_2(t_k) = \alpha \ln(1 + (e - 1)t_k)$$

approximate Pocock's and O'Brien and Fleming's group sequential boundaries, respectively. $\alpha^*_1(t_k)$ is derived from the crossing probabilities of a Brownian motion (see Proschan et al. 2006) whereas $\alpha^*_2(t_k)$ is intuitively derived from the logarithmic shape of the significance level spent over the stages when using constant boundaries for equally sized stages (note that $\alpha_2^*(0) = 0$ and $\alpha_2^*(1) = \alpha$).

Kim and DeMets (1987b) proposed a family of α-spending functions indexed by a parameter $\varrho > 0$:

$$\alpha_3^*(\varrho, t_k) = \alpha\, t_k^{\varrho}\ .$$

Hwang et al. (1990) introduced the one-parameter family

$$\alpha_4^*(\gamma, t_k) = \begin{cases} \alpha\, \frac{1-e^{-\gamma t_k}}{1-e^{-\gamma}} & \text{for} \quad \gamma \neq 0 \\ \alpha\, t_k & \text{for} \quad \gamma = 0\ , \end{cases}$$

and showed that the use of $\alpha_4^*(\gamma, t_k)$ yields approximately optimal plans similar to the Δ-class of Wang and Tsiatis (1987). Alternatively, Jennison (1987) proposed a four-parameter family of approximately optimal α-spending functions. Li and Geller (1991) investigated general conditions for α-spending functions and suggested the use of piecewise linear convex functions (see Geller 1994).

The α-spending functions α_1^*, α_2^* and $\alpha_3^*(\varrho)$ are illustrated in Fig. 3.2. $\alpha_3^*(\varrho)$ is displayed for $\varrho = 1.0$, 1.5, and 2.0. These functions lie in between the functions α_1^* and α_2^* that approximate O'Brien & Fleming's and Pocock's designs, respectively. Notice that the α-spending function α_2^* is not linear in t_k, which means that constant boundaries do not correspond with a linear shape of the α-spending function.

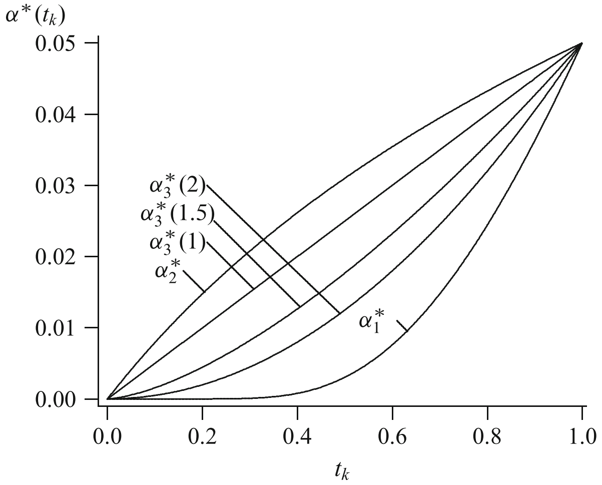

Fig. 3.2 Examples of α-spending functions. α_1^* and α_2^* approximate O'Brien & Fleming's and Pocock's design, respectively. $\alpha_3^*(\varrho)$ is plotted for $\varrho = 1.0$, 1.5, and 2.0; $\alpha = 0.05$

We illustrate the successive calculation of the critical values by the use of the α-spending function α_1^*. Suppose, in a two-sided design at significance level $\alpha = 0.05$, the first interim analysis is conducted after having performed $n_1 = 30$ observations from a maximum of $N = 100$ observations. Hence, $t_1 = 0.30$ and $\alpha_1^*(0.3) = 0.00009$, and therefore $u_1 = \Phi^{-1}(1 - 0.00009/2) = 3.929$ is the critical value for the first interim analysis. If the second interim analysis is placed at $t_2 = 0.60$ (i.e., after having observed a total of 60 patients), the second stage critical value is $u_2 = 2.670$. This value is obtained numerically under the requirement that the Type I error rate at the second stage should be $\alpha_1^*(0.6) - \alpha_1^*(0.3) = 0.00762 - 0.00009 = 0.00753$. If, at the next stage, $t_3 = 1$, then the critical value is $u_3 = 1.981$. Recall that the critical values are determined successively and the information rates are not fixed prior to the start of the study. If, for example, there were only 90 observations available for the third interim analysis, then $t_3' = 0.90$ and the critical value for the third interim analysis would be $u_3' = 2.121$. In this case, if $t_4' = 1$, the last stage critical value is $u_4' = 2.063$. Hence, although the interim analyses can be performed at arbitrary time points of analyses, the number of actually performed stages has an impact on the critical values. We will see below how the calculation of the critical values is performed if the last information is not equal to one.

The sequence of critical values $u_1 = 3.929$, $u_2 = 2.670$, $u_3 = 2.121$, $u_4 = 2.063$ corresponds to a four-stage group sequential test performed at information rates $t_1 = 0.30$, $t_2 = 0.60$, $t_3 = 0.90$, $t_4 = 1.0$. The sequence of critical values obtained by (3.7) with $\Delta = 0$ and $\boldsymbol{V} = (0.30, 0.60, 0.90, 1)$ is given by $u_1 = 3.735$, $u_2 = 2.641$, $u_3 = 2.157$, $u_4 = 2.046$. Therefore, indeed, the α-spending function α_1^* approximates the classical group sequential boundaries also for the case of unequally spaced stages. Table 3.4 shows that the α-spending function α_1^* yields critical values close to those of O'Brien and Fleming's test for some "typical" $\boldsymbol{V}$ yielding $K = 2$, 3, and 4, respectively. This is true for a wide range of information rates. Thus, it is reasonable to use this function to produce boundaries very similar

Table 3.4 Critical values u_k for Z_k^* for different K and information rates $t_1, \ldots, t_K$ using the critical values defined by (3.7) as compared to critical values obtained from the α-spending function α_1^* (O'Brien and Fleming type); significance level $\alpha = 0.05$, two-sided

		Sample sizes fixed	α-spending approach
K	$t_1, \ldots, t_K$	$u_1, \ldots, u_K$	$u_1, \ldots, u_K$
2	0.30, 1	3.581, 1.961	3.929, 1.960
	0.50, 1	2.797, 1.977	2.963, 1.969
	0.90, 1	2.135, 2.026	2.094, 2.053
3	0.30, 0.90, 1	3.700, 2.136, 2.027	3.929, 2.094, 2.053
	0.33, 0.67, 1	3.471, 2.454, 2.004	3.710, 2.511, 1.993
	0.80, 0.90, 1	2.300, 2.168, 2.057	2.250, 2.177, 2.072
4	0.20, 0.40, 0.90, 1	4.539, 3.209, 2.140, 2.030	4.877, 3.357, 2.097, 2.054
	0.25, 0.50, 0.75, 1	4.049, 2.863, 2.337, 2.024	4.333, 2.963, 2.359, 2.014
	0.30, 0.60, 0.90, 1	3.735, 2.641, 2.157, 2.046	3.929, 2.670, 2.121, 2.063

Table 3.5 Critical values u_k for Z_k^* for different K and information rates $t_1, \ldots, t_K$ using constant critical values as compared to critical values obtained from the α-spending function α_2^* (Pocock type); significance level $\alpha = 0.05$, two-sided

		Sample sizes fixed	α-spending approach
K	$t_1, \ldots, t_K$	$u_1, \ldots, u_K$	$u_1, \ldots, u_K$
2	0.30, 1	2.206, 2.206	2.312, 2.124
	0.50, 1	2.178, 2.178	2.157, 2.201
	0.90, 1	2.072, 2.072	1.989, 2.241
3	0.30, 0.90, 1	2.263, 2.263, 2.263	2.312, 2.162, 2.342
	0.33, 0.67, 1	2.289, 2.289, 2.289	2.279, 2.295, 2.296
	0.80, 0.90, 1	2.152, 2.152, 2.152	2.021, 2.271, 2.332
4	0.20, 0.40, 0.90, 1	2.362, 2.362, 2.362, 2.362	2.438, 2.427, 2.224, 2.376
	0.25, 0.50, 0.75, 1	2.361, 2.361, 2.361, 2.361	2.368, 2.368, 2.358, 2.350
	0.30, 0.60, 0.90, 1	2.334, 2.334, 2.334, 2.334	2.312, 2.321, 2.318, 2.412

to O'Brien and Fleming's test. It is also worth mentioning that using the one-sided version of α_1^* for the one-sided cases improves this approximation.

The approximation of the constant boundaries for Pocock's design when using the α-spending function α_2^* is somewhat worse if unequally spaced information rates are considered. This is illustrated in Table 3.5. We already mentioned that, for Pocock's test, the critical values defined by (3.6) and (3.7) coincide, and the α-spending function α_2^* should produce constant critical values for any information rate vector $\boldsymbol{V}$. This, however, is only true in a limited sense. Especially, for later looks at the data as compared to equal spacing, i.e., $t_k > k/K$, the function α_2^* even produces (slightly) increasing boundaries which are difficult to justify. Indeed, for unequal stage sizes, certain values of ϱ, for example, $\varrho = 0.80$, for the Kim and DeMets (1987b) α-spending function $\alpha_3^*(\varrho)$ provide a better approximation to Pocock's constant boundaries. Nevertheless, the approximation for equally sized stages behaves quite well.

We will now describe how a design with an α-spending function approach is implemented if random overrunning or random underrunning takes place. This happens if the maximum information is not met exactly but $t_K < 1$ or $t_K > 1$, respectively. This t_K and the other observed information rates need to be random in the sense that they are not allowed to be data-driven. To illustrate, imagine if close to the end of the trial it is observed that significance is not reached by far and hence the next stage is planned a long time after the maximum number of patients is expected to reach. As expected the Type I error rate is not controlled any more (see Proschan et al. 1992). The same holds true if, say, a result is near to showing significance and therefore it is decided that the next analysis is performed very soon and declared to be the final one. If the reason for not exactly meeting the planned maximum information is not data-driven, however, one can exactly account for this by defining $\alpha^*(t_K) = \alpha$ even if $t_K \neq 1$. To obtain a valid sequence of rejection boundaries the updated correlation structure must be taken into consideration. This is illustrated by an example.

Assume as above (page 76) that α_1^* was chosen at two-sided $\alpha = 0.05$ and a maximum of $N = 100$ observations were planned. We saw that $u_1 = 3.929$ at $t_1 = 0.30$ and $\alpha_1^*(0.30) = 0.00009$, and $u_2 = 2.670$ at $t_2 = 0.60$ and $\alpha_1^*(0.60) = 0.00762$ were derived from this. Now suppose the next stage is performed after 120 patients instead of the planned 100. That is, $t_K > 1$ and obviously the covariances between the test statistic at this stage (which is based on 120 observations) and the others are different from the ones that were obtained if $t_K = 1$. A solution is to "adjust" the information rates t_k observed so far as follows.

Set $t_1' = 30/120 = 0.25$ and $t_2' = 60/120 = 0.5$ for a newly defined α-spending function with $\alpha_1^{*'}(0.25) = 0.00009$ and $\alpha_1^{*'}(0.5) = 0.00762$. This produces the same rejection boundaries $u_1 = 3.929$ and $u_2 = 2.670$ for the first two stages since the covariance between Z_1^* and Z_2* depends only on the ratio between the information rates which is the same for both situations. The final rejection boundary at $t_3' = 120/120 = 1$ is $u_3' = 1.989$, i.e., somewhat larger than $u_3 = 1.981$. In other words, one must account for the fact that an overrunning took place.

For random underrunning, the same way of adjusting the information rates yields valid critical values u_3'. These are smaller than the ones for $t_K = 1$. For example, if it is decided to terminate the study after 80 observations then, as above, a newly defined α-spending function with $\alpha_1^{*'}(0.375) = 0.00009$ and $\alpha_1^{*'}(0.75) = 0.00762$ yields the same rejection boundaries for the first two stages, but $u_3' = 1.969$. That is, the use of the boundaries with $t_K = 1$ yields conservative bounds which can be improved. It is worth mentioning that in this case it might be difficult from an organizational perspective to use such improved bounds because it is difficult to authenticate that this stage was fixed to be the final one, for example, through a protocol amendment or an amendment of the Statistical Analysis Plan.

A group sequential test using the α-spending function approach is implemented using the sequence of information rates actually observed. However, when investigating the spending function approach in terms of, say, power and ASN_{H_1}, the information rates t_k and the maximum number of stages, K, must be given. That is, when planning such a trial, K and a specific choice of the information rate vector $\boldsymbol{V}$ are fixed where, conveniently, one can assume equally spaced information rates. Given K and a specific α-spending function $\alpha^*(t_k)$, the shape of the decision boundaries and hence the test characteristics can be calculated as described in the last section. Consequently, given the overall power $1 - \beta$ and the vector of information rates $\boldsymbol{V}$ the maximum and average sample size can be calculated to achieve power $1 - \beta$. Like for the other test designs, the inflation factor I can be used for sample size calculations and the expected reduction in sample size under H_1, ASN_{H_1}, can be used to assess the performance of the test. Table 3.6 shows I and ASN_{H_1} relative to the fixed sample size design for different K, $\alpha = 0.01$ and 0.05, $1 - \beta = 0.80$ and 0.90, and for the α-spending function α_1^* (O'Brien and Fleming type), α_2^* (Pocock type), and $\alpha_3^*(\varrho)$ for $\varrho = 1.0, 1.5, 2.0$. The values in the table were calculated assuming equally spaced information rates.

Table 3.6 Inflation factor I and expected reduction in sample size under H_1, relative to n_f, for different K, significance level α, and power $1-\beta$ using the α-spending functions α_1^* (O'Brien and Fleming type), α_2^* (Pocock type), and $\alpha_3^*(\varrho)$ under the assumption of equally spaced information rates

		$1-\beta = 0.80$		$1-\beta = 0.90$	
	K	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$
O'Brien and Fleming type	1	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
	2	1.001 (0.959)	1.004 (0.921)	1.001 (0.930)	1.003 (0.877)
	3	1.005 (0.894)	1.013 (0.866)	1.004 (0.847)	1.012 (0.811)
	4	1.009 (0.868)	1.020 (0.839)	1.008 (0.815)	1.018 (0.777)
	5	1.012 (0.853)	1.025 (0.824)	1.012 (0.796)	1.023 (0.759)
	10	1.022 (0.822)	1.038 (0.794)	1.021 (0.758)	1.035 (0.722)
$\varrho = 2.0$	1	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
	2	1.028 (0.882)	1.028 (0.867)	1.025 (0.822)	1.025 (0.805)
	3	1.045 (0.839)	1.045 (0.823)	1.042 (0.768)	1.041 (0.750)
	4	1.056 (0.817)	1.056 (0.801)	1.052 (0.740)	1.051 (0.722)
	5	1.064 (0.804)	1.063 (0.788)	1.059 (0.723)	1.058 (0.705)
	10	1.082 (0.780)	1.081 (0.762)	1.076 (0.692)	1.075 (0.672)
$\varrho = 1.5$	1	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
	2	1.045 (0.874)	1.047 (0.856)	1.041 (0.808)	1.042 (0.788)
	3	1.068 (0.834)	1.070 (0.814)	1.062 (0.757)	1.064 (0.734)
	4	1.082 (0.815)	1.085 (0.794)	1.075 (0.731)	1.077 (0.707)
	5	1.091 (0.803)	1.094 (0.782)	1.084 (0.716)	1.086 (0.692)
	10	1.112 (0.780)	1.116 (0.759)	1.103 (0.686)	1.106 (0.662)
$\varrho = 1.0$	1	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
	2	1.076 (0.871)	1.082 (0.850)	1.070 (0.799)	1.075 (0.777)
	3	1.108 (0.836)	1.117 (0.812)	1.099 (0.750)	1.107 (0.722)
	4	1.126 (0.820)	1.137 (0.795)	1.116 (0.727)	1.124 (0.698)
	5	1.138 (0.810)	1.150 (0.785)	1.126 (0.714)	1.136 (0.684)
	10	1.163 (0.792)	1.177 (0.766)	1.150 (0.688)	1.161 (0.657)
Pocock type	1	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)	1.000 (1.000)
	2	1.111 (0.875)	1.123 (0.855)	1.101 (0.798)	1.111 (0.777)
	3	1.153 (0.845)	1.170 (0.819)	1.140 (0.751)	1.154 (0.721)
	4	1.176 (0.831)	1.196 (0.804)	1.160 (0.730)	1.178 (0.697)
	5	1.190 (0.823)	1.212 (0.796)	1.173 (0.717)	1.192 (0.684)
	10	1.220 (0.808)	1.247 (0.780)	1.201 (0.694)	1.224 (0.660)

In parentheses: expected reduction in sample size

The effect on the inflation factor and on ASN_{H_1} when actually using unequally spaced information rates is only moderate and hence Table 3.6 can be used for planning a trial that is based on the α-spending function approach (see Jennison and Turnbull 2000; §7.2). A comparison with Tables 2.3 and 2.5 shows that the properties of the tests based on the respective α-spending function are very similar. The Wang and Tsiatis Δ-class tests with $\Delta = 0.10$, 0.25, and 0.40 have properties similar to the Kim and DeMets α-spending function with $\varrho = 2.0$, 1.5, and 1.0, respectively. Furthermore, $\varrho = 3$ yields values very near to O'Brien and Fleming's design and, as it was shown for the Wang and Tsiatis Δ-class test with $\Delta = 0.40$, it turns out that $\alpha_3^*(\varrho)$ with $\varrho = 1.0$ even yields better test characteristics than the Pocock type test. Hence, there is little reason to use α_1^* or α_2^*.

The α-spending function approach can be generalized in several ways. First, it is easy to implement an asymmetric procedure for the determination of an upper and a lower bound in a two-sided test situation. Then, two α-spending functions must be given specifying the Type I error rate for the lower and the upper bounds, respectively. Second, for planning purposes, it is quite natural to consider a function describing how the power of the procedure should be spent during the stages of the study. This yields the power-spending approach that was proposed by Bauer (1992). A similar β-spending function approach was proposed by Chang et al. (1998) and Pampallona et al. (2001), see also Anderson and Clark (2010). Here, the Type II error rate is controlled. Essentially, these designs provide an alternative method for deriving futility bounds as it was described in the last chapter. Finally, Cook (1996) describes "coupled error spending functions" for, say, multiple endpoint testing where each outcome variable is marginally monitored by specified α-spending functions.

It is tempting to use the results of an interim analysis to modify the schedule of interim looks. This is particularly true for the α-spending function approach since the time points of analyses and even the maximum number of analyses needs not to be prespecified. For example, if the test result is very near to showing significance, it could be decided to plan the next interim analysis earlier than originally planned and hence change the frequency of future analyses (Lan and DeMets 1989a). From a theoretical point of view, however, a data-driven analysis strategy is not allowed for the α-spending function approach. Indeed, there are cases in which the Type I error rate is seriously inflated, as was shown by several authors (Jennison and Turnbull 1991b; Proschan et al. 1992). In this case, therefore, one should use *adaptive* or *flexible designs* that are designed specifically for a data-driven analysis strategy and offer an even larger degree of flexibility. Part II of this book contains a description of these methodologies.

Chapter 4
Confidence Intervals, *p*-Values, and Point Estimation

It is generally agreed that the result of a clinical trial should not consist solely in a decision to reject (or not to reject) the null hypothesis, but also in further measures of evidence against the null hypothesis and in measures of the effect size. This dictum certainly applies for group sequential designs as well. The acceptability of such measures, however, is affected by the repeated significance testing nature of group sequential designs. This influence is not surprising when p-values are considered. Where the decision rules for the rejection of the null hypothesis are dictated by repeated significance testing, p-values will have to be chosen accordingly. This also holds true for confidence intervals, as they are directly linked to significance testing. Additionally, confidence intervals based on a point estimate of the effect size suffer from the poor performance of usual estimators when employed in group sequential test designs. In this context, even estimators which fulfill certain optimality criteria in fixed sample size designs usually show a considerable lack of accuracy. A prominent example is the maximum likelihood (ML) estimate for a normal mean. The sample mean does not lose its ML property in a group sequential design but it will no longer be unbiased.

Various approaches have been made to overcome these difficulties. As p-values, point estimators and confidence intervals should serve as conjoint means for interpreting study results, it is desirable that they are chosen coherently. There is a conceptual difference between confidence intervals and p-values which are intended to be used only once after the end of a trial and such measures that can be calculated at each stage of the trial. The latter are therefore called repeated confidence intervals and p-values. Point estimators are not categorized in this way as they do not suffer from multiplicity restrictions. Thus, generally there is no need to distinguish between repeated and non-repeated point estimates. Still, due

G. Wassmer, W. Brannath, *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*, Springer Series in Pharmaceutical Statistics,
DOI 10.1007/978-3-319-32562-0_4

to construction there can be limitations on their applicability. In Sect. 4.1, we will present confidence intervals and p -values both for analyses at the end of trial as well as a means for monitoring a group sequential trial. Point estimates, which mainly can be used in both situations, will be presented in Sect. 4.2.

4.1 Confidence Intervals and p-Values

This section deals with various approaches to construct confidence intervals and p-values at the end and during the course of a group sequential trial. We will first deal with analyses based on orderings of the sample space—being the most famous approach to confidence interval construction in this setting—which are designated for use after a trial has been terminated, be it due to early rejection or non-rejection of H_0 or due to reaching the last planned stage. We will then introduce another type of confidence intervals which can be employed for monitoring a group sequential trial and p -values associated with them.

4.1.1 Sample Space Orderings

When constructing confidence intervals and p-values, sample space orderings become necessary as their construction involves the task of determining the probability of obtaining a value from the sample space more extreme than a given one. In fixed sample size designs, this task is easily fulfilled if there is only one outcome variable. In this case, the sample space is one-dimensional, and the ordering underlying the construction of confidence intervals is simply the ordering on the real numbers. In group sequential designs, the sample space becomes two-dimensional, one of its dimensions describing the test statistic and the other the number of stages performed (see Sect. 1.4). Therefore, no intrinsic ordering on the sample space is given. Various suggestions have been made for imposing an ordering on such a two-dimensional sample space. The simplest, of course, is to ignore the discrete dimension of the sample space and to order the outcomes in terms of the test statistic only. Even in this simple case, however, there is a difference between ordering by the test statistic S_k and by its standardized version Z_k^*, as we will see further on. In the following, we will first describe the general procedure of deriving overall p-values and confidence intervals from an ordering and then introduce the most common orderings in the context of group sequential test designs. For an overview of the procedures, see also Jennison and Turnbull (2000), Chap. 8, and Proschan et al. (2006), Chap. 7.

One-sided overall p-values for H_0 can be calculated as follows. Recall that in the fixed sample design the p-value is given by $P_{H_0}(Z \geq z)$ or $P_{H_0}(Z \leq z)$ depending on which direction the hypothesis is formulated for the one-sided case, or as twice the minimum of these numbers in the two-sided case. The ordering on the real numbers enters this definitions evidently through the conditions of $Z \geq z$ and $Z \leq z$, respectively.

The same scheme works for group sequential designs as well, replacing Z by (Z_M^*, M) where M denotes the stage number as a random variable. For the determination of an overall p-value an ordering on the two-dimensional sample space $(\mathbb{R}, \mathbb{N}^+)$ is needed to determine whether a given observation $(z_{k'}^{*\prime}, k')$ is more or less extreme than another observation (z_k^*, k). Accordingly, the one-sided p-value is obtained by calculating

$$P_{H_0}((Z_M^*, M) \succeq (z_k^*, k)) \tag{4.1}$$

or

$$P_{H_0}((Z_M^*, M) \preceq (z_k^*, k)) \tag{4.2}$$

(Emerson and Fleming 1990), the two-sided p-value being given by twice the minimum of (4.1) and (4.2) (Chang et al. 1995) in analogy to the fixed sample size design.

In much the same way, confidence intervals can be constructed. Consider a fixed sample size design testing a normal mean where $H_0 : \mu = \mu_0$ is tested against $H_1 : \mu \neq \mu_0$. For convenience, throughout this section we will derive confidence intervals for the standardized effect $\delta = (\mu - \mu_0)/\sigma$ instead of μ as δ. A $(1 - \alpha)100\,\%$ confidence interval for δ is then given by $(\delta^L(z); \delta^U(z))$ where $\delta^L(z)$ and $\delta^U(z)$ are determined such that

$$P_{\delta^L(z)}(Z \geq z) = \alpha/2 \tag{4.3}$$

and

$$P_{\delta^U(z)}(Z \leq z) = \alpha/2\ , \tag{4.4}$$

where $Z = \sqrt{n}\,(\bar{X} - \mu_0)/\sigma$. Solving (4.3) and (4.4) yields the usual interval $((\bar{x} - \mu_0)/\sigma \pm u/\sqrt{n})$ with $u = \Phi^{-1}(1 - \alpha/2)$. This procedure is applicable to all fixed sample size designs with a one-dimensional normally distributed test statistic T.

Again, in order to solve (4.3) and (4.4), an ordering in the sample space is needed. If such an ordering is defined, a $(1 - \alpha)\,100\,\%$ confidence interval for δ can be obtained by finding $(\delta^L(z_k^*, k); \delta^U(z_k^*, k))$ such that

$$P_{\delta^L(z_k^*, k)}((Z_M^*, M) \succeq (z_k^*, k)) = \alpha/2$$

and

$$P_{\delta^U(z_k^*,k)}((Z_M^*, M) \preceq (z_k^*, k)) = \alpha/2$$

(see Tsiatis et al. 1984; Chang 1989), where $(z_{k'}^{*\prime}, k') \succeq (z_k^*, k)$ denotes that $(z_{k'}^{*\prime}, k')$ is larger or equal than (z_k^*, k). Note that both probabilities must be monotone in δ^L and δ^U, respectively, otherwise there is no unique determination of the confidence interval.

It is also possible to define point estimators on the basis of orderings of the sample space. These point estimators can be defined in two ways: Whitehead (1997) proposed to choose $\tilde{\delta}(z_k^*, k)$ such that

$$P_{\tilde{\delta}(z_k^*,k)}((Z_M^*, M) \succeq (z_k^*, k)) = P_{\tilde{\delta}(z_k^*,k)}((Z_M^*, M) \preceq (z_k^*, k)) = 0.5\ . \tag{4.5}$$

Point estimators obtained by this method fulfill the criterion of median unbiasedness. Kim (1988) alternatively suggested to use the midpoint of a confidence interval as a point estimate. With this approach, different confidence levels can lead to slightly different point estimates in the case of asymmetry.

Stage-Wise Ordering

A quite intuitive *stage-wise ordering* was originally proposed by Armitage (1957), and later by Siegmund (1978), Fairbanks and Madsen (1982), and Tsiatis et al. (1984). It basically attributes a higher level of evidence against the null hypothesis to rejection in earlier stages and a lower level of evidence against the null to stopping in later stages of the trial. For a K-stage group sequential trial with continuation regions $\mathscr{C}_k^* = (a_k; b_k)$ for $k = 1, \ldots, K-1$, $(z_{k'}^{*\prime}, k')$ is said to be larger than (z_k^*, k) if one of the following conditions is fulfilled:

1. $k' = k$ and $z_{k'}^{*\prime} > z_k^*$, i.e., within the stage the ordering is determined by the values of $z_{k'}^{*\prime}$ and z_k^*;
2. $k' < k$ and $z_{k'}^{*\prime} \geq b_{k'}$, i.e., crossing the upper bound at an earlier stage is considered a larger outcome than stopping at a later stage;
3. $k' > k$ and $z_k^* \leq a_k$, i.e., crossing the lower bound at an earlier stage is considered a smaller outcome than stopping at a later stage.

This ordering is illustrated in Fig. 4.1 for the example of a four-stage two-sided O'Brien and Fleming design with critical values u_k, i.e., with $(a_k; b_k) = (-u_k; u_k)$, $k = 1, \ldots, K$, arrows pointing from smaller to larger values. Following the arrows explains why this ordering is also known as the anti-clockwise ordering. Note that the definition of the stage-wise ordering does not cover group sequential plans with non-interval continuation regions, as, for example, those arising from the two-sided design of Pampallona and Tsiatis (1994).

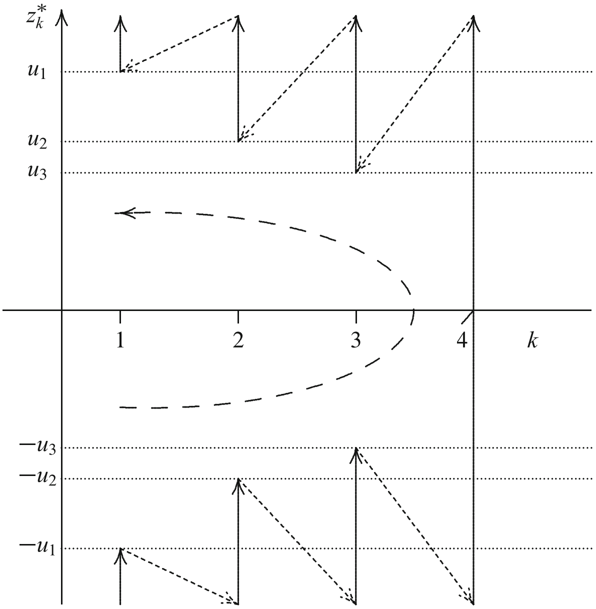

Fig. 4.1 Illustration of the stage-wise ordering of the sample space in a two-sided group sequential design with O'Brien and Fleming's boundaries ($K = 4$, $n_k \equiv n$ for $k = 1, \dots, 4$). The *elliptical arrow* illustrates the anti-clockwise direction of the ordering

Consider a four-stage design for testing $H_0 : \mu = 0$. The continuation regions are given by $\mathscr{C}_k^* = (-u_k; u_k) = (-4.049/\sqrt{k}; 4.049/\sqrt{k})$, $k = 1, \dots, 4$, in the O'Brien and Fleming case and by $\mathscr{C}_k^* = (-u_k; u_k) = (-2.361; 2.361)$, $k = 1, \dots, 4$, in the Pocock case. Let the variance be known as $\sigma = 1$ and let $n_k \equiv 22$, yielding a power of approximately $1 - \beta = 0.80$ for $\delta = 0.31$ using O'Brien and Fleming type boundaries and for $\delta = 0.33$ using Pocock type boundaries, respectively. Suppose the study continues after the first stage and the second stage yields $Z_2^* = 3$, permitting the rejection of H_0 for both designs as $3 > 4.049/\sqrt{2} = 2.863$ and $3 > 2.361$. In order to derive a 95 %-confidence interval for δ, we need to find δ^L and δ^U such that

$$P_{\delta^L(3,2)}((Z_M^*, M) \succeq (3, 2)) = 0.025$$

and

$$P_{\delta^U(3,2)}((Z_M^*, M) \preceq (3, 2)) = 0.025. \tag{4.6}$$

$P_{\delta^L(3,2)}((Z_M^*, M) \succeq (3, 2))$ consists of two addends: The first addend is the probability of stopping in stage 1 with positive Z_1^* and second one is the probability of stopping in stage 2 with $Z_2^* > 3$. Writing the equation in terms of the group sequential density (see Sect. 1.4) leads to the sum of integrals

$$\int_{u_1}^{\infty} f_{\delta^L(3,2)}(s_1, 1)\, ds_1 + \int_{\sqrt{2}\,3}^{\infty} f_{\delta^L(3,2)}(s_2, 2)\, ds_2 = 0.025 \tag{4.7}$$

(recall that the density is expressed in terms of S_k and that $S_k = \sqrt{k}\, Z_k^*$ in the case of equal stage sizes). Solving this for $\delta^L(3,2)$ yields 0.157 for O'Brien and Fleming type boundaries and 0.0737 for Pocock type boundaries.

$P_{\delta^U(3,2)}((Z_M^*, M) \preceq (3,2))$ consists of more addends. The first part is given by the probability of stopping in stage 2 with Z_2^* positive but smaller than 3. The second part consists of the probability of stopping in stages 3 or 4. The third part is given by the probability of stopping in stages 1 or 2 with negative Z_2^*. Altogether, (4.6) is given by

$$\begin{aligned} &\int_{u_2}^{\sqrt{2}\,3} f_{\delta^U(3,2)}(s_2, 2)\, ds_2 \\ &\quad + \int_{-\infty}^{\infty} f_{\delta^U(3,2)}(s_3, 3)\, ds_3 + \int_{-\infty}^{\infty} f_{\delta^U(3,2)}(s_4, 4)\, ds_4 \\ &\quad + \int_{-\infty}^{-u_1} f_{\delta^U(3,2)}(s_1, 1)\, ds_1 + \int_{-\infty}^{-u_2} f_{\delta^U(3,2)}(s_2, 2)\, ds_2 = 0.025 \end{aligned} \tag{4.8}$$

(note that the second integral has a discontinuous integration domain, consisting of the upper and lower parts of the rejection region $\mathscr{R}_3$). Solving this for $\delta^U(3,2)$ yields 0.748 for the O'Brien and Fleming case and 0.729 for the Pocock case. Considering that, given we are in stage 2, stopping in stages 3 or 4 is equivalent to not stopping at stage 2, (4.8) reduces to two integrals, leaving the equation

$$\int_{-\infty}^{-u_1} f_{\delta^U(3,2)}(s_1, 1)\, ds_1 + \int_{-\infty}^{\sqrt{2}\,3} f_{\delta^U(3,2)}(s_2, 2)\, ds_2 = 0.025$$

to be solved. This transformation makes it clear why confidence intervals following the stage-wise ordering can be derived independently of the subsequent information levels and even independently of the number of subsequent interim analyses. Therefore, this ordering enables obtaining confidence intervals not only for fixed boundaries at given information levels but also for the α-spending approach and other types of flexible monitoring (Kim and DeMets 1987a). This property is limited to the stage-wise ordering and does not apply to the other orderings presented below. Putting the results of solving (4.7) and (4.8) together, we obtain the confidence intervals $(0.157; 0.748)$ (OBF) and $(0.074; 0.729)$ (P), respectively. Note that the OBF-boundaries are more conservative at stage 2 than the P-boundaries are and therefore stopping at this stage in an OBF-design implies a stronger evidence against the null, represented here by a smaller confidence interval that is more shifted to the right.

For the O'Brien and Fleming design, the upper one-sided p-value is given by $P_{H_0}((Z^*_M, M) \succeq (3,2)) = 0.0014$ and the lower one-sided p-value is given by $P_{H_0}((Z^*_M, M) \preceq (3,2)) = 0.9986$, both probabilities consisting of the same integrals as the probabilities in (4.7) and (4.8) (again, (4.8) can be simplified as shown above). The two-sided p-value is $2 \times 0.0014 = 0.0028$. For the Pocock design, the one-sided p-values are 0.0098 and 0.9902, respectively, and the two-sided p-value is $2 \times 0.0098 = 0.0196$. Again, the evidence against the null is stronger in the OBF case.

As to point estimates, in the O'Brien and Fleming case both the midpoint of the confidence interval (0.157; 0.748) and the median unbiased estimate following (4.5) take the value of 0.452. For the Pocock case, we find different estimates, the midpoint of the confidence interval being 0.401 and the median unbiased estimate 0.419. Consistently with the remarks above, both estimates are smaller than 0.452.

We finally note that the calculated confidence intervals and point estimates can also be provided in a form that is independent of an actually realized sample size. Recall that the focus was on the estimation of the standardized effect size $(\mu - \mu_0)/\sigma$, as mentioned earlier, and the estimates in the example were calculated with $n_1 = n_2 = n = 22$. Corresponding estimates for other $n'_1 = n'_2 = n'$ are easily obtained through multiplying the values with $\sqrt{n'}/\sqrt{n}$. Equivalently, a possible way is to provide the estimates for

$$E(Z^*_k) = \frac{\mu - \mu_0}{\sigma} \sqrt{\sum\nolimits_{\tilde{k}=1}^{k} n_{\tilde{k}}}\,,$$

and retransform these for the parameter of interest. In our example, the confidence intervals for $E(Z^*_k)$ are (1.038; 4.959) (OBF) and (0.489; 4.836) (P), and the median unbiased point estimates for $E(Z^*_k)$ are 2.999 (OBF) and 2.776 (P). This is always possible and opens a more general way for providing estimates for the parameter of interest.

Likelihood Ratio Ordering

Chang and O'Brien (1986) proposed an ordering based on the likelihood ratio test for a binomial proportion which was extended to the normal case by Chang (1989) and Rosner and Tsiatis (1988). For the normal case, the likelihood ratio test (LRT) statistic for testing $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ is given by Z^*_k. Two pairs of observations $(z^{*\prime}_{k'}, k')$ and (z^*_k, k) can then be ordered according to the values of the test statistic Z^*_k itself, not considering the relation between k and k'. While for the stage-wise ordering all results of subsequent stages are rated equally as giving weaker evidence against H_0 than z^*_k, they are viewed in a more differentiated way here, assessing whether or not they lie below or above z^*_k, the final value of the test statistic when stopping at stage k. Therefore, deriving estimates for this ordering is only possible if the critical boundaries of subsequent stages are known.

Simulations by Emerson and Fleming (1990) showed that the interval based on the likelihood ratio ordering tends to be wider than the interval that is based on the stage-wise ordering for O'Brien and Fleming's design but shorter for Pocock's design. Chang et al. (1995) showed that the two-sided likelihood ratio based p-values for stopping at earlier stages and with a small overshot (test statistic only slightly above the critical value) are larger than the respective two-sided stage-wise order based p-values, whereas the order is reversed in cases of later stopping or larger overshot.

Sample Mean Ordering

Emerson and Fleming (1990) proposed a one-parameter family of orderings. For normally distributed observations, this ordering defines $(z^{*\prime}_{k'}, k')$ to be more extreme than (z^*_k, k) if

$$\frac{z^{*\prime}_{k'}}{(\sum_{\tilde{k}=1}^{k'} n_{\tilde{k}})^{\Delta-\frac{1}{2}}} > \frac{z^*_k}{(\sum_{\tilde{k}=1}^{k} n_{\tilde{k}})^{\Delta-\frac{1}{2}}} \, .$$

Emerson and Fleming (1990) proposed choosing $\Delta = 1$ which they report to have found optimal with respect to certain criteria. This choice leads to ordering by the (standardized) difference between the sample mean and μ_0 in the case of normally distributed observations and hence is based on the maximum likelihood estimate for normally distributed data.

While the stage-wise ordering is mainly based on the stopping time, this ordering, like the likelihood ratio ordering, relies mainly on the extremity of the observed effect and less on the stage it was obtained at. Emerson and Fleming (1990) showed the resulting confidence intervals to be of shorter expected length than those obtained by the stage-wise ordering in a variety of settings and in many, though not all, cases to be of shorter expected length than the likelihood ratio ordering. For the same reasons as for the likelihood ratio ordering, critical boundaries of subsequent stages have to be fixed in order to calculate confidence intervals.

Score Test Ordering

Due to the fact that the sample mean is not always covered by the confidence intervals obtained by the stage-wise ordering, Rosner and Tsiatis (1988) proposed an ordering that is based on the score test statistic S_k (see Sect. 1.4). This ordering is quite similar to the likelihood ratio ordering, including the fact that the statistic to be ordered has to be computed anew for each hypothesized parameter value. Chang et al. (1995) reported that in early stages of the trial, a rejection of the null hypothesis is often accompanied by an inconsistent p-value based on this ordering when Pocock boundaries are used, i.e., the p-value may be larger than α even if the study stopped with rejection of H_0. In this case the corresponding confidence interval would cover μ_0 in spite of rejecting H_0. For later stages, such discrepancies are rarer. Generally they do not occur for O'Brien and Fleming boundaries with constant critical boundaries for S_k.

Which ordering to choose for a given analysis is a difficult question. Many comparisons have been made on this behalf, but there is no unanimous answer. Emerson and Fleming (1990) compared the stage-wise, likelihood ratio, and sample mean orderings for one-sided procedures with respect to the length of the resulting confidence intervals. They showed that the stage-wise ordering yields wider confidence intervals than the sample mean ordering, the rank of the likelihood ratio ordering depending on the chosen type of boundaries. On the other hand, results of Chang (1989) show a slight advantage for the likelihood ratio intervals as compared to the stage-wise ordering intervals in terms of length in the case of two-sided designs. In addition to such a comparison, Rosner and Tsiatis (1988) compared the probabilities of covering a wrong mean for the stage-wise and the likelihood ratio ordering. They found a slight advantage for the likelihood ratio ordering but not for all combinations of true and hypothesized means. Chang et al. (1995) and Cook (2002) compared the p-values obtained by all four orderings presented in this section. Both agreed on preferring the likelihood ratio ordering to the other orderings. Both Chang et al. (1995) and Cook (2002) recommended refraining from the score test ordering because of its possible inconsistency with the test decision. We think a decisive advantage of the stage-wise ordering is its independence of future outcomes and its applicability for more flexible designs such as the α-spending function approach. In Part II of this book we will see that this property actually enables the use of this technique in adaptive designs.

4.1.2 Monitoring a Trial

In addition to analyzing the results of a group sequential trial after its completion, it will be monitored along its interim analyses. The confidence intervals presented in the last section cannot be used for this purpose as they are only intended for use after the completion of a group sequential trial and therefore their validity is also restricted to this situation. Other approaches to obtain confidence intervals and also p-values have been proposed for monitoring purposes. The concept of *repeated confidence intervals* (RCIs) was introduced by Jennison and Turnbull (1984, 1989) and Lai (1984). It exploits the duality of statistical tests and confidence intervals, including such values of the parameter of interest in a confidence interval which do not lead to the rejection of the null hypothesis. Since this is done in compliance with the group sequential testing procedure, the resulting confidence intervals ensure protection of the multiple significance level. They can therefore be calculated at each interim analysis allowing an accompanying analysis tool in the course of the trial.

Formally, repeated confidence intervals with confidence level $1 - \alpha$ form a sequence of intervals I_k fulfilling

$$P_\delta\left(\bigcap_{k=1}^{K}\{I_k \ni \delta\}\right) = 1 - \alpha \ ,$$

δ denoting the parameter of interest. This condition ensures that the overall coverage probability $1 - \alpha$ is protected in spite of the multiple looks. The intervals I_k are constructed using all observations obtained up to stage k. The easiest way of obtaining such a sequence of confidence intervals is by inverting the group sequential plan employed to conduct the trial. That is, the null hypothesis is shifted and all values of δ which do not lead to a rejection of the shifted null hypothesis up to the current stage are included in the confidence interval.

Consider the one-sample case with normally distributed observations and consider employing a two-sided group sequential plan with continuation regions $\mathscr{C}_k^* = (-u_k; u_k)$ for testing the hypothesis $H_0 : \mu = \mu_0$ at overall significance level α. After each stage k, the null hypothesis is not rejected if

$$-u_k < \frac{\sum_{\tilde{k}=1}^{k} \sqrt{n_{\tilde{k}}}\, Z_{\tilde{k}}}{\sqrt{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}}} < u_k \ ,$$

(see (1.3) in Sect. 1.2). Inverting this sequence of tests yields the sequence of confidence intervals

$$\frac{\bar{X}^{(k)} - \mu_0}{\sigma} \pm u_k / \sqrt{\sum\nolimits_{\tilde{k}=1}^{k} n_{\tilde{k}}} \ , \ k = 1, \ldots, K, \tag{4.9}$$

for the standardized effect $(\mu - \mu_0)/\sigma$ with a global coverage probability of $1 - \alpha$. Note that $(\bar{X}^{(k)} - \mu_0)/\sigma$ will always be the midpoint of the confidence interval, in contrast to the confidence intervals based on orderings of the sample space (see above) where it may occur that the sample mean is not included in the resulting confidence interval. A one-sided confidence interval can be obtained in much the same way using the critical value of a one-sided group sequential test design (without stopping for futility) instead of the two-sided one as above.

The confidence interval defined by (4.9) has the same structure as a confidence interval for the fixed sample size test, replacing a normal quantile by the critical value for the underlying group sequential test and the fixed sample size n by $\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}$. Replacing the normal quantile by the larger group sequential critical boundary reflects the price to be paid for the multiple inference. The difference in width as compared to the fixed sample size confidence interval varies, of course, with the type of group sequential boundaries employed: For example, for earlier stages, a Pocock design will yield much smaller confidence intervals than an

O'Brien and Fleming design, and vice versa for later stages, according to the allocation the Type I error rate to the stages.

Every sequence of continuation regions maintaining a Type I error rate α is suitable for constructing a sequence of repeated confidence intervals with global coverage probability $1 - \alpha$. There is no restriction with respect to the α -spending approach or flexible monitoring as the continuation regions of subsequent stages are not needed for the calculation of (4.9). The possibility of freely choosing the employed critical values even ensures the repeated confidence intervals to be independent from the test decision and therefore to be valid however the decision to terminate the trial was reached. They are valid at each stage of the trial, even if the stopping rule was not adhered to. This, again, is in contrast to the confidence intervals based on the ordering of the sample space as those require strict adherence to the stopping rules set beforehand.

As a disadvantage, repeated confidence intervals are prone to be conservative: They include all parameters which do not lead to a rejection of the null hypothesis up to the current stage, but as the future development of the trial is unknown, they inevitably will also include parameters which would have led to a rejection of the null hypothesis in subsequent stages and therefore should not have been included in the confidence interval in the first place. This problem is largest in early stages, where there are many stages to follow, and diminishes in later stages of the trial. Another problem may occur when using the sequence of repeated confidence intervals obtained over the stage of the trial: It may happen that the intersection of the confidence intervals is empty which will be difficult to interpret. It will, however, be only the case for very heterogeneous results over the stages.

A further monitoring instrument associated with the use of repeated confidence intervals is the use of *repeated p-values* as proposed by Jennison and Turnbull (2000). They are defined as the largest level α for which the $(1 - \alpha)100\,\%$ RCI at stage k contains the hypothesized value of the null hypothesis. Equivalently, it is the smallest overall significance level for which the group sequential test with the data obtained so far reaches rejection of H_0 provided the same class of critical values is used.

Consequently, a repeated p-value is lower than α if and only if the parameter of the null hypothesis is not within the $(1 - \alpha)100\,\%$ RCI or, equivalently, if the test leads to the rejection of H_0 at the given stage. It can therefore be used to make the test decision and provides some measure of evidence for or against the hypothesis that is adjusted for the multiple looks effect. Furthermore, they can also be used for making test decisions in one-sided or two-sided equivalence tests (Jennison and Turnbull 2000, Chap. 6). As the RCIs they can be calculated repeatedly at each stage of the trial irrespective of whether the hypothesis was rejected at (or before) the given stage.

In practice, repeated p-values are generally found by varying the significance level of the group sequential procedure employed (adhering to the chosen type of boundary) until rejection regions are found that only just allow the rejection of the null hypothesis using the actual test statistic (for more details, see Sect. 8.1). If the group sequential test is defined through the shape of the rejection boundaries (for example, the Wang and Tsiatis power family including the Pocock and O'Brien & Fleming design) it can be accomplished by one integral calculation. For example, if $z^*_{k'}$ was observed at stage k' in a K stage design with equally sized stages, the repeated p-value is

$$P_{H_0}\left(\bigcup_{k=1}^{K}\left\{|Z^*_k| \geq \left(\frac{k}{k'}\right)^{\Delta-0.5} z^*_{k'}\right\}\right) .$$

So if $k' = 1$ we have

$$P_{H_0}\big(|Z^*_1| \geq z^*_1 \text{ or } |Z^*_2| \geq z^*_1\ 2^{\Delta-0.5} \text{ or} \ldots \text{or } |Z^*_K| \geq z^*_1\ K^{\Delta-0.5}\big) ,$$

for $k' = 2$ we have

$$P_{H_0}\big(|Z^*_1| \geq z^*_2\ 2^{0.5-\Delta} \text{ or } |Z^*_2| \geq z^*_2 \text{ or} \ldots \text{or } |Z^*_K| \geq z^*_2\ K^{\Delta-0.5}\ 2^{0.5-\Delta}\big) ,$$

and so on. Finally for $k' = K$

$$P_{H_0}\big(|Z^*_1| \geq z^*_K\ K^{0.5-\Delta} \text{ or } |Z^*_2| \geq z^*_K\ 2^{\Delta-0.5}\ K^{0.5-\Delta} \text{ or} \ldots \text{or } |Z^*_K| \geq z^*_K\big)$$

needs to be calculated. Note that it is important to keep in mind that the same Δ-class of critical values is used for deriving these p-values. It becomes clear that the repeated p-value at stage k' exactly corresponds with the rejection rule of the group sequential design. By definition, if $z^*_{k'} = u_{k'}$, i.e., the test statistic coincides with the rejection boundary at stage k', the repeated p-value equals α. This is in contrast to, for example, the stage-wise ordering which takes into account at which stage a rejection of the null hypothesis was possible. To illustrate, in a three-stage O'Brien and Fleming design at two-sided $\alpha = 0.05$, if the first stage p-value yields 0.0005, the repeated p-value is 0.05 whereas the stage-wise p-value is 0.0005 (see Table 2.2). This is a somewhat strange result but illustrates the idea behind two very different approaches.

The repeated p-value is consistent with the repeated confidence interval in the following sense: whenever the repeated p-value is less than α, the confidence interval excludes all parameter values of the null hypothesis. Repeated p-values also define an ordering in the sample space as do repeated confidence intervals. It is interesting to recognize, however, that these two orderings are not the same. This was shown in Posch et al. (2008). They illustrated for the one-sided testing situation that the following property does not hold in general: the smaller the repeated p-value the larger is the lower bound of the RCI. So there are inconsistencies, a solution is to define ordering consistent repeated p-values having a different family of stopping boundaries as the ones from where the test was derived.

4.2 Point Estimation

We will now briefly consider the problem of point estimates for the unknown parameter in the group sequential context. Since the stopping boundaries have no influence on the derivation of the maximum likelihood (ML) estimate (see Chang 1989), ML estimates known from the fixed sample size settings retain their ML property. They do, however, not retain other properties in the group sequential setting: The ML estimate for, for example, one sample of normally distributed variables is a uniformly minimum variance unbiased estimate in fixed sample size designs, but it is not even unbiased in group sequential designs. The reason for the biasedness of the ML estimate lies in the selective nature of the sampling procedure: Samples leading to the termination of a group sequential trial are always selected if they are extreme.

To illustrate, consider the two-sided group sequential design with continuous rejection regions for testing $H_0 : \mu = \mu_0$ in the one-sample case. Again, similar as in Sect. 4.1, we will consider point estimates for $\delta = (\mu - \mu_0)/\sigma$ rather than for μ itself. It can easily be shown that the ML estimate is given by (see (1.17) in Sect. 1.4)

$$\hat{\delta}_{ML} = \frac{\sqrt{n_1} S_k}{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}} = \frac{\bar{X}^{(k)} - \mu_0}{\sigma} . \tag{4.10}$$

Although all individual observations follow a normal distribution, the distribution of $\hat{\delta}_{ML}$ is generally not normal and its mean is unequal to δ. This is because only samples yielding a mean with an extremely large absolute value allow an early termination of the trial, whereas samples leading to a moderate mean stipulate continuation of the trial and therefore do not entail a final point estimation. $\hat{\delta}_{ML}$ will therefore overestimate the true δ when the upper boundary is crossed and it will underestimate the true δ when the lower boundary is crossed.

Figure 4.2 shows the bias of the ML estimate in a four-stage group sequential design with both Pocock and O'Brien & Fleming type boundaries for testing $H_0 : \mu = \mu_0$ and $n_1 = \cdots = n_4 = n = 10$ in the case of $\mu_0 = 0$ and $\sigma = 1$. This bias is calculated using the group sequential density from Sect. 1.4 using the fact that $\hat{\delta}_{ML}$ can be expressed in terms of a function $h(S_k)$ of the test statistic S_k as given in (4.10). The global bias can then be calculated by

$$B_\delta(\hat{\delta}_{ML}) = \sum_{k=1}^{K-1} \int_{\mathbb{R}\setminus\mathscr{C}_k} h(s_k) f_\delta(s_k, k)\, ds_k + \int_{\mathbb{R}} h(s_K) f_\delta(s_K, K)\, ds_K - \delta . \tag{4.11}$$

For both curves in Fig. 4.2 we note a positive bias for positive values of δ and a negative bias for negative values of δ although for extreme values of δ, the bias disappears. The reason for the bias is as explained above: For moderate values of δ, only extreme samples (extreme being defined relatively to the mean of the

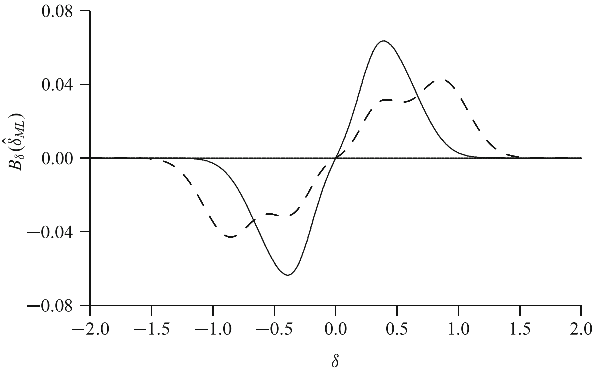

Fig. 4.2 Bias of the ML estimate in a two-sided four-stage design for testing $H_0 : \mu = 0$ with continuation regions $\mathscr{C}_k^* = (-2.361; 2.361)$ (*solid lines*) and $\mathscr{C}_k^* = (-4.049/\sqrt{k}; 4.049/\sqrt{k})$ (*dashed lines*), $k = 1, \ldots, 4$. Sample sizes are $n_k \equiv 22$, $\sigma = 1$

distribution) will lead to early stopping of the trial and thereby to an extreme point estimate. However, the more extreme the true parameter value is, the less extreme (relative to the true parameter) a sample has to be in order to allow the trial to be terminated and a point estimation to be performed. The bias of the ML estimate therefore becomes smaller for absolutely large values of δ. The smaller bias for parameter values near to 0, in comparison, is kind of an artefact: If the true parameter value is 0, in the first stages there can be no precise estimation of δ as a test statistic with values close to 0 would not lead to termination of the trial. Instead, overestimation and underestimation cancel out in this situation as it is equally probable to stop because of crossing the upper boundary as it is because of crossing the lower boundary. In the last stage, however, a precise estimation is possible in this situation. As δ becomes more extreme, the probability shifts to one of the sides and the bias becomes more and more pronounced until reaching its extreme value and then decaying to 0 again. The symmetry of the bias is a result of the symmetry of the continuation regions in this two-sided situation. The shape of the boundaries is reflected in the shape of the bias: Since the O'Brien and Fleming continuation regions $\mathscr{C}_k^*$ are non-constant in k, there is no constant increase in the bias as opposed to that belonging to the Pocock boundaries.

Figure 4.2 tells only half the tale of the bias in group sequential estimation. Looking deeper inside the mechanism of the estimation at termination it becomes clear that the bias of the ML estimate consists of a weighted mean of the conditional biases of every stage, the weight being the probability to stop at a given stage. Those single stage-wise biases exhibit a quite different behaviour as is shown in Fig. 4.3 for Pocock boundaries. The calculation of stage-wise bias again utilizes the fact that the ML estimate can be specified in terms of a function $h(S_k)$. In contrast to (4.11)

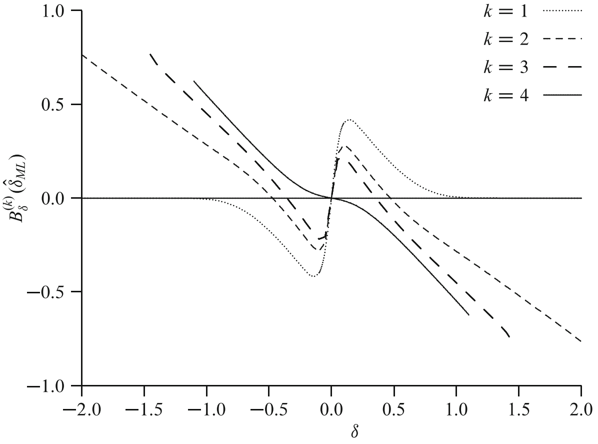

Fig. 4.3 Conditional stage-wise bias of the ML estimate in a two-sided four-stage design for testing $H_0 : \mu = 0$ with continuation regions $\mathscr{C}_k^* = (-2.361; 2.361)$, $k = 1, \ldots, 4$. Sample sizes are $n_k \equiv 22$, $\sigma = 1$

here the conditional density of S_k is needed, yielding

$$B_\delta^{(k)}(\hat{\delta}_{ML}) = \frac{\int_{\mathbb{R}\setminus\mathscr{C}_k} h(s_k) f_\delta(s_k, k)\, ds_k}{\int_{\mathbb{R}\setminus\mathscr{C}_k} f_\delta(s_k, k)\, ds_k} - \delta \tag{4.12}$$

for $k = 1, \ldots, K - 1$, and

$$B_\delta^{(K)}(\hat{\delta}_{ML}) = \frac{\int_{\mathbb{R}} h(S_K) f_\delta(S_K, K)\, ds_K}{\int_{\mathbb{R}} f_\delta(S_K, K)\, ds_K} - \delta\ . \tag{4.13}$$

The shape of the stage-wise bias is most similar to the global bias for the first stage as in the first stage the mechanism of bias induction is exactly as described above: Extreme observations lead to termination of the trial and therefore to point estimation to be performed, leading to biased estimates if the true parameter value is moderate. In the second and third stages, another mechanism is added to that first one: for extreme values of δ no bias was induced in the first stage because no selection of samples has to take place in order to allow termination of the trial. Yet, if the true parameter is extreme and the trial was nevertheless continued, there must have been a selection of moderate samples which led to overly moderate estimation in the second and third stages, resulting in an overestimation of negative parameter values and an underestimation for positive ones. The resulting bias for the second and third stages is a mixture of these two mechanisms: the first one for moderate parameter values and the second for extreme parameter values. In the forth stage, only the second mechanism plays a role as the termination of the trial is independent from the sample here (although, of course, rejection of the null is not). Therefore, the bias is a decreasing function in the true parameter

We will now consider methods of reducing the bias of the ML estimate. A widely investigated approach has been suggested by Whitehead (1986). In the case of triangular designs, he proposed using the bias adjusted estimate

$$\hat{\delta}_{\mathrm{adj}} = \hat{\delta}_{ML} - B_{\hat{\delta}_{\mathrm{adj}}}(\hat{\delta}_{ML}) \tag{4.14}$$

with $B_\delta(\hat{\delta}_{ML})$ given by (4.11). Considering that the expected value $E_\delta(\hat{\delta}_{ML})$ is given by

$$E_\delta(\hat{\delta}_{ML}) = \delta + B_\delta(\hat{\delta}_{ML}) ,$$

Eq. (4.14) corresponds to choosing the adjusted estimate $\hat{\delta}_{\mathrm{adj}}$ such that the expected value $E_\delta(\hat{\delta}_{ML})$ would take the value $\hat{\delta}_{ML}$ if the true parameter value was $\hat{\delta}_{\mathrm{adj}}$. Equation (4.14) can be solved, for example, by the Newton–Raphson algorithm (see Whitehead 1986). This approach has been transferred to group sequential designs by Todd et al. (1996). They show that there is a considerable reduction in global bias although some bias remains. This is mainly due to the fact that $\hat{\delta}$ will not vary symmetrically about its expected value due to the distribution being cut off at one side by the critical boundary.

Application to other settings than that of normally distributed observations is also possible (see, for example, Todd and Whitehead 1997, for the case of binary responses). The only requirement is a suitable initial estimate (as, for example, $(\bar{X}^{(k)} - \mu_0)/\sigma$ in the case of one sample of normally distributed data) which can be expressed as a function of the test statistic and has therefore a calculable bias. The (global) bias adjusting procedure for a parameter δ is then given by the following procedure:

1. Calculate an initial estimate $\hat{\delta}$
2. Calculate the bias $B_\delta(\hat{\delta})$ of the initial estimate
3. Calculate the adjusted estimate $\hat{\delta}_{\mathrm{adj}}$ following (4.14).

The prerequisite of an initial bias expressable in terms of a test statistic can even be relaxed when considering evaluating the bias of the estimate using resampling methods (see Pinheiro and DeMets 1997; Wang and Leung 1997).

Figures 4.2 and 4.3 suggest a slight modification to the bias adjusting procedure by Whitehead (1986). Since the global bias is a somewhat rough instrument for describing the behaviour of an estimate—ignoring which stage the estimate comes from—the bias adjusting procedure can be refined using the stage-wise conditional bias $B^k_\delta(\hat{\delta}_{ML})$ from (4.12) and (4.13) instead of the global bias $B_\delta(\hat{\delta}_{ML})$. The stage-wise adjusted estimate is found by the equation

$$\hat{\delta}^{(k)}_{\mathrm{adj}} = \hat{\delta}^{(k)}_{ML} - B_{\hat{\delta}^{(k)}_{\mathrm{adj}}}(\hat{\delta}^{(k)}_{ML}) \tag{4.15}$$

(see Troendle and Yu 1999; Coburger and Wassmer 2001, 2003). The adjusting procedure is therefore changed to

1. Calculate an initial estimate $\hat{\delta}^{(k)}$
2. Calculate the bias $B_{\delta}^{(k)}(\hat{\delta}^{(k)})$ belonging to the specific stage k
3. Calculate the adjusted estimate $\hat{\hat{\delta}}^{(k)}$ following (4.15).

Note that $\hat{\delta}$ and $\hat{\delta}^{(k)}$ are different in notation only. The ML estimate, for example, can be chosen as an initial estimate for both estimation procedures.

This stage-wise adjusting procedure reduces the stage-wise conditional bias of the initial estimate for all stages considerably (see Coburger and Wassmer 2001). Yet it works best for the last stage and for extreme parameter values for stages $k = 2, \ldots, K - 1$; for $k = 1$ and moderate parameter values in stages $k = 2, \ldots, K - 1$ it shows a similar improvement in stage-wise bias as the globally adjusted estimate in global bias. Both adjustment procedures therefore reach their aim: The global adjusted estimate shows a reduced global bias and the stage-wise adjusted estimate achieves a reduction of the stage-wise conditional bias. Evaluating the estimators crosswisely (i.e., considering the stage-wise conditional bias for the global adjusted estimate and vice versa) shows that good performance with respect to one bias type does not necessarily imply a similar behaviour with respect to the other bias type: The global adjusted estimate shows an improvement as compared to the initial estimate in the stage-wise conditional bias for $k = 1$ and a deterioration for later stages, and there is an intermediate parameter range where the stage-wise adjusted estimate has a larger global bias than the initial estimate. Still, considering again Figs. 4.2 and 4.3, the stage-wise conditional bias can be considered a finer instrument for judging the behaviour of an estimate as it is not an assembly of different situations (i.e., different stopping stages) of which only one at a time can occur.

Confidence intervals on the basis of adjusted estimators can be calculated in different ways. Whitehead (1986) gives an approximation of the standard error of the global bias adjusted estimate and constructs a classical confidence interval around the adjusted estimate. He notes, however, that the distribution of the adjusted estimate is non-normal and therefore this confidence interval may be inaccurate sometimes. Due to the fact of the non-normality of the adjusted estimate, Todd et al. (1996) make use of a proposal by Woodroofe (1992) and calculate a double standardisation of the test statistic, yielding a quantity which they claim to be normally distributed. A confidence interval around this quantity can then be calculated following classical theory. This method does not utilize the adjusted estimate but was proposed by Todd et al. (1996) as an accompanying interval for the adjusted estimate. Another suggestion was made by Pinheiro and DeMets (1997). They start with a naive confidence interval around the adjusted estimate and then apply the adjusting procedure to the endpoints of this confidence interval. They investigate their procedure and that of Whitehead (1986) but find no clear advantage for one of them. Both of these suggestions, however, have the advantage that they always include the bias adjusted estimate itself which may not be the case for the proposal by Todd et al. (1996).

A completely different approach for point estimation in group sequential trials was followed by Emerson and Fleming (1990). They employed Rao–Blackwell's theorem (see Lehmann and Casella 1998) which states that, given an unbiased estimate $\hat{\vartheta}$ for an unknown parameter and a statistic T which is sufficient and complete for ϑ, the conditional expectation $E(\hat{\vartheta}|T)$ is a uniformly minimum variance unbiased estimate. In the case of a group sequential trial with one sample of normally distributed variables with mean μ, taking $(\bar{X}_1 - \mu_0)/\sigma$ as an estimate irrespective of whether the trial was stopped or not yields a globally unbiased estimate for δ. As a sufficient and complete statistic, Emerson and Fleming (1990) identify (S_M, M). They then propose the estimate

$$\hat{\delta}_{RB} = E(\frac{\sqrt{n_1}S_1}{n_1}|(S_M, M)) = E(\frac{\bar{X}_1 - \mu_0}{\sigma}|(S_M, M)) \ .$$

Emerson (1993) derived a computationally quite extensive formula for calculating the conditional expectation, a faster algorithm was proposed by Emerson and Kittelson (1997). As Liu and Hall (1999) point out, however, (S_M, M) is not complete due to the curvature of the exponential family and therefore $\hat{\delta}_{RB}$ is not a uniformly minimum variance unbiased estimate (see also, Proschan et al. 2006, §7.1). It is, however, globally unbiased (the unbiasedness of $(\bar{X}_1 - \mu_0)/\sigma$ is not lost) but its variance is sometimes higher than that of the ML estimate and sometimes even higher than those of the stage-wisely unbiased estimators which generally tend to have a higher variance than the ML estimate.

We finally note that bias is an important but not the only important statistical property of a point estimate. An obviously more important property is precision, often quantified in terms of the mean squared error (MSE). We all know that bias may well be small while precision can be too large (as is, for example, the case for the mean of the first stage data only). Therefore, every unbiased or conditionally unbiased estimate should be investigated with regard to precision, for example, by the consideration of the MSE. In the case that the bias reduction substantially increases the mean squared error compared to the ML, the use of the bias reduced estimate is more than questionable. We will come back to this issue in the second part of this book (see Sect. 8.3).

Chapter 5
Applications

So far, we have considered the problem of testing the hypothesis $H_0 : \mu = \mu_0$ in a one-sample testing situation with normally distributed observations assuming the variance, σ^2, to be known. We already mentioned that this situation serves as a prototype for many other testing situations for which group sequential designs can be applied. Particularly, the inflation factor I, introduced in Sect. 2.1, will be used to perform the calculation of the maximum necessary sample size within a group sequential design. I depends on

- the maximum number K of stages,
- the (one-sided or two-sided) significance level α,
- the power $1 - \beta$,
- the anticipated sequence of information rates summarized in the vector $\boldsymbol{V}$,
- the chosen design, for example, on Δ when using a design within the Wang and Tsiatis family of boundaries.

Tables of the inflation factor for different designs were supplied in Chaps. 2 and 3. The average sample size either under H_1 or under another parameter value of interest can be calculated and used for the assessment of a specific design. As will be shown, this is possible for many different designs which are practically relevant.

This chapter focuses on specific characteristics and properties of group sequential tests in the most popular testing situation which we encounter in clinical trials. These are one- and two-sample tests for normally distributed observations where σ^2 is unknown, one- and two-sample tests for binary data, and statistical inference on time-to-event (survival) data. For survival data analysis, we will slightly modify the notation in order to account for the different meaning of "information" in this case.

G. Wassmer, W. Brannath, *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*, Springer Series in Pharmaceutical Statistics,
DOI 10.1007/978-3-319-32562-0_5

5.1 Normal Response

Consider testing the mean of normally distributed observations in the one-sample design if σ^2 is unknown. As in the preceding sections, let the responses X_i be independently $N(\mu, \sigma^2)$ distributed. In the fixed sample design with n_f observations $X_1, \ldots, X_{n_f}$, σ^2 is estimated by

$$\hat{\sigma}^2 = \frac{1}{n_f - 1} \sum_{i=1}^{n_f} (X_i - \bar{X})^2 \, .$$

Let $H_0 : \mu = \mu_0$ be tested against the two-sided alternative H_1 (the one-sided case is treated analogously). The test statistic for testing H_0 is given by

$$T = \frac{\bar{X} - \mu_0}{\hat{\sigma}} \sqrt{n_f} \, ,$$

which is central t distributed with $n_f - 1$ degrees of freedom. Let $G_{\vartheta,df}(\cdot)$ denote the cdf of the noncentral t distribution with non-centrality parameter ϑ and df degrees of freedom, and let $G_{df}^{-1}(\cdot)$ denote the inverse of the central t cdf with df degrees of freedom. H_0 is rejected if $|T| > G_{n_f-1}^{-1}(1 - \alpha/2)$, or, equivalently, if the two-sided p-value

$$p = \min\{2(1 - G_{0,n_f-1}(|T|)), 1\}$$

is smaller than or equal to α. In a fixed sample size design with unknown variance, the sample size n_f to meet the power $1 - \beta$ is iteratively found by searching n_f that fulfills

$$G_{\vartheta,n_f-1}\big(-G_{n_f-1}^{-1}(1 - \alpha/2)\big) + 1 - G_{\vartheta,n_f-1}\big(G_{n_f-1}^{-1}(1 - \alpha/2)\big) = 1 - \beta \, ,$$

where $\vartheta = \sqrt{n_f}(\mu - \mu_0)/\sigma$. That is, by specifying the standardized effect size $\delta = (\mu - \mu_0)/\sigma$, the sample size n_f necessary to achieve power $1 - \beta$ subject to the significance level α can be calculated by use of the central and the noncentral t distributions.

This test can be applied in the group sequential setting in an approximate sense as follows. Let the sample sizes per stage be given by $n_1, \ldots, n_K$. At stage k, the test statistic is

$$T_k^* = \frac{\bar{X}^{(k)} - \mu_0}{\hat{\sigma}^{(k)}} \sqrt{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}} \, , \qquad (5.1)$$

where $\bar{X}^{(k)}$ denotes the cumulative mean up to stage k given by (1.1) (see Sect. 1.2), and the sample variance, $\hat{\sigma}^{(k)2}$, is calculated from data obtained up to stage k through

$$\hat{\sigma}^{(k)2} = \frac{1}{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}} - 1} \sum_{i=1}^{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}} (X_i - \bar{X}^{(k)})^2 \, . \tag{5.2}$$

Consider a group sequential test design with critical values $u_1, \ldots, u_K$ which can be any sequence of critical values defined in Chaps. 2 and 3. The hypothesis H_0 is rejected if the (two-sided) p-value calculated at stage k falls below or is equal to the adjusted significance level, α_k, of the group sequential test. Equivalently, H_0 is rejected if

$$|T_k^*| \geq G^{-1}_{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}} - 1}(\Phi(u_k)) \ , \ k = 1, \ldots, K.$$

This *significance level approach* was proposed by Pocock (1977). Since the sequence of test statistics $T_1^*, \ldots, T_K^*$ fails to possess the independent and normally distributed increments structure, the Type I error rate of this test is only approximately equal to α. It can be shown, however, that the approximation is stupendously accurate. Particularly, it considerably improves the approach that directly compares the test statistic (5.1) with u_k, $k = 1, \ldots, K$.

The maximum sample size, N, necessary to achieve power $1 - \beta$ is calculated by multiplying the inflation factor, I, with the sample size of the fixed sample size design, n_f. The average sample size under H_0 or under H_1 is calculated analogously. So the vector of accumulated sample sizes is

$$(n_1, n_1 + n_2, \ldots, N) = \boldsymbol{V} I \, n_f \ ,$$

and one is able to assess the design in terms of the expected and maximum sample sizes.

We illustrate the sample size calculation in this testing situation by an example. Suppose one wishes to plan a four-stage group sequential design with constant bounds (Pocock's design) and equally sized stages. Specifying $\delta = 0.50$, $\alpha = 0.05$ and $1 - \beta = 0.80$ yields $n_f = 33.4$ (for known variance, $n_f = 31.4$, see Sect. 2.1), and

$$\boldsymbol{V} I \, n_f = (0.25, 0.50, 0.75, 1) \, 1.202 \times 33.4 = (10.0, 20.1, 30.1, 40.1) \ ,$$

where I is found from Table 2.3. Hence, if interim analyses take place after each $n_1 = \cdots = n_4 = n = 10$ observations, the power is approximately equal to 80 % when using the significance level approach. With this approach, H_0 is rejected at stage k if

$$|T_k^*| \geq G^{-1}_{k\,10-1}(\Phi(2.361)) = G^{-1}_{k\,10-1}(0.9909) \ , \ k = 1, \ldots, 4 \ ,$$

or, equivalently, if the two-sided p-value at stage k is smaller than or equal to 0.0182 (see Table 2.2). If $\delta = 0.50$, the average sample size under H_1 is $0.805 \times 33.4 = 26.9$.

Occasionally, it could be of interest to calculate the power (and the ASN) at given maximum sample size, N, for different standardized effect sizes δ. An approximate solution is to fix

$$\vartheta_k = \delta \sqrt{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}}\ , \ k = 1, \ldots, K, \tag{5.3}$$

for different δ, and to calculate the multiple integral with suitably shifted decision regions. Since this does not account for the unknown variance it will overestimate the power, especially for small to moderate sample sizes. A better solution is to find the value ζ for which

$$\begin{aligned} &G_{\vartheta,N-1}\big(G_{N-1}^{-1}(1-\alpha/2)\big) - G_{\vartheta,N-1}\big(-G_{N-1}^{-1}(1-\alpha/2)\big) \\ &\quad = \Phi\big(\Phi^{-1}(1-\alpha/2) - \sqrt{N}\zeta\big) - \Phi\big(-\Phi^{-1}(1-\alpha/2) - \sqrt{N}\zeta\big)\ , \end{aligned}$$

where $\vartheta = \delta\sqrt{N}$. This means finding the standardized effect size ζ for which the test with known variance has the same power as the test assuming the variance to be unknown. ζ can be found iteratively or can be well approximated by

$$\zeta = \frac{\operatorname{sign}(\vartheta)\,\Phi^{-1}(1-\alpha/2) - \Phi^{-1}\Big(G_{\vartheta,N-1}\big(\operatorname{sign}(\vartheta)\,G_{N-1}^{-1}(1-\alpha/2)\big)\Big)}{\sqrt{N}}\ , \tag{5.4}$$

where $\operatorname{sign}(\vartheta)$ denotes the sign of ϑ. In the one-sided case, essentially, $\alpha/2$ is replaced by α. In order to obtain the power and the ASN, the multiple integral is calculated by setting

$$\vartheta_k = \zeta \sqrt{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}}\ , \ k = 1, \ldots, K.$$

This method provides a good approximation to the power in the group sequential test design, which is illustrated in Table 5.1. For different δ, it compares the power obtained with the approximation obtained from (5.3) with the approximation obtained from (5.4) for different $n_1 = \ldots = n_K = n$ when using a two-sided four-stage Pocock design at significance level $\alpha = 0.05$. Additionally, the true power is estimated by simulation, using one million replicates, with standard error smaller than 0.0005.

The table shows that the power approximation using (5.4) considerably improves the estimation of the true power. Unless the sample size per stage, n, is very small

Table 5.1 Approximate power for testing the mean of normally distributed observations when the variance is unknown using formula (5.3) and (5.4), respectively, and true power, estimated by simulation (one million replicates), for different δ and stage sample sizes $n_1 = \cdots = n_4 = n$ in Pocock's four-stage group sequential design, $\alpha = 0.05$ (two-sided)

	Power			Power			Power		
δ	(5.3)	(5.4)	True	(5.3)	(5.4)	True	(5.3)	(5.4)	True
	$n = 3$			$n = 6$			$n = 10$		
0.0	0.050	0.050	0.055	0.050	0.050	0.053	0.050	0.050	0.052
0.3	0.146	0.129	0.126	0.250	0.233	0.226	0.390	0.374	0.366
0.4	0.227	0.196	0.185	0.413	0.383	0.370	0.626	0.603	0.592
0.5	0.333	0.284	0.265	0.596	0.558	0.539	0.824	0.803	0.793
0.6	0.457	0.390	0.361	0.762	0.724	0.704	0.939	0.927	0.920
0.7	0.587	0.507	0.469	0.883	0.852	0.836	0.985	0.980	0.977
0.8	0.709	0.624	0.582	0.952	0.933	0.923	0.997	0.996	0.995
0.9	0.812	0.731	0.689	0.984	0.975	0.969	1.000	0.999	0.999
1.0	0.889	0.820	0.782	0.996	0.992	0.990	1.000	1.000	1.000

$\delta = 0$ refers to the Type I error rate

($n = 3$), the estimated power comes close to the true power. But also for very small n the difference between the true power and the estimated power is about one third of the difference between the true power and the estimated power when using the crude approximation (5.3). The table additionally shows that the departure from the desired Type I error is quite small for all considered n. Since Pocock's design has the greatest chance to stop the trial at early stages when the sample sizes are small and hence the estimate for σ is more imprecise, the departure from the desired Type I error rate becomes even smaller for other designs (for example, O'Brien and Fleming's design). For a more systematic view of this issue, see Jennison and Turnbull (2000), §3.8. Note that the approximation of the power described in their book is different from the approach considered here, though the results are quite the same.

Confidence Intervals

It is straightforward to derive confidence intervals for μ when σ^2 is unknown. Since the estimation procedures described in Chap. 4 were essentially derived for the test statistic Z_k^*, the estimates for μ can be found from replacing the unknown variance σ^2 in Z_k^* by its estimate (5.2). For example, the sequence of $(1 - \alpha)100\,\%$ RCIs is found by inverting the test statistic (5.1), which yields

$$\bar{X}^{(k)} \pm G^{-1}_{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}-1}(\Phi(u_k))\, \frac{\hat{\sigma}^{(k)}}{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}}\,, \tag{5.5}$$

where $u_1, \ldots, u_K$ are the critical values of the two-sided test at significance level α. Note that, at stage k, one might replace the term $G^{-1}_{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}-1}(\Phi(u_k))$ with the original critical value u_k to achieve an approximate $(1-\alpha)100\,\%$ RCI. The form (5.5), however, directly corresponds with the test decision in the sense that, at stage k, H_0 is rejected exactly if and only if the RCI at stage k does not contain μ_0. Thus, (5.5) is clearly preferable.

Paired Comparisons

In practice, the situation of testing the mean of normally distributed observations in the one-sample design applies to the paired comparison design. Let the responses in pair i be denoted by X_{iA} and X_{iB} for treatments A and B, respectively. The differences $D_i = X_{iA} - X_{iB}$, $i = 1, 2, \ldots$, are assumed to be independently $N(\tilde{\mu}, \sigma^2)$ distributed, where $\tilde{\mu} = \mu_A - \mu_B$. In this setting, one usually tests $H_0 : \tilde{\mu} = 0$. At stage k, the test statistic is

$$T_k^* = \frac{\bar{D}^{(k)}}{\hat{\sigma}^{(k)}} \sqrt{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}}\ ,$$

where $\bar{D}^{(k)}$ denotes the cumulative mean difference up to stage k given by

$$\bar{D}^{(k)} = \frac{1}{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}} \sum_{i=1}^{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}} D_i\ ,$$

and the sample variance, $\hat{\sigma}^{(k)2}$, is calculated through

$$\hat{\sigma}^{(k)2} = \frac{1}{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}} - 1} \sum_{i=1}^{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}} (D_i - \bar{D}^{(k)})^2\ .$$

Since the differences D_i play the same role as the observations X_i in the one-sample t test design described above, the one-sample group sequential t test can be directly applied.

Two-Sample Comparisons

Consider two samples with independent observations $X_{1j}, X_{2j}, \ldots$, which are assumed to be $N(\mu_j, \sigma_j^2)$ distributed, $j = 1, 2$. Usually, the two samples refer to two treatment groups and one is interested in testing hypotheses concerning the means μ_1 and μ_2. Consider testing the hypothesis

$$H_0 : \mu_2 - \mu_1 = \lambda_0$$

against the one-sided and the two-sided alternative, respectively. This includes testing equality of means as well as non-inferiority testing problems, for example, in active-controlled trials (see, for example, Wang et al. 2001; D'Agostino et al. 2003; Hung et al. 2003). In non-inferiority testing problems, λ_0 needs to be predefined and reflects the mean difference which is clinically meaningless and, hence, the acceptable degree of inferiority. With such trials it is possible to establish that, say, a new treatment is at most irrelevantly worse than the control.

Within a K-stage group sequential design with $n_{j1}, \ldots, n_{jK}$ observations per stage and treatment $j, j = 1, 2$, the application of the two-sample t test is straightforward and easy to use in practice. The usage of this test depends on the assumption about the variances σ_1^2 and σ_2^2. If the variances are assumed to be equal, i.e., $\sigma_1^2 = \sigma_2^2$, the usual two-sample t test can be applied whereas elsewhere the Welch approximation is appropriate.

Equal Variances

Under the assumption of equal variances the test statistic at stage k, $k = 1, \ldots, K$, is

$$T_k^* = \frac{\bar{X}_2^{(k)} - \bar{X}_1^{(k)} - \lambda_0}{\hat{\sigma}^{(k)}} \left(\frac{1}{\sum_{\tilde{k}=1}^{k} n_{1\tilde{k}}} + \frac{1}{\sum_{\tilde{k}=1}^{k} n_{2\tilde{k}}} \right)^{-1/2},$$

where $\bar{X}_j^{(k)}$ denotes the overall mean from treatment $j, j = 1, 2$, and

$$\hat{\sigma}^{(k)2} = \frac{\left(\sum_{\tilde{k}=1}^{k} n_{1\tilde{k}} - 1\right)\hat{\sigma}_1^{(k)2} + \left(\sum_{\tilde{k}=1}^{k} n_{2\tilde{k}} - 1\right)\hat{\sigma}_2^{(k)2}}{\sum_{\tilde{k}=1}^{k}(n_{1\tilde{k}} + n_{2\tilde{k}}) - 2}$$

is the pooled variance, where $\hat{\sigma}_1^{(k)2}$ and $\hat{\sigma}_2^{(k)2}$ are the overall sample variances at stage k from the two treatment groups. Disregarding the group sequential setting, at stage k, the test statistic T_k^* is, under H_0, t distributed with $df = \sum_{\tilde{k}=1}^{k}(n_{1\tilde{k}} + n_{2\tilde{k}}) - 2$ degrees of freedom and the significance level approach can be applied with the sequence of test statistics $T_1^*, T_2^*, \ldots,$ and properly chosen degrees of freedom.

Welch Approximation

For unequal variances the Welch approximation for the degrees of freedom at stage k, df_k, is

$$df_k = \frac{\left(\frac{\hat{\sigma}_1^{(k)2}}{\sum_{\tilde{k}=1}^{k} n_{1\tilde{k}}} + \frac{\hat{\sigma}_2^{(k)2}}{\sum_{\tilde{k}=1}^{k} n_{2\tilde{k}}}\right)^2}{\frac{1}{\sum_{\tilde{k}=1}^{k} n_{1\tilde{k}}-1}\left(\frac{\hat{\sigma}_1^{(k)2}}{\sum_{\tilde{k}=1}^{k} n_{1\tilde{k}}}\right)^2 + \frac{1}{\sum_{\tilde{k}=1}^{k} n_{2\tilde{k}}-1}\left(\frac{\hat{\sigma}_2^{(k)2}}{\sum_{\tilde{k}=1}^{k} n_{2\tilde{k}}}\right)^2},$$

and the test statistic is given by

$$T_k^* = \left(\bar{X}_2^{(k)} - \bar{X}_1^{(k)} - \lambda_0\right)\left(\frac{\hat{\sigma}_1^{(k)2}}{\sum_{\tilde{k}=1}^{k} n_{1\tilde{k}}} + \frac{\hat{\sigma}_2^{(k)2}}{\sum_{\tilde{k}=1}^{k} n_{2\tilde{k}}}\right)^{-1/2}.$$

Under H_0, T_k^* is approximately t distributed with df_k degrees of freedom (disregarding the group sequential setting), and the group sequential test can be conducted with the adjusted significance levels obtained from a specific design.

For sample size calculations, we first assume balanced treatment groups over the stages of the trial. Let n_k denote the sample size per stage and treatment in the balanced case, i.e., $n_{1k} = n_{2k} = n_k$, $k = 1, \ldots, K$. Assuming equal sample sizes in the treatment groups, in a fixed sample size design with unknown variance the sample size n_f per treatment group to meet the power $1 - \beta$ is found by searching n_f that fulfills

$$1 - G_{\vartheta,2n_f-2}\big(G_{2n_f-2}^{-1}(1-\alpha)\big) = 1 - \beta$$

in the one-sided case, and

$$G_{\vartheta,2n_f-2}\big(-G_{2n_f-2}^{-1}(1-\alpha/2)\big) + 1 - G_{\vartheta,2n_f-2}\big(G_{2n_f-2}^{-1}(1-\alpha/2)\big) = 1 - \beta$$

in the two-sided case, where $\vartheta = \sqrt{n_f/2}\,(\mu_2 - \mu_1 - \lambda_0)/\sigma$. The solution for n_f is computed iteratively. As for the one-sample t test situation, the maximum sample size of the group sequential design is $I n_f$, and the vector of accumulated sample sizes per treatment group is

$$(n_1, n_1 + n_2, \ldots, N) = V I n_f\ ,$$

where I is calculated under a specific group sequential test design. The average sample size can be computed analogously using the expected reduction in sample size relative to the fixed sample size design, which can be obtained from the respective tables.

If a sample size allocation ratio $r = n_{2k}/n_{1k} \neq 1$ is specified, the total maximum sample size and the average sample size are multiplied by

$$\frac{(1+r)^2}{4\,r}\ ,$$

which is in analogy to the fixed sample size design. This term is derived from equating the non-centrality parameter ϑ in the unbalanced design with the non-centrality parameter in the balanced design. That is, from

$$\frac{\mu_2 - \mu_1 - \lambda_0}{\sigma}\sqrt{\frac{r\,n_{1k}}{1+r}} = \frac{\mu_2 - \mu_1 - \lambda_0}{\sigma}\sqrt{\frac{n_k}{2}}$$

it follows that the sample size per stage in the unbalanced design is

$$n_{1k} + r\,n_{1k} = \left(\frac{1+r}{2\,r} + \frac{r(1+r)}{2\,r}\right) n_k = \frac{(1+r)^2}{4\,r} 2\,n_k\ ,$$

$k = 1, \ldots, K$. Note that the latter term is minimum for $r = 1$, which means that equally sized treatment groups require the smallest number of observations.

Occasionally, it is more appropriate to express the non-inferiority margin in terms of the ratio of the means μ_1 and μ_2. The hypothesis to be tested is then given by $H_0 : \mu_2/\mu_1 = \lambda_0$, where λ_0 is a properly chosen non-inferiority margin. In the non-inferiority setting, H_0 is tested against the one-sided alternative. In a single stage design, it is straightforward to apply the two-sample t test (Hauschke et al. 1999), and this can be easily extended to the group sequential setting. The confidence intervals are found by finding the values λ_0 that do not lead to a rejection of the corresponding null hypothesis when using the appropriate test statistic. Note that, generally, Fieller's theorem (Fleiss 1986) applies, which means that there are cases for which no finite interval exists. We also note that in this setting equally sized treatment groups are, in general, not optimum and one can find an optimizing sample size allocation rate r^* that reduces the necessary total sample size (for an extension to three-arm clinical trials in the fixed sample size situation, see Pigeot et al. 2003; Röhmel and Pigeot 2010; Schlömer and Brannath 2013).

Extensions

There are many other situation for normally distributed data where the group sequential t test can be applied. The most prominent example is the normal linear model where a parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is estimated from the data using the maximum likelihood technique. In the fixed sample size case, it is well known that for testing, say, $H_0 : \beta_1 = 0$, the t test with appropriate chosen degrees of freedom and properly estimated variance can be used. Under certain regularity conditions, Jennison and Turnbull (1997a,b) showed that the sequence of test statistics obtained in a group sequential sampling scheme asymptotically fulfills the independent and normally distributed increments structure and hence the theory of the group sequential designs derived for the normal case with known σ^2 can be applied. The normal linear model applies to, for example, the cross-over design and it follows that the group sequential theory can be easily adopted for this situation (for a two-stage crossover design, see Cook 1995). It also applies to treatment comparisons that are adjusted for covariates, such as stratification variables or confounding factors. The application of the group sequential theory is even possible for the normal linear model with dependent observations or, more generally, for generalized linear models where even different types of response variables can be modelled (Jennison and Turnbull 1997a). The work of Jennison and Turnbull hence serves as a basis for a wide range of applications in group sequential testing. Exact critical values for the t test, the F test, and the χ^2 test situation were also derived (Jennison and Turnbull 1991a). For further details, the reader is referred to the respective chapters in Jennison and Turnbull (2000) and the references therein.

5.2 Binary Response

In many trials the primary endpoint is the occurrence of an event, for example, the event that the treatment was successful. That is, the outcome is a dichotomous measure and one wants to perform statistical tests on the proportion of an event. In this section, we briefly describe some group sequential tests in proportion trials for the one-sample and the two-sample testing situation.

5.2.1 Testing a Single Rate

If an event rate for one sample of observations is considered, the hypothesis is specified through

$$H_0 : \pi = \pi_0 \ ,$$

where π denotes the probability that the event occurs. For testing H_0, the exact binomial test or an approximate test can be performed, and H_0 can be tested against a one-sided or a two-sided alternative. As an example, this case appears in the situation of paired samples with dichotomous data where one wants to perform a McNemar test. This testing situation is also common in Phase II clinical trials (Fleming 1982).

Consider first the case with a fixed number, n_f, of observations. If the power is directed towards $\pi > \pi_0$, the exact (one-sided) upper p-value, p_U, is computed as

$$p_U = \sum_{i=m}^{n_f} B(i, n_f, \pi_0) \ ,$$

where m denotes the observed number of events, and

$$B(x, n, \pi) = \binom{n}{x} \pi^x (1-\pi)^{n-x} \quad \text{for } 0 \leq x \leq n$$

is the binomial probability function. If the power is directed towards $\pi < \pi_0$, the exact lower (one-sided) p-value, p_L, is

$$p_L = \sum_{i=0}^{m} B(i, n_f, \pi_0) \ .$$

In the two-sided case, the p-value can be calculated as $p = 2\min\{p_U, p_L\}$. In the one-sided and the two-sided case, H_0 is rejected if the corresponding p-value falls short of α.

The approximate test statistic Z for testing H_0 is

$$Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)}} \sqrt{n_f} \ , \tag{5.6}$$

where $\hat{\pi} = m/n_f$ is the observed event rate. H_0 is rejected if Z exceeds the specified critical value $u = \Phi^{-1}(1-\alpha)$ in the one-sided setting (or falls short of $-u$), or if $|Z|$ exceeds $u = \Phi^{-1}(1-\alpha/2)$ in the two-sided case.

Fixing the alternative $H_1 : \pi = \pi_1$, in a fixed sample size design the required sample size n_f to achieve power $1-\beta$ is approximately given by

$$n_f = \left(\frac{\Phi^{-1}(1-\alpha)\sqrt{\pi_0(1-\pi_0)} + \Phi^{-1}(1-\beta)\sqrt{\pi_1(1-\pi_1)}}{\pi_1 - \pi_0} \right)^2 \tag{5.7}$$

in the one-sided case, and α replaced by $\alpha/2$ in the two-sided case. This easily follows from the approximate normality of (5.6) and

$$E(Z) = \frac{\pi_1 - \pi_0}{\sqrt{\pi_0(1-\pi_0)}} \sqrt{n_f} \quad \text{and} \quad \text{Var}(Z) = \frac{\pi_1(1-\pi_1)}{\pi_0(1-\pi_0)} \ .$$

Note that it is possible to calculate the exact power with the help of the binomial cdf. Then, it is possible to "adjust" the necessary sample size, n_f, accordingly. Even so, the approximation (5.7) works quite well, except if π_0 comes very close to 0 or 1. Hence, (5.7) is reasonable and widely used in practice.

Adopting the approximate test to the group sequential design, at stage k the standardized test statistic

$$Z_k^* = \frac{\hat{\pi}^{(k)} - \pi_0}{\sqrt{\pi_0(1-\pi_0)}} \sqrt{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}}\ ,\ k = 1, \ldots, K,$$

where $\hat{\pi}^{(k)}$ is the total success rate in the first k stages, is considered. As for the prototype of normally distributed data with known variance, the test statistics Z_k^*, $k = 2, \ldots, K$, are weighted sums of the stage-wise test statistics which are approximately normally distributed. Hence, approximately, the theory of the group sequential tests derived for the prototype case directly applies. That is, the maximum sample size, N, necessary to achieve power $1 - \beta$ is calculated by multiplying the inflation factor, I, with the sample size of the fixed sample size design, n_f. The vector of accumulated sample sizes is

$$(n_1, n_1 + n_2, \ldots, N) = \boldsymbol{V}\, I\, n_f\ ,$$

and the average sample size under H_0 or under H_1 is approximately given by multiplying n_f with the average sample size reduction relative to n_f.

To illustrate these calculations, suppose we wish to plan a three-stage design with O'Brien and Fleming boundaries and equally sized stages at one-sided significance level $\alpha = 0.025$, yielding critical values $u_k = 3.4711/\sqrt{k}$, $k = 1, 2, 3$ (see Table 2.1). For power $1-\beta = 0.80$, the inflation factor I and the expected reduction in sample size under H_1 relative to n_f are given by 1.017 and 0.856, respectively, which is found from Table 2.3. Let $H_0 : \pi = 0.40$, and the alternative of interest be given by $H_1 : \pi = 0.20$. Since

$$n_f = \left(\frac{\Phi^{-1}(1-0.025)\sqrt{0.40 \times 0.60} + \Phi^{-1}(0.80)\sqrt{0.20 \times 0.80}}{-0.20} \right)^2 = 42.0\ ,$$

the vector of accumulated sample sizes in the group sequential design becomes

$$\boldsymbol{V}\, I\, n_f = (0.333, 0.667, 1)\, 1.017 \times 42.0 = (14.2, 28.4, 42.7)\ ,$$

and the ASN under H_1 is $0.856 \times 42.0 = 36.0$. Hence, it might be reasonable to conduct interim analyses after 15 and 29 observations, respectively, and the final analysis after 43 observations. Note that, if no success was observed in the first 15 observations, the test statistic is $Z_1^* = 3.1623 < 3.4711$. It is therefore not possible at all to reject H_0 in favor of H_1 in the first interim analysis under the prescribed plan. Therefore, it could be wise to skip the first interim analysis, and to use a plan with unequally sized stages (see Chap. 3).

Alternatively, one might use the exact p-values for testing H_0 and apply the significance level approach introduced in the last section. That is, in the example the p-values p_L after 15, 29, and 43 observations are compared with 0.00026, 0.0071, and 0.0225, respectively (see Table 2.2). Note that the differences are only small, especially if π_0 is not near 0 or 1. The decision regions differ slightly, if at all (due to the discreteness of the test statistic), and it is reasonable to use the approximate test.

For computing the power and the ASN at given maximum sample size N and rate π, it is possible to compute the multiple integral at given

$$\vartheta_k = \zeta \sqrt{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}}\ ,\ k = 1, \ldots, K,$$

similarly as described in Sect. 5.1. In the two-sided case, the standardized effect ζ is estimated by finding ζ for which

$$\begin{aligned}&\Phi\big(\Phi^{-1}(1-\alpha/2) - \zeta\sqrt{N}\big) - \Phi\big(-\Phi^{-1}(1-\alpha/2) - \zeta\sqrt{N}\big)\\ &\quad = \Phi\left(\frac{\Phi^{-1}(1-\alpha/2) - E(Z)}{\sqrt{\mathrm{Var}(Z)}}\right) - \Phi\left(\frac{-\Phi^{-1}(1-\alpha/2) - E(Z)}{\sqrt{\mathrm{Var}(Z)}}\right),\end{aligned}$$

where

$$E(Z) = \frac{\pi - \pi_0}{\sqrt{\pi_0(1-\pi_0)}}\sqrt{N}\ ,\ \mathrm{Var}(Z) = \frac{\pi(1-\pi)}{\pi_0(1-\pi_0)}$$

are expected value and variance, respectively, of the test statistic

$$Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)}}\sqrt{N}\ .$$

One can use the exact solution for ζ, or ζ can be well approximated by

$$\zeta = \frac{\mathrm{sign}(\pi - \pi_0)\,\Phi^{-1}(1-\alpha/2)\left(\sqrt{\pi(1-\pi)} - \sqrt{\pi_0(1-\pi_0)}\right)\sqrt{1/N} + \pi - \pi_0}{\sqrt{\pi(1-\pi)}}\ , \tag{5.8}$$

where, in the one-sided case, $\alpha/2$ is replaced by α. This technique provides a good approximation to the power (and the ASN) at given maximum sample size N. Clearly, as N gets large,

$$\zeta \approx \frac{\pi - \pi_0}{\sqrt{\pi(1-\pi)}}\ , \tag{5.9}$$

which can also be used for the two-sided case. This is a simpler form, though the approximation performs badly, especially for π_0 close to 0 or 1.

In the binary case, exact power calculations are possible with only little effort. For this purpose, let $m^{(k)}$ denote the total number of successes observed in the first k stages. The trial is continued if the approximate test statistic $Z_k^* = Z_k^*(m^{(k)})$ lies within the continuation region $\mathscr{C}_k^*$, $k = 1, \ldots, K-1$, and H_0 is rejected if $Z_k^* \notin \mathscr{C}_k^*$ for some k, $k = 1, \ldots, K$. Thus,

$$P_\pi(\text{reject } H_0) = 1 - \sum_{\{i: Z_K^*(i) \in \mathscr{C}_K^*\}} H_K(i, \pi) \ , \tag{5.10}$$

where

$$H_k(m^{(k)}, \pi) = \sum_{\{i: Z_{k-1}^*(i) \in \mathscr{C}_{k-1}^*\}} H_{k-1}(i, \pi)\, B(m^{(k)} - i, n_k, \pi) \ ,$$
$$m^{(k)} = 0, \ldots, n_1 + \cdots + n_k \ , \ k = 2, \ldots, K,$$

and

$$H_1(m^{(1)}, \pi) = B(m^{(1)}, n_1, \pi) \ , \ m^{(1)} = 0, \ldots, n_1,$$

recursively defines the function H_K, analogous to the recursive formula (1.21) in the normal case (see Elfering and Schultz 1973).

For illustrating the performance of using formula (5.8) and (5.9), respectively, Table 5.2 supplies the approximation of the power together with the exact power for some values of π_0 and $\delta = \pi - \pi_0$. In the table, a two-sided four-stage design with constant critical values at significance level $\alpha = 0.05$ is considered.

Table 5.2 illustrates that the Type I error rate is close to the nominal level even for small n. Except for $\pi_0 = 0.05$, the Type I error hardly exceeds α. Nevertheless, by exact calculation it is possible to change the decision regions slightly in order to achieve a valid level α testing procedure. Note that this might involve a departure from the original design (for example, a design with constant critical values), but the obtained test might behave more satisfactory. The table also illustrates that the approximation (5.9) is useless for small π_0. Only if π_0 comes close to 0.50, the power approximation is satisfactory. On the other hand, although not optimal, the approximation based on (5.8) behaves much better and might be used in practice.

Fleming (1982) supplied tables for binomial tests that have Type I error probabilities and power close to α and $1 - \beta$, respectively. Similar calculations for binary data were also described in Schultz et al. (1973).

Table 5.2 Approximate power for testing a single rate using formula (5.8) and (5.9), respectively, and true power, calculated with the recursive formula (5.10), for different π_0, $\delta = \pi - \pi_0$, and stage sample sizes $n_1 = \cdots = n_4 = n$ in Pocock's four-stage group sequential design, $\alpha = 0.05$ (two-sided)

		Power			Power			Power		
π_0	δ	(5.9)	(5.8)	True	(5.9)	(5.8)	True	(5.9)	(5.8)	True
		$n = 3$			$n = 6$			$n = 10$		
0.05	0.0	0.050	0.050	0.073	0.050	0.050	0.059	0.050	0.050	0.034
	0.1	0.133	0.333	0.424	0.223	0.478	0.549	0.346	0.628	0.633
	0.2	0.290	0.641	0.736	0.526	0.841	0.902	0.757	0.950	0.966
	0.3	0.495	0.843	0.908	0.803	0.970	0.989	0.958	0.997	0.999
	0.4	0.714	0.949	0.976	0.954	0.997	0.999	0.998	1.000	1.000
	0.5	0.892	0.990	0.996	0.996	1.000	1.000	1.000	1.000	1.000
	0.6	0.982	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.10	0.0	0.050	0.050	0.051	0.050	0.050	0.056	0.050	0.050	0.041
	0.1	0.115	0.219	0.259	0.186	0.327	0.417	0.284	0.454	0.497
	0.2	0.263	0.498	0.557	0.479	0.724	0.811	0.706	0.884	0.916
	0.3	0.473	0.745	0.802	0.780	0.934	0.969	0.947	0.991	0.996
	0.4	0.709	0.905	0.937	0.952	0.993	0.997	0.997	1.000	1.000
	0.5	0.902	0.979	0.987	0.997	1.000	1.000	1.000	1.000	1.000
	0.6	0.988	0.998	0.999	1.000	1.000	1.000	1.000	1.000	1.000
0.40	0.0	0.050	0.050	0.028	0.050	0.050	0.033	0.050	0.050	0.041
	0.1	0.091	0.096	0.082	0.134	0.142	0.126	0.195	0.205	0.193
	0.2	0.235	0.235	0.237	0.428	0.428	0.401	0.644	0.644	0.626
	0.3	0.528	0.477	0.504	0.835	0.800	0.769	0.970	0.960	0.949
	0.4	0.889	0.787	0.801	0.996	0.986	0.971	1.000	1.000	0.999
	0.5	1.000	0.988	0.975	1.000	1.000	1.000	1.000	1.000	1.000

$\delta = 0$ refers to the Type I error rate

Confidence Intervals

As for the normal case it is straightforward to derive confidence intervals from the test statistic Z_k^*. Since (5.6) is monotone in π_0, it is easy to calculate parameter values from the corresponding conditions for $Z_k^* = Z_k^*(\pi_0)$. For example, the $(1 - \alpha)100\,\%$ RCI is defined by the values π_0 that do not lead to a rejection of the corresponding null hypothesis at stage k when using the critical values u_k, $k = 1 \ldots, K$, from the group sequential test with two-sided level α. That is, the sequence of RCIs is given by

$$I_k = \{\pi_0 : Z_k^*(\pi_0) \in (-u_k; u_k)\}\ ,\ k = 1, \ldots, K,$$

which involves a quadratic inequality. Approximately,

$$I_k = \left(\hat{\pi}^{(k)} \pm u_k \sqrt{\frac{\hat{\pi}^{(k)}(1-\hat{\pi}^{(k)})}{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}}} \right) \ , \ k = 1, \ldots, K.$$

Note that, when using the significance level approach with the exact p-values introduced at the beginning of this section, the RCIs are Clopper–Pearson type intervals which can also be solved analytically.

Different types of RCIs were compared by Coe and Tamhane (1993). Jennison and Turnbull (1983) and Duffy and Santner (1987) proposed exact confidence intervals and p-values upon termination of the trial. Some of these approaches are briefly described in Jennison and Turnbull (2000), §12.1.

5.2.2 Parallel Group Design

Consider now comparing two treatment arms in a parallel group design. Let the hypothesis to be tested be given by

$$H_0 : \pi_1 = \pi_2 \ ,$$

where $\pi_j, j = 1, 2$, are the event rates in treatment j. Let H_0 be tested against the one-sided or two-sided alternative. Consider first the fixed sample size case with sample size n_{fj} in treatment $j, j = 1, 2$. The approximate test statistic for testing H_0 is

$$Z = \frac{\hat{\pi}_2 - \hat{\pi}_1}{\sqrt{\hat{\bar{\pi}}(1-\hat{\bar{\pi}})}} \left(\frac{1}{n_{f1}} + \frac{1}{n_{f2}} \right)^{-1/2} , \qquad (5.11)$$

where $\hat{\pi}_1$ and $\hat{\pi}_2$ are the observed rates in the two treatment groups and $\hat{\bar{\pi}} = (n_{f1}\hat{\pi}_1 + n_{f2}\hat{\pi}_2)/(n_{f1} + n_{f2})$ is the observed overall rate. This test statistic coincides with the χ^2-test in four-fold tables for the two-sided case and is an extension of it for the one-sided case. Z is approximately standard normal distributed. Hence, one can use the standard normal percentiles as for the one-sample case.

If a sample size allocation ratio $r = n_{f2}/n_{f1}$ is specified, the required sample size n_{f1} for the first treatment group to achieve power $1 - \beta$ is approximately given by (see Machin and Campbell 1987)

$$\begin{aligned} n_{f1} = \Bigg(& \Phi^{-1}(1-\alpha)\sqrt{\left(1+\frac{1}{r}\right)\bar{\pi}(1-\bar{\pi})} \\ & + \Phi^{-1}(1-\beta)\sqrt{\pi_1(1-\pi_1) + \frac{\pi_2(1-\pi_2)}{r}} \Bigg)^2 \Big/ (\pi_2 - \pi_1)^2 \end{aligned} \qquad (5.12)$$

in the one-sided case, and α replaced by $\alpha/2$ in the two-sided case. The sample size of the second treatment group is $n_{f2} = r n_{f1}$. As for the one-sample case, formula (5.12) follows from the approximate normality of (5.11) and from the fact that $E(Z)$ and $\text{Var}(Z)$ are approximately given by

$$E(Z) = \frac{\pi_2 - \pi_1}{\sqrt{\bar{\pi}(1-\bar{\pi})}}\sqrt{\frac{r\,n_{f1}}{r+1}}$$

and

$$\text{Var}(Z) = \frac{r\pi_1(1-\pi_1) + \pi_2(1-\pi_2)}{\bar{\pi}(1-\bar{\pi})(1+r)} \ ,$$

where $\bar{\pi} = (\pi_1 + r\pi_2)/(1+r)$.

Note that the sample size formula (5.12) depends on both parameters, π_1 and π_2. For sample size calculations, it is therefore necessary to specify not only the effect $\delta = |\pi_2 - \pi_1|$ which is worthwhile to detect, but also the rate π_1 (or π_2). For example, when comparing a new treatment with a control, besides the clinically meaningful improvement in the success rate, the success rate in the control group must be specified. This is associated with the problem that the test statistic Z depends, under H_0, on the nuisance parameter $\pi = \pi_1 = \pi_2$.

Furthermore, equal sample sizes for each treatment group, i.e., $r = 1$, does not result in a minimum sample size requirement. In fact, it can be shown that unequally sized treatment groups may yield a smaller required total sample sizes $n_{f1} + n_{f2}$. This is due to the fact that the variances of the success rate differ between the two treatment groups. Nevertheless, for most cases, the gain when searching for optimum designs with respect to this issue is only very small and hardly practically relevant.

Within a K-stage group sequential design with $n_{j1}, \ldots, n_{jK}$ observations at stage k, $k = 1, \ldots, K$, and treatment j, $j = 1, 2$, consider the test statistic

$$Z_k^* = \frac{\hat{\pi}_2^{(k)} - \hat{\pi}_1^{(k)}}{\sqrt{\hat{\bar{\pi}}^{(k)}(1 - \hat{\bar{\pi}}^{(k)})}}\left(\frac{1}{\sum_{\tilde{k}=1}^{k} n_{1\tilde{k}}} + \frac{1}{\sum_{\tilde{k}=1}^{k} n_{2\tilde{k}}}\right)^{-1/2} , \tag{5.13}$$

where $\hat{\bar{\pi}}^{(k)}$ is the overall rate in the first k stages. It must be recognized that—as for the t test situation—the sequence of test statistics $Z_1^*, Z_2^*, \ldots$ fails to possess the independent increment structure since Z_k^* cannot be written as a sum of independent test statistics. Nonetheless, the theory of the group sequential tests derived for the normal case with known variance can be applied in an approximate sense and it can be shown that the test behaves quite well. That is, the maximum sample sizes, N_j, $j = 1, 2$, necessary to achieve power $1 - \beta$ are calculated by multiplying the inflation factor, I, with the sample sizes of the fixed sample size design, n_{fj}, $j = 1, 2$.

The vectors of accumulated sample sizes are

$$(n_{j1}, n_{j1} + n_{j2}, \ldots, N_j) = \boldsymbol{V} I n_{fj} \ , \ j = 1, 2 \ ,$$

and the average sample size under H_0 or under H_1 per treatment group is approximately given by multiplying $n_{f1} + n_{f2}$ with the average sample size reduction relative to n_f provided in the respective tables.

As an example, consider a four-stage group sequential design with O'Brien and Fleming boundaries and equally sized stages at one-sided significance level $\alpha = 0.025$ yielding critical values $u_k = 4.0486/\sqrt{k}, \ k = 1, \ldots, 4$ (see Table 2.1). For power $1 - \beta = 0.90$, the inflation factor I and the expected reduction in sample size under H_1 relative to n_f are given by 1.022 and 0.767, respectively, which is found from Table 2.3. Suppose it is desired to find the necessary sample size if the minimum clinically relevant effect is $\delta = \pi_2 - \pi_1 = 0.30$ and $\pi_1 = 0.10$. If $r = 1$, the sample sizes $n_{f1} = n_{f2} = n_f$ per treatment group in a fixed sample size design are given by

$$\begin{aligned} n_f &= \left(\frac{\Phi^{-1}(0.975)\sqrt{2 \times 0.25 \times 0.75} + \Phi^{-1}(0.90)\sqrt{0.10 \times 0.90 + 0.40 \times 0.60}}{0.20} \right)^2 \\ &= 41.5 \ . \end{aligned}$$

The vector of accumulated sample sizes per treatment group in the group sequential design becomes

$$\boldsymbol{V} I n_f = (0.25, 0.50, 0.75, 1)\, 1.022 \times 41.7 = (10.7, 21.3, 32.0, 42.6) \ ,$$

and it is reasonable to perform interim analyses after each 11 observations per treatment group. The ASN per treatment group under H_1 is $0.767 \times 41.7 = 32.0$. In this example, an optimum allocation rate r^* is found to be 0.92 yielding a total sample size in the fixed design of $n_{f1} + n_{f2} = 43.4 + 39.8 = 83.2$. This is essentially the same as for $r = 1$, though it might be desirable to have fewer observations in the second treatment group.

Similar to the one-sample case, exact power calculations are possible through the use of a recursive formula. Let $m_j^{(k)}$ denote the total number of successes in treatment group $j, j = 1, 2$, observed in the first k stages. The trial is continued if the approximate test statistic $Z_k^* = Z_k^*(m_1^{(k)}, m_2^{(k)})$ lies within the continuation region $\mathscr{C}_k^*$, $k = 1, \ldots, K - 1$, and H_0 is rejected if $Z_k^* \notin \mathscr{C}_k^*$ for some k, $k = 1, \ldots, K$. Thus,

$$P_{\pi_1,\pi_2}(\text{reject}\, H_0) = 1 - \sum_{\{i_1, i_2 : Z_K^*(i_1, i_2) \in \mathscr{C}_K^*\}} \!\!\!\! H_K(i_1, i_2, \pi_1, \pi_2) \ , \qquad (5.14)$$

where

$$\begin{aligned}&H_k(m_1^{(k)}, m_2^{(k)}, \pi_1, \pi_2)\\ &\quad = \sum_{\{i_1,i_2: Z^*_{k-1}(i_1,i_2)\in \mathscr{C}^*_{k-1}\}}\sum H_{k-1}(i_1, i_2, \pi_1, \pi_2)\, B(m_1^{(k)}- i_1, m_2^{(k)}- i_2, n_{1k}, n_{2k}, \pi_1, \pi_2)\,,\\ &\quad m_1^{(k)} = 0, \dots, n_{11} + \cdots + n_{1k}\ ,\ m_2^{(k)} = 0, \dots, n_{21} + \cdots + n_{2k}\ ,\ k = 2, \dots, K,\end{aligned} \tag{5.15}$$

and

$$\begin{aligned}H_1(m_1^{(1)}, m_2^{(1)}, \pi_1, \pi_2) = B(m_1^{(1)}, n_{11}, \pi_1)\ B(m_2^{(1)}, n_{21}, \pi_2)\ ,\\ m_1^{(1)} = 0, \dots, n_{11},\ m_2^{(1)} = 0, \dots, n_{21},\end{aligned}$$

recursively defines the function H_K. Since, from (5.14), the Type I error rate depends on the nuisance parameter $\pi = \pi_1 = \pi_2$, the actual size of the test is defined as

$$\sup_{\pi\in(0;\,1)} P_{\pi_1,\pi_2}(\text{reject } H_0)\ , \tag{5.16}$$

which can be found numerically. This technique enables the assessment of a group sequential test found from approximate considerations.

By fixing the total sample sizes N_j, $j = 1, 2$, it is possible to approximate the power and the ASN at given π_1 and π_2. Assume that the total sample size, N, is equal between the two treatment groups, i.e., $N = N_1 = N_2$. Using the same argument as in the one-sample case, the standardized effect ζ is estimated, in the two-sided case, by

$$\begin{aligned}\zeta =& \operatorname{sign}(\pi_2 - \pi_1)\\ &\times \frac{\Phi^{-1}(1-\alpha/2)\Big(\sqrt{\big(\pi_1(1-\pi_1) + \pi_2(1-\pi_2)\big)/2} - \sqrt{\bar\pi(1-\bar\pi)}\ \Big)\sqrt{2/N}}{\sqrt{\big(\pi_1(1-\pi_1) + \pi_2(1-\pi_2)\big)/2}}\\ &\quad + \frac{\pi_2 - \pi_1}{\sqrt{\big(\pi_1(1-\pi_1) + \pi_2(1-\pi_2)\big)/2}}\ ,\end{aligned} \tag{5.17}$$

and $\alpha/2$ replaced by α in the one-sided case. If N gets large,

$$\zeta \approx \frac{\pi_2 - \pi_1}{\sqrt{\big(\pi_1(1-\pi_1) + \pi_2(1-\pi_2)\big)/2}}\ , \tag{5.18}$$

which can be used as an approximation for large N, analogously to the one-sample case. Then, the multiple integral is calculated at given

$$\vartheta_k = \zeta \sqrt{\sum_{\tilde{k}=1}^{k} n_{\tilde{k}}/2}\ ,\ k = 1, \ldots, K.$$

We illustrate the approximate use of the standardized effect ζ in the two-sample binomial case and compare it to the exact values calculated with the use of (5.14). Table 5.3 contains the power for a two-sided four-stage group sequential design with Pocock boundaries ($\alpha = 0.05$) and some values π_1 and $\delta = \pi_2 - \pi_1$. Note that, although conceptually not different, the calculation of the exact values is much more time consuming in the two-sample case as compared to the one-sample case because of the double sum in (5.15) for each recursive step.

Table 5.3 Approximate power for testing rates in two independent samples using formula (5.17) and (5.18), respectively, and true power, calculated with the recursive formula (5.14), for different π_1, $\delta = \pi_2 - \pi_1$, and stage sample sizes $n_1 = \cdots = n_4 = n$ per treatment group in Pocock's four-stage group sequential design, $\alpha = 0.05$ (two-sided)

		Power			Power			Power		
π_1	δ	(5.18)	(5.17)	True	(5.18)	(5.17)	True	(5.18)	(5.17)	True
		$n = 3$			$n = 6$			$n = 10$		
0.05	0.0	0.050	0.050	0.001	0.050	0.050	0.004	0.050	0.050	0.006
	0.1	0.109	0.105	0.017	0.173	0.167	0.109	0.263	0.254	0.193
	0.2	0.239	0.216	0.107	0.436	0.406	0.385	0.654	0.625	0.620
	0.3	0.421	0.366	0.290	0.720	0.668	0.689	0.915	0.889	0.906
	0.4	0.633	0.541	0.520	0.913	0.868	0.893	0.992	0.984	0.989
	0.5	0.830	0.722	0.732	0.988	0.969	0.976	1.000	0.999	0.999
	0.6	0.956	0.877	0.884	1.000	0.997	0.997	1.000	1.000	1.000
0.10	0.0	0.050	0.050	0.005	0.050	0.050	0.021	0.050	0.050	0.025
	0.1	0.091	0.088	0.033	0.134	0.131	0.104	0.195	0.191	0.164
	0.2	0.195	0.180	0.125	0.352	0.330	0.321	0.544	0.520	0.516
	0.3	0.359	0.315	0.285	0.636	0.589	0.603	0.857	0.826	0.840
	0.4	0.569	0.487	0.489	0.869	0.818	0.834	0.981	0.969	0.972
	0.5	0.784	0.677	0.694	0.977	0.952	0.955	0.999	0.998	0.998
	0.6	0.939	0.851	0.858	0.999	0.995	0.993	1.000	1.000	1.000
0.40	0.0	0.050	0.050	0.064	0.050	0.050	0.056	0.050	0.050	0.046
	0.1	0.070	0.070	0.089	0.092	0.090	0.096	0.121	0.119	0.112
	0.2	0.138	0.131	0.160	0.235	0.223	0.225	0.365	0.351	0.343
	0.3	0.274	0.246	0.285	0.499	0.463	0.464	0.728	0.697	0.689
	0.4	0.499	0.429	0.467	0.807	0.754	0.758	0.959	0.941	0.935
	0.5	0.784	0.677	0.694	0.977	0.952	0.955	0.999	0.998	0.998

$\delta = 0$ refers to the Type I error rate

Table 5.3 shows that the use of the approximation (5.18) performs better than in the one-sample case although, again, (5.17) is preferable. Except for $n = 3$ and small δ, the power approximation using (5.17) is satisfactory and might be used for practical purposes. For small π_1, the Type I error rate is very small for all considered n. On the other hand, the actual size of the test defined by (5.16) can by found by a grid search. These are given by 0.0683, 0.0562, and 0.0484 for $n = 3$, 6, and 10, respectively. That is, the actual size of the test only slightly exceeds α for $n = 3$ and 6 (and might exceed α for $n > 10$), and hence the test is adequate. Clearly, if one can exclude values $\pi_1 > 0.10$, the decision regions can be modified in order to better exhaust the significance level. In all other cases, one must accept the conservatism of the test procedure for small nuisance parameter values π.

Confidence Intervals

When comparing two event rates in independent samples, the determination of confidence intervals for the difference $\pi_2 - \pi_1$ and for the relative risk π_2/π_1 might be of interest. One way to derive confidence intervals is to define appropriate test statistics $Z_k^* = Z_k^*(\delta_0)$ for $H_0 : \pi_2 - \pi_1 = \delta_0$ and $H_0 : \pi_2/\pi_1 = \delta_0$, respectively. Then, the confidence interval is given by solving the condition for $Z_k^*(\delta_0)$, valid in the group sequential context, for δ_0. Note that the corresponding tests also apply to non-inferiority designs where the non-inferiority margin is expressed in terms of the difference and the relative risk, respectively.

A number of approaches were proposed for defining a test statistic $Z_k^*(\delta_0)$ for testing the difference and the risk ratio, respectively. One of these is the application of the maximum likelihood technique for estimating the unknown rates, π_1 and π_2, under the respective null hypothesis. This technique was originally proposed by Koopman (1984) and Miettinen and Nurminen (1985). Explicit formulas for the maximum likelihood estimates were supplied by Farrington and Manning (1990). In a review article, Gart and Nam (1988) showed that this method is preferable compared to other techniques (see also, Newcombe 1998a,b, 2013). Generally, the confidence intervals can be found by iterative methods. Nonetheless, closed form analytical expressions for the lower and upper confidence limits of the confidence interval for the relative risk in single stage designs were found by Nam (1995). Interestingly, unlike for Fieller's intervals, there always exists a solution for the confidence interval, which is also the case for the group sequential situation.

Other approaches for calculating the confidence interval (and p-values) based on exact calculations are provided in Lin et al. (1991) and Coe and Tamhane (1993). These are briefly described in Jennison and Turnbull (2000), §12.2, where also some extensions of the binomial two-sample situation to case-control and stratified studies are discussed.

5.3 Survival Data

The analysis of survival data comprises a wide field of statistical techniques. It is concerned with the time elapsed from some fixed starting point to the occurrence of a particular event. In clinical trials, survival analyses are often conducted in oncological studies, where the event is the death of a patient or a relapse. As usual we use the term "survival" although the event needs not to be the death of the patient. The starting point is, for example, the surgery (i.e., the date of randomization in case of the comparison between two types of surgery) or the date of diagnosis. In the latter case, it might be of interest if different diagnoses have influence on the survival time. Typically, survival times are censored, which means that for some or many patients the survival time was not (yet) observed. This might be due to "loss to follow-up" or simply to the fact that the period of observation was at the end of the study before the event occurred. For these patients it is only known that the time is at least as long as the difference between the end date and the starting point date of that patient. The number of uncensored data increases with study duration, therefore these trials typically take a long time. Hence, it is desirable to plan interim looks to possibly shorten the trial providing a statistically significant result before the planned end of the trial.

Before we will be describing the group sequential test for survival data, we illustrate the pattern of patient entries within a survival trial. Typically, patients are recruited successively yielding different start times per patient. Figure 5.1 illustrates the different ways how ten (hypothetical) patients can enter and proceed through the study. It is assumed that the patients approximately enter the study at uniformly distributed time points (dashed bisecting line) though the third and the fourth patient

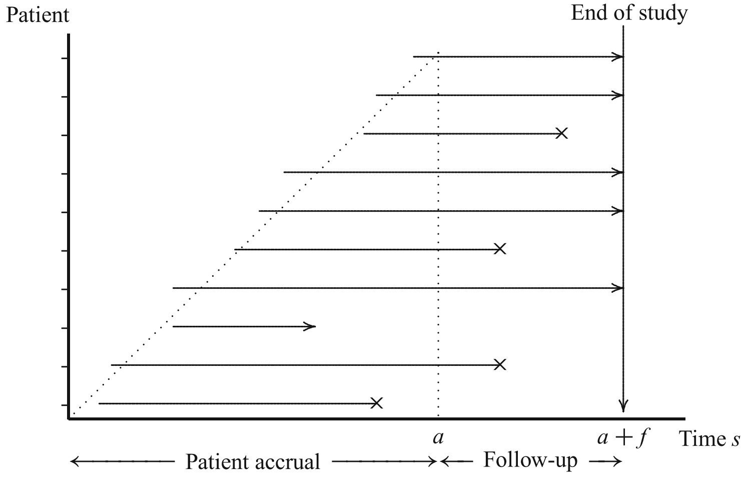

Fig. 5.1 Schematic diagram of patient entries at different times in a survival trial with censored observations. *Times*: occurrence of an event; *right arrow*: censored survival time

enter the study at the same time. The survival time of the third patient is right censored due to loss to follow-up, five survival times are right censored since patients are still alive (i.e., no event occurred) by the end of the trial.

A survival time X is a positive real-valued random variable with density function $f(x)$. The survival function is defined as

$$S(x) = P(X > x) = \int_x^{\infty} f(u)\,du\,,$$

and the hazard function is

$$\lambda(x) = \lim_{\varepsilon \to 0} \frac{P(x < X \leq x + \varepsilon \,|\, X > x)}{\varepsilon} = \frac{f(x)}{S(x)} = \frac{\partial \log(S(x))}{\partial x}\,.$$

The hazard function at time x specifies the instantaneous risk of the event, given the event did not occur up to time x. It follows that

$$S(x) = \exp\big(-\int_0^x \lambda(u)\,du\big)\,.$$

Hence, the hazard function determines the survival function (and vice versa), which clarifies the prominent role of the hazard function in survival analysis. Particularly, one can define characteristics for $\lambda(x)$ to obtain modeling assumptions for survival data. For example, a constant hazard function λ results in exponentially distributed survival times. In this case, the maximum likelihood estimate for λ is given by

$$\hat{\lambda} = \frac{d}{\sum_{i=1}^{d} x_i}\,,$$

if d events and hence d uncensored survival times $x_1, \ldots, x_d$ were observed. If an estimate $\hat{\pi}$ of the event rate in a period of length l is given, λ is estimated by

$$\hat{\lambda} = \frac{-\log(1 - \hat{\pi})}{l}\,. \tag{5.19}$$

One can also estimate λ from the (estimated) median survival time, $\hat{m}$, by

$$\hat{\lambda} = \frac{\log(2)}{\hat{m}}\,.$$

As another example, when comparing two treatment groups, a proportional hazards assumption declares that the hazard ratio

$$\omega(x) = \frac{\lambda_2(x)}{\lambda_1(x)}$$

is constant over time (i.e., $\omega(x) = \omega$).

Of particular interest in survival trials is the comparison of two survival functions. The hypothesis to be tested is $H_0 : S_1(x) = S_2(x)$ for all x. Under the proportional hazards assumption, this hypothesis is equivalent to $H_0 : \omega = 1$ and a rejection of H_0 can be interpreted in terms of the hazard ratio being smaller or greater than 1. The most commonly used test for H_0 is the log-rank test. Let n_j patients be randomized to treatment group $j, j = 1, 2$. Assume that only different survival times were observed, i.e., there were no ties. Let d be the number of occurred events, and let n_{1i}, n_{2i} be the number of patients at risk in treatment groups 1 and 2, respectively, when the ith event occurred. The test statistic of the log-rank test is then given by

$$LR = \frac{\sum_{i=1}^{d} \left(I_{2i} - \frac{n_{2i}}{n_{1i}+n_{2i}} \right)}{\sqrt{\sum_{i=1}^{d} \frac{n_{1i} n_{2i}}{(n_{1i}+n_{2i})^2}}} \,, \tag{5.20}$$

where $I_{2i} = 1$ if the ith event was in the treatment group 2 and $I_{2i} = 0$ otherwise. This is a standardized and one-sided version of the log-rank test statistic which is sensitive for rejecting H_0 against $H_1 : \omega > 1$. Note that there also exist quadratic forms of the test statistic LR, which are χ^2 distributed, but the test statistic (5.20) directly fits into the group sequential approach. This will be shown below. Note that the log-rank test is a non-parametric test since (5.20) does not depend on the observed survival times $x_1, \ldots, x_d$, it only depends on the number of patients who were at risk at the ith event.

For the moment, let d be fixed and let the patients be recruited until d events were observed. Then the log-rank test statistic is approximately normally distributed with

$$E(LR) = \sqrt{d}\, \frac{\sqrt{r}}{1+r} \log(\omega) \quad \text{and} \quad \mathrm{Var}(LR) = 1 \,, \tag{5.21}$$

where ω denotes the true hazard ratio, and $r = n_2/n_1$ is the allocation rate specifying the proportion of patients randomized to treatment group 1 and 2, respectively (Schoenfeld 1981).

Sample size calculations in a fixed sample size design can be performed as follows. The first step is to determine the required number of events. From (5.21) it follows that, in order to achieve power $1 - \beta$ for $H_1 : \omega = \omega_1$ at two-sided significance level α, the total number of events, d_f, required in a fixed sample size design is approximately given by

$$d_f = \frac{\left(\Phi^{-1}(1-\alpha/2) + \Phi^{-1}(1-\beta)\right)^2}{r/(1+r)^2 \left(\log(\omega_1)\right)^2} \,, \tag{5.22}$$

where $\alpha/2$ is replaced by α in the one-sided case.

The total sample size $n_{f1} + n_{f2} = n_{f1}(1 + r)$ required is d_f/ψ, where ψ denotes the combined probability of an event in the two treatment groups. There are several methods to estimate ψ under reasonable assumptions concerning the accrual time a

and follow-up time f of patients (see Fig. 5.1). For example, assume exponentially distributed survival times and assume that patients enter the trial uniformly over the recruitment period a. That is, at time $s \in (0; a + f)$ the time under observation, C, is uniformly distributed on the interval $(\max\{0, s - a\}; s)$ with density function $g(c) = 1/a$ if $c > 0$ and $P(C = 0) = 1 - s/a$. Hence, for $s < a$, the distribution of C is a mixture of a continuous and a discrete random variable. Given $\lambda_j, j = 1, 2$, the probability of an event is

$$\begin{aligned}\psi_{\lambda_j}(s) &= P_{\lambda_j}(X < C)\\ &= \int_{\max\{0,s-a\}}^{s} P_{\lambda_j}(X < c \,|\, C = c)\, g(c)\, dc\\ &= \int_{\max\{0,s-a\}}^{s} \big(1 - \exp(-\lambda_j c)\big)\, \frac{1}{a}\, dc\,.\end{aligned}$$

Simple integration calculus yields

$$\psi_{\lambda_j}(s) = \begin{cases} \dfrac{s}{a} - \dfrac{1 - \exp(-\lambda_j s)}{\lambda_j a} & \text{if } s \le a \\[2ex] 1 - \exp(-\lambda_j s)\, \dfrac{\exp(\lambda_j a) - 1}{\lambda_j a} & \text{if } s > a\,. \end{cases} \tag{5.23}$$

Particularly, by the end of the trial,

$$\psi_{\lambda_j}(a + f) = 1 - \exp(-\lambda_j (a + f))\, \frac{\exp(\lambda_j a) - 1}{\lambda_j a}\,.$$

The combined probability of an event under H_1 is

$$\psi = \psi_{(1)}(a + f) = \big(\psi_{\lambda_1}(a + f) + r\, \psi_{\lambda_2}(a + f)\big)/(1 + r)\,,$$

where $\lambda_j, j = 1, 2$, are the assumed hazards under H_1 (see Schumacher and Schulgen 2002). So the sample size for the first treatment group is given by

$$n_{f1} = \frac{\big(\Phi^{-1}(1 - \alpha/2) + \Phi^{-1}(1 - \beta)\big)^2}{\psi\ r/(1 + r)\,\big(\log(\omega_1)\big)^2}\,, \tag{5.24}$$

and the sample size for the second treatment group is $n_{f2} = r\, n_{f1}$.

Before we proceed to the group sequential setting, we illustrate the sample size calculation in the fixed design by an example. Suppose in a trial with exponentially distributed survival times it is assumed that the probabilities of an event after 12 months are 30 % and 50 % in treatment group 1 and 2, respectively. From (5.19), this translates to hazards $\lambda_1 = 0.0297$, $\lambda_2 = 0.0578$, and to a hazard ratio $\omega_1 = 1.943$. The required total number of events to achieve power 80 % at one-sided level

$\alpha = 0.025$ is $4 \times 2.802^2/\big(\log(1.943)\big)^2 = 71.1$. Assume now that the accrual period is 6 months and the follow-up time is 3 months in a balanced design. That is, $r = 1, a = 6$ and $f = 3$, yielding $\psi_{0.0297}(9) = 0.162$, $\psi_{0.0578}(9) = 0.289$, and $\psi = \psi_{(1)}(9) = 0.226$. From (5.24), the required sample size per treatment group is 157.3, and hence a total of $2 \times 158 = 316$ patients are required. Of course, with increasing length $a + f$ of the study, the required total sample size decreases and, given any a, it converges to d.

In the group sequential setting, it is planned to recruit a maximum number N_j of patients in treatment group j, $j = 1, 2$, yielding a total maximum sample size $N = N_1 + N_2$. At observation times $s_k \in (0;\ a+f)$, $k = 1, \ldots, K$, interim analyses are conducted. These analyses can take place in the accrual period (i.e., $s_k \in (0;\ a)$) or in the follow-up period (i.e., $s_k \in (a;\ a+f)$) (see Fig. 5.1). In the latter case, the maximum sample size is already reached, hence the stopping rule of a group sequential design can only reduce the duration of the trial. At each interim analysis, a specific number of events is recorded. Specifically, during the stages of the trial, a sequence of *accumulated* events $d_1, \ldots, d_K$ can be observed. At each stage k of the test procedure the log-rank test statistic

$$LR_k = \frac{\sum_{i=1}^{d_k} \big(I_{2ik} - \frac{N_{2ik}}{N_{1ik}+N_{2ik}}\big)}{\sqrt{\sum_{i=1}^{d_k} \frac{N_{1ik}N_{2ik}}{(N_{1ik}+N_{2ik})^2}}}\ ,\ k = 1, \ldots, K,$$

is calculated, where N_{1ik} and N_{2ik} are the number of patients at risk at stage k in treatment groups 1 and 2, respectively, when the ith event occurred. $I_{2ik} = 1$ if the ith event until the end of stage k was observed in treatment group 2 and $I_{2ik} = 0$ otherwise. This is simply to calculate the log-rank test statistic with all data available at stage k. Approximately, for fixed d_k, LR_k has unit variance and

$$E(LR_k) = \sqrt{d_k}\,\frac{\sqrt{r}}{1+r}\,\log(\omega)\ ,\ k = 1, \ldots, K, \tag{5.25}$$

where $r = N_2/N_1$ is the allocation rate. Most notably, it was shown by several authors that the sequence of test statistics $LR_1, \ldots, LR_K$ approximately has the independent and normally distributed increments structure (Jones and Whitehead 1979; Sellke and Siegmund 1982; Tsiatis 1981, 1982; Olschewski and Schumacher 1986). Therefore, the group sequential test designs described in Chaps. 2 and 3 can be applied in the usual way.

Reconsider the example from above to illustrate this. Suppose it is planned to conduct a four-stage O'Brien and Fleming design at one-sided significance level $\alpha = 0.025$ yielding critical values $u_k = 4.0486/\sqrt{k}$, $k = 1, \ldots, 4$ (see Table 2.1). Hence, the test stops with the rejection of H_0 at stage k if

$$LR_k > u_k\ ,\ k = 1, \ldots, 4\,.$$

For power $1 - \beta = 0.80$, the inflation factor I is 1.024 (see Table 2.3). As for the fixed sample size case, the relevant information is contained in the number of events which is the "effective sample size" preserving the power. Since $1.024 \times 71.1 = 72.8$, with the assumptions from the above example a total of $d = d_K = 73$ events are required by the end of the study to fulfill the power requirement. Translating this to the maximum number of patients, a total of $N = 324$ patients are needed by the end of the trial since $1.024 \times 2 \times 157.3 = 322.1$. Using the stopping criteria of the group sequential design, the expected number of events under H_1 is $0.831 \times 71.1 = 59.1$.

This test has Type I error rate close to 0.025 and power close to 80 % if the interim analyses take place after equally spaced information rates, and if 73 events occurred by the end of the study. The information levels, however, are formulated in terms of the observed events rather than in terms of the recruited patients. It is a valid test if the information rates are given by

$$t_k = \frac{k}{K} d \ , \ k = 1, \ldots, 4 \, ,$$

since then the correlation between LR_k and $LR_{k'}$ $(k < k')$ is $\sqrt{k/k'}$. Otherwise, the correlation is $\sqrt{d_k/d_{k'}}$ and this does not ensure that the test keeps the level nor that it has the desired power.

If the interim analyses take place after equally spaced calendar times, it will hardly be the case that the information rates are equally spaced. We will see below that it is possible to choose analysis times and patient recruitment such that it can be expected to produce the desired information rates. But also in this case, in a concrete application the actually observed events need not to produce the desired information rates. An elegant and attractive solution is the use of the α-spending approach (see Sect. 3.3). The information rates are determined during the stages of the trial yielding a specific sequence of critical values. This test keeps the Type I error rate, given the observed pattern of events. We note that it is also possible to define the α-spending approach in terms of calendar time. For a discussion of these conceptually different approaches, we refer to DeMets and Gail (1985), Lan and DeMets (1989b), Lan and Lachin (1990), and Kim et al. (1995).

In the following, we briefly present a strategy to estimate the observation times such that the information rates are equal to a sequence of specified information rates $t_1, \ldots, t_K$. We thereby closely follow the lines of Kim and Tsiatis (1990) who provide a unified design procedure for exponentially distributed survival response and uniform patient entry. Under these assumptions, given the maximum sample size N of a group sequential design, at time s the expected number of events, under H_1, is

$$N \ \psi_{(1)}(s) = N \ \left(\psi_{\lambda_1}(s) + r \, \psi_{\lambda_2}(s)\right)/(1 + r) \ ,$$

where $\psi_{\lambda_j}(s)$ is defined by (5.23). Having fixed a, by the end of the trial the expected number of events should be equal to number of events d required by the power condition. That is,

$$N\,\psi_{(1)}(a+f)=d\;.$$

So the follow-up time f can be determined by

$$f=\psi_{(1)}^{-1}(d/N)-a\;, \tag{5.26}$$

where $\psi_{(1)}^{-1}(\cdot)$ denotes the inverse of the function $\psi_{(1)}(\cdot)$.

Due to the linearity of the information rates to the number of events, the real (calendar) times s_k, $k=1,\ldots,K$, when the kth analyses should take place are, under H_1, given by

$$s_k=\psi_{(1)}^{-1}\big((t_k\,\psi_{(1)}(a+f)\big)\;. \tag{5.27}$$

Together with the stopping probabilities of a specific design, this enables the calculation of the expected study duration. Finally, the expected patient accrual at calendar time s_k, $k=1,\ldots,K$, is given by

$$a(s_k)=\begin{cases}\dfrac{N}{a}\,s_k & \text{if } s_k<a\\ N & \text{if } s_k\geq a\;,\end{cases}$$

and $a(s_k)$ together with the stopping probabilities of a specific design can be used to calculate the expected number of patients.

To illustrate this, we again reconsider the example. It was found that 73 events are necessary by the end of the trial to achieve power 80 %. With accrual time $a=6$ and follow-up time $f=3$, the maximum necessary sample size was $N=324$. Equation (5.27) yields the observation (calendar) times $s_1=4.1$, $s_2=5.8$, $s_3=7.3$, and $s_4=9$. This sequence results in the expected number of events $324\,\psi_{(1)}(4.1)=18.2$, $324\,\psi_{(1)}(5.8)=36.4$, $324\,\psi_{(1)}(7.3)=54.6$, and $324\,\psi_{(1)}(9)=72.8$, which indeed corresponds with equally spaced information rates. The expected study duration, under H_1, is found to be 7.8. Furthermore, from the stopping probabilities of the chosen design and the sample sizes $a(s_1)=222$, $a(s_2)=314$, $a(s_3)=324$, and $a(s_4)=324$, one finds the expected number of patients to be 320.2. This sample size is only slightly smaller than the maximum sample size $N=324$.

Suppose now that it is desired to recruit only a maximum of 2×100 patients in the trial. To achieve power 80 %, from (5.26) one finds that the follow-up time should be $f=7.7$. Equation (5.27) now yields the observation times $s_1=5.2$, $s_2=7.7$, $s_3=10.5$, and $s_4=13.7$. Due to the later analyses times, the maximum sample size, 200, is already reached at stage 2. Nevertheless, although the expected number of patients under H_1 is 199.9 and hence lower than in the first case, the expected study duration, under H_1, is found to be 11.4 and hence substantially larger than in the first case.

Table 5.4 Approximate power for the log-rank test using the observation times defined by (5.27) with $a = 6$ and $f = 3$, and true power, estimated by simulation (one million replicates), for different π_1, π_2 at time $l = 12$, hazard ratio $\omega = \log(1-\pi_2)/\log(1-\pi_1)$, and maximum sample sizes N in an O'Brien and Fleming four-stage group sequential design, $\alpha = 0.025$ (one-sided)

			Power		Power		Power	
π_1	π_2	ω	Approximate	True	Approximate	True	Approximate	True
			$N = 2 \times 20$		$N = 2 \times 40$		$N = 2 \times 100$	
0.10	0.10	1.000	0.025	0.006	0.025	0.019	0.025	0.021
	0.20	2.118	0.095	0.055	0.150	0.120	0.309	0.279
	0.30	3.385	0.236	0.172	0.419	0.334	0.795	0.733
	0.40	4.848	0.446	0.342	0.734	0.615	0.984	0.960
	0.50	6.579	0.680	0.539	0.931	0.844	1.000	0.998
0.30	0.30	1.000	0.025	0.024	0.025	0.023	0.025	0.024
	0.40	1.432	0.071	0.066	0.103	0.095	0.194	0.187
	0.50	1.943	0.164	0.150	0.284	0.264	0.596	0.578
	0.60	2.569	0.324	0.292	0.568	0.532	0.922	0.908
	0.70	3.376	0.552	0.497	0.842	0.805	0.997	0.995
0.50	0.50	1.000	0.025	0.024	0.025	0.023	0.025	0.024
	0.60	1.322	0.071	0.067	0.104	0.099	0.197	0.192
	0.70	1.737	0.179	0.168	0.313	0.302	0.646	0.637
	0.80	2.322	0.394	0.371	0.669	0.651	0.967	0.964
	0.90	3.322	0.736	0.707	0.957	0.949	1.000	1.000

$\omega = 1$ refers to the Type I error rate

Finally suppose that a maximum of 2 × 100 patients should be observed with a maximum follow-up of $f = 3$. The calculation of the test characteristics is straightforward. Clearly, the power is reduced. Using $\psi_{(1)}(9) = 0.226$, the expected number of events at the final analysis is $2 \times 100 \times 0.226 = 45.2$. Consequently, equally spaced information levels $d_k = k/4 \times 45.2$, $k = 1, \ldots, 4$, can be used in (5.25) to find the standardized effect size for computing the power which, in this case, is 59.6 %.

In Table 5.4 we illustrate the accuracy of the power calculation at given maximum sample size, N, for different N and some values π_1 and π_2 in the one-sided O'Brien and Fleming design. Depending on π_1 and π_2, the choice of $a = 6$ and $f = 3$ results in different patterns of observation times. The true power is estimated by simulation with uniform patient entry, using one million replicates (standard error smaller than 0.0005).

Table 5.4 shows that, except for very small event rates and sample sizes, the power calculation using the recursive integration formula is quite accurate. Note that for small event rates the (maximum) number of observed events is small, too, and hence the "effective sample size" is too small to produce a satisfactory behavior of the log-rank test. There is a trend towards overestimating the true power but, for suitably large event rates, this effect is not dramatic. We also realize that in these cases the Type I error rate is almost accurately equal to the desired level. This is even more impressive since, as before, the number of events is considerably smaller than N.

By varying the design, the information rates, the time of accrual, and the follow-up observation time one finds different operating characteristics of the group sequential design. Kim and Tsiatis (1990) gave illustrative examples and show how this might help to select an appropriate design. However, it is difficult to find optimizing strategies. This becomes even more important as these considerations do not account for losses to follow-up or competing risks.

Extensions

There is a huge literature on sample size determination in survival studies for the fixed sample size situation. For a comprehensive overview, we refer to Oellrich et al. (1997). Much work is dealing with special survival distributions, effects of losses to follow-up, other patient accrual patterns, and so on. When using the log-rank test, many techniques for sample size calculation in the fixed sample size situation can also be applied to the group sequential setting. There are situations, however, where the log-rank test should not be used and other tests for testing the equality of survival curves are perhaps better suited. For example, if differences between survival curves are expected to be more pronounced at the beginning of the observation time, a weighted log-rank statistic with higher weights at early time points might be appropriate. For example, the G^ρ family due to Harrington and Fleming (1982) places more weight on early time points if $\rho > 0$. Some caution is necessary as Slud and Wei (1982) showed that such weighted test statistics possibly fail to possess the independent increments structure when patient entry is staggered.

An important issue in survival trials is the stratified analysis. In the fixed sample size design, for stratified analyses one can use the Mantel–Haenszel version of the log-rank test or a generalization of it. More generally, the proportional hazards regression model (Cox 1972) enables taking into account several factors including quantitative covariables. It was shown in a very general sense that the sequence of test statistics or the sequence of estimates obtained from this model can be embedded in the independent and normally distributed increments structure (for example, Tsiatis et al. 1985, 1995; Jennison and Turnbull 1997b; Scharfstein et al. 1997). This shows that the theory of group sequential tests derived for the prototype case can be applied to a wide range of testing situations. Particularly, it also enables the determination of estimation procedures for the parameters of interest. For some further details, the reader is referred to Chap. 13 in Jennison and Turnbull (2000) and the references therein.

Part II
Confirmatory Adaptive Designs with a Single Hypothesis

Chapter 6
Adaptive Group Sequential Tests

The calculation of the sample size in clinical trials requires the specification of the treatment effect for which the study is powered for. This treatment effect must be a realistic projection of the treatment's efficiency in order to avoid an underpowered study. At the same time it must correspond to a clinically relevant change in the primary endpoint such that overly large sample sizes and statistically significant but clinically irrelevant study results are avoided. The treatment for the sample size calculation is often specified from previous clinical trials or pilot studies. If the prior experience with the study treatments is insufficient for a persistent pre-specification of relevant and realistic treatment effects, the effect and corresponding sample sizes may be reassessed at an interim analysis from unblinded interim data. In this case designs are required that guarantee Type I error rate control even though the sample size is adjusted based on unblinded interim treatment effect estimates. Bauer (1989), Bauer and Köhne (1994), on the one hand, and Proschan and Hunsberger (1995), on the other hand, have independently suggested designs that control Type I error rates after such data-driven types of sample size adjustments.

A major advantage of these methods is that any sample size adaptation rule can be used and no specific rule needs to be assumed for Type I error control. This permits an update of the relevant and realistic treatment effects in an arbitrarily complex manner and not only based on the interim data, but also on emerging new external information and expert knowledge. This is particularly valuable in long term clinical trials. Additionally, the sample size can also be adjusted to meet other issues than the power of the primary hypothesis test. For instance, one can increase the sample size to obtain more information on important secondary efficacy or safety endpoints. Such issue may have not been foreseen and could come up only while reviewing the unblinded interim data in an independent Data Monitoring Committee (iDMC) (see Sect. 11.4).

G. Wassmer, W. Brannath, *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*, Springer Series in Pharmaceutical Statistics,
DOI 10.1007/978-3-319-32562-0_6

We note that *blinded* sample size recalculation is an important topic in adaptive sample size recalculation, too. It was introduced as the "internal pilot design" (Wittes and Brittain 1990; Birkett and Day 1994). The blinded sample size reassessment design determines the sample size of the second stage using only the estimate of nuisance parameters such as the variance, the overall response rate, or the overall survival pattern (see, for example, Gould 1992, 1995; Kieser and Friede 2000). For a review of these methods, see Friede and Kieser (2006). No unblinding of the data is necessary, and no effect size is calculated. Hence, no early assessment of efficacy is possible which makes this design conceptually different from an adaptive extension of group sequential designs. Although these designs are clearly considered as adaptive, we do not consider them in this monograph.

The current chapter is devoted to adaptive designs as suggested by these pioneering papers as well as newer developments and investigations. We will focus in this chapter on two-stage adaptive designs with an adaptation of the sample size, although the method allows for other types of adaptations, such as changing test statistics or changing hypotheses, as well as for the multi-stage generalization. Such types of adaptations and the multi-stage case will be considered later in this book. Nevertheless, this chapter provides the foundation of all the remaining chapters inclusive Part III where adaptive designs with multiple hypotheses will be discussed.

6.1 Basic Principle and Assumptions

The two approaches introduced by Bauer (1989) and Bauer and Köhne (1994), on the one hand, and Proschan and Hunsberger (1995), on the other hand, are often denoted as *combination test* and *conditional error function approach*, respectively. Many of the later approaches follow or extend these approaches. Although different in their appearance, combination tests and the conditional error function approach are based on a common principle which is called the *conditional invariance principle* (see, for example, Brannath et al. 2007).

The conditional invariance principle is as follows. Assume that we want to test a null hypothesis H_0, for example, non-superiority of an experimental treatment to a control. Think of a trial with two sequential stages, where design characteristics of the second stage are chosen at an interim analysis based on the data from the first stage as well as external information. The design of the first stage is pre-fixed and remains unaltered. Assume further that the first and second stage data are from independent cohorts of patients. If the trial continues to the second stage, let T_2 be the statistics for H_0 from the second cohort recruited after the interim analysis. Due to the data-driven choice of design features, the null distribution of T_2 will in general depend on the interim data. However, we can often transform T_2 in a way that the conditional null distribution of T_2 given the interim data and the second stage design equals a fixed pre-specified null distribution, and hence is *invariant* with respect to the interim data and mid-trial design adaptations.

An invariant conditional distribution is typically achieved by transforming T_2 to a p-value p_2 which is uniformly distributed under H_0 conditionally on the interim data and second stage design. If T_2 is normally distributed with mean 0 and known standard error σ_{T_2}, invariance of the conditional distribution can also be achieved by standardizing T_2 to $Z_2 = T_2/\sigma_{T_2}$ which is always standard normal.

The invariance of the conditional null distribution of p_2 or Z_2 implies that they are stochastically independent from the interim data. Since the null distribution of the interim data is usually known (as the first stage design is fixed), the common distribution of the interim data and p_2 is known and invariant with respect to the (often unknown) mid-trial adaptation rule. Hence, we can specify a level α rejection region in terms of the interim data and invariant second stage test statistic p_2. This gives a test with Type I error rate α independently from the adaptation rule. For a mathematical rigorous verification of the conditional invariance principle, see Brannath et al. (2012) (see also Hommel 2001; Liu et al. 2002). The combination test and conditional error function approach just differ in the way how the invariant rejection region is specified.

In most parts of the book we will assume as above that the second stage test statistic T_2 is computed from a cohort of patients that is independent from the first stage cohort whose observations are used at the interim analysis for the design adaptations. We will discuss the relaxation of this assumption in later chapters.

6.2 Combination Tests

With a combination test we combine a p-value from the first and a p-value from the second stage by a pre-specified combination function. This method, which is a convenient way to implement the conditional invariance principle, will be the topic of the following sections.

6.2.1 General Methodology

As for the conditional invariance principle, we consider a null hypothesis H_0 which is tested in a trial with two stages where the data in the first and second stage are from independent cohorts of patients. Unless otherwise noted, H_0 is assumed to be one-sided. Note that we do not make any distributional assumptions nor assumptions on the type of design. That is, we can apply this method for testing means, rates, hazard ratios, etc., in a parallel group design, a repeated measures designs, or whatever. We will also be able to apply this method if nuisance parameters (for example, an unknown variance) are present.

Let p_1 and p_2 denote the one-sided p-values for H_0 computed from the first and second stage cohort, respectively. A two-stage combination test is defined by a combination function $C(p_1, p_2)$ which is non-decreasing in both arguments and continuous in p_2, boundaries α_1 and α_0 for early stopping, and a critical value c for the final analysis. The trial is stopped in the interim analysis if either $p_1 \leq \alpha_1$ or $p_1 > \alpha_0$, whereby H_0 is rejected if $p_1 \leq \alpha_1$ and retained if $p_1 > \alpha_0$. This is similar to a group sequential design where the null hypothesis can be rejected or retained at the interim analysis as well. Retainment of H_0 at the first stage is often denoted by *stopping for futility* (see Sect. 2.3). The trial proceeds to the second stage if $\alpha_1 < p_1 \leq \alpha_0$. In this case the null hypothesis H_0 is rejected at the end of the second stage if $C(p_1, p_2) \leq c$. Figure 6.1 provides a schematic illustration of combination tests. Note that $\alpha_1 = 0$ implies that the trial is never stopped with a rejection of H_0 and $\alpha_0 = 1$ means that no stopping for futility is foreseen.

Usually p_1 is uniformly distributed under H_0 and p_2 has the same conditional distribution for given stage 1 data and second stage trial design. Since this is true for any interim data under H_0, p_1 and p_2 are independent and uniformly distributed. Therefore, in order to achieve Type I error control the boundaries $\alpha_1 < \alpha_0$ and c must satisfy the condition

$$\alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_0^1 \mathbf{1}_{\{C(p_1,p_2) \leq c\}} dp_2 \, dp_1 = \alpha \ , \tag{6.1}$$

where the indicator function $\mathbf{1}_{\{\cdot\}}$ equals 1 if $C(p_1, p_2) \leq c$ and 0 otherwise.

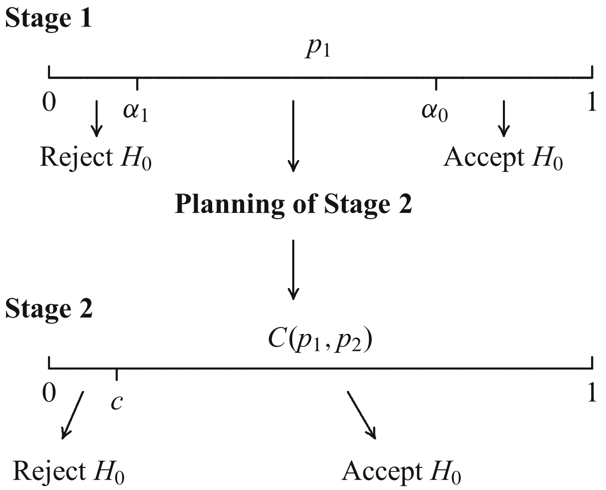

Fig. 6.1 Two-stage combination tests. For the planning, fix first stage sample sizes, test, α_1, α_0, and the combination function $C(p_1, p_2)$ with critical value c. After stage 1, compute the p-value p_1 from the stage 1 data. Then, either stop or fix the design for stage 2 based on data from stage 1. After stage 2, compute the p-value p_2 from the stage 2 data and reject H_0 if $C(p_1, p_2) \leq c$

According to the invariance principle the p-values p_1 and p_2 are independent and uniformly distributed irrespective of the sample size adaptation, and so the combination test has Type I error rate α for all sample size adaptation rules.

The Type I error control remains true if under H_0 the distribution of p_1 is (unconditionally) stochastically larger than the uniform distribution, and the conditional distribution of the second stage p-value p_2 given p_1 is stochastically larger than or equal to the uniform distribution as well. More formally, if P_{H_0} denotes the probability distribution of p_1 and p_2 under H_0, then

$$P_{H_0}(p_1 \leq u) \leq u \quad \text{and} \quad P_{H_0}(p_2 \leq u | p_1 = v) \leq u \quad \text{for all } 0 \leq u, v \leq 1 \qquad (6.2)$$

is sufficient for Type I error rate control of the adaptive two-stage design. We will denote this property *p-clud* (Brannath et al. 2002). Property (6.2) is satisfied if, for example, the stage-wise cohorts are independent and conservative tests are used for p_k at each stage k, $k = 1, 2$. A mathematical rigorous discussion of the p-clud condition and a prove of Type I error control with the p-clud property can be found in Brannath et al. (2012).

At the planning stage of an adaptive clinical trial one has to specify the combination function, the decision boundaries α_1, α_0, and c, and the design for the first stage, including the first stage sample size and first stage test statistic. The second stage design does not have to be specified in advance, but it must at the latest be specified at the interim analysis.

6.2.2 Fisher's Product Test

In Bauer (1989) and Bauer and Köhne (1994), Fisher's product test (Fisher 1932) was proposed as combination test for use in adaptive designs. With this test the combination function is $C(p_1, p_2) = p_1 p_2$. The choice of this type of combination function was motivated by meta-analysis where p-values from different studies are combined to an overall significance test (see, for example, Hedges and Olkin 1985; Sonnemann 1991).

Combining significance tests has been a research topic already in the 1930s. Since p-values from independent studies are independent and uniformly distributed under the null hypothesis the distribution of their product is easily determined. For a uniformly distributed p-value U, $-2\log(U)$ has the exponential distribution with rate parameter $\lambda = 1/2$ which corresponds to the χ^2-distribution with two degrees of freedom. Since the sum of two independent χ^2-distributed random variables is χ^2-distributed as well, we obtain that

$$-2\log(p_1 p_2) = -2\log(p_1) - 2\log(p_2)$$

is χ^2-distributed with 4 degrees of freedom. Hence, the rejection rule is

$$-2\log(p_1 p_2) \geq \chi^2_{4,1-\alpha} ,$$

where $\chi^2_{4,1-\alpha}$ denotes the $(1-\alpha)$-quantile of the χ^2 distribution with 4 degrees of freedom. Equivalently, the rejection rule

$$p_1 p_2 \leq c_\alpha \text{ with } c_\alpha = \exp(-\chi^2_{4,1-\alpha}/2) \tag{6.3}$$

provides a level α-test. Fisher's product test is also named the "inverse chi-squared method" (Hartung 2001; Hartung and Knapp 2003).

To understand an important feature of Fisher's product test let us consider an interim sample point with $p_1 \leq c_\alpha$. Then $p_1 p_2 \leq c_\alpha$ for any $0 \leq p_2 \leq 1$. Hence, we will reject H_0 at stage 2 with any second stage p-value p_2 and therefore can stop the trial and reject H_0 already at stage 1. This shows that Fisher's product test has a built-in early rejection boundary, namely $\alpha_1 = c_\alpha$. This phenomenon is denoted as *non-stochastic curtailment* because the stopping rule is based on non-stochastic reasoning.

Generally, with Fisher's product combination function and $c \leq \alpha_1 < \alpha_0$ the level condition (6.1) becomes

$$\alpha_1 + c \cdot \big(\log(\alpha_0) - \log(\alpha_1)\big) = \alpha . \tag{6.4}$$

This follows from

$$\int_{\alpha_1}^{\alpha_0} \int_0^1 \mathbf{1}_{\{p_1 p_2 \leq c\}} dp_2\, dp_1 = \int_{\alpha_1}^{\alpha_0} (c/p_1) dp_1 = c \cdot \big(\log(\alpha_0) - \log(\alpha_1)\big) .$$

Hence, any triple set of values α_1, α_0, and c that fulfills (6.4) can be used within the two-stage adaptive approach and provides a valid level-α test.

For example, if $c = c_\alpha = 0.0087$ and $\alpha_0 = 0.50$ is chosen, a simple numerical root finding yields $\alpha_1 = 0.0233$. Figure 6.2 illustrates the rejection region of Fisher's product test for this case. The shaded area is equal to α and the determination of α_1 can be geometrically interpreted as shifting the area under the curve where $p_1 > \alpha_0$ to the top area under the curve where $p_1 \leq \alpha_1$. The area which is defined through the condition "$p_1 \leq \alpha_1$" can be the more enlarged the larger the area defined through "$p_1 > \alpha_0$" is.

Several ways to choose the boundaries α_0, α_1, and c have been suggested in the literature. Bauer and Köhne (1994) considered at first the case where no futility stopping is foreseen ($\alpha_0 = 1$) and $\alpha_1 = c_\alpha$ as defined in (6.3). This corresponds to using the full level α for the second stage critical value while accounting for the non-stochastic curtailment phenomenon. Bauer and Köhne (1994) also considered designs with $c = c_\alpha$ and a pre-planned futility stopping rule α_0 ($c_\alpha < \alpha_0 < 1$). $\alpha_1 > c_\alpha$ is found by a numerical search such that level condition (6.4) is met for the chosen

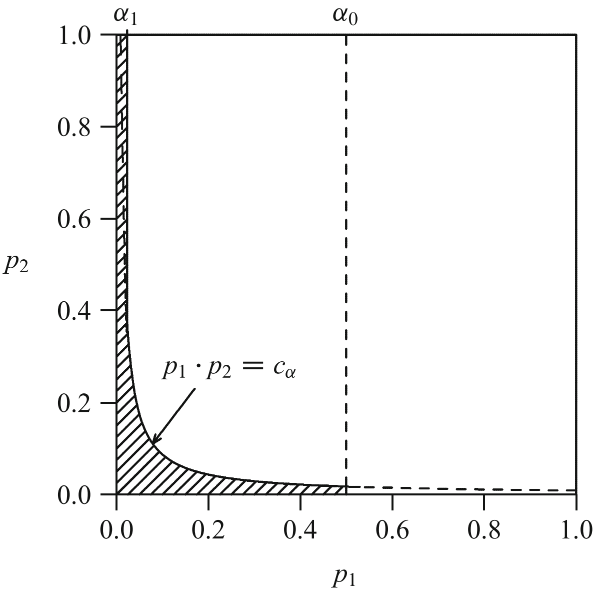

Fig. 6.2 Rejection region (*dashed region*) for the two-stage design of Bauer and Köhne; $\alpha = 0.05$, $\alpha_0 = 0.50$, $\alpha_1 = 0.0233$

Table 6.1 First stage rejection level α_1 at different significant levels α with different futility boundaries $\alpha_0 = 0.1, 0.15, \ldots, 1.00$ and the second stage critical value $c_\alpha = \exp(-\chi^2_{4,1-\alpha}/2)$ of Fisher's product test

	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	l$\alpha = 0.005$
α_0	$c_\alpha = 0.00870$	$c_\alpha = 0.00380$	$c_\alpha = 0.00131$	$c_\alpha = 0.00059$
0.10	0.0426	0.0186	0.0064	0.0029
0.15	0.0381	0.0166	0.0057	0.0026
0.20	0.0348	0.0152	0.0052	0.0024
0.25	0.0321	0.0140	0.0048	0.0022
0.30	0.0299	0.0131	0.0045	0.0020
0.40	0.0263	0.0115	0.0040	0.0018
0.50	0.0233	0.0102	0.0035	0.0016
0.60	0.0207	0.0090	0.0031	0.0014
0.70	0.0183	0.0080	0.0027	0.0012
0.80	0.0159	0.0069	0.0024	0.0011
0.90	0.0133	0.0058	0.0020	0.0009
1.00	0.0087	0.0038	0.0013	0.0006

The row $\alpha_0 = 1.0$ corresponds to the non-stochastic curtailment case

α_0. In Table 6.1 the critical values α_1 are provided for $\alpha_0 = 0.1, 0.15, \ldots, 1.00$ and several α.

We illustrate the method with a numerical example. Consider Fisher's product test at level $\alpha = 0.025$ with $c = c_\alpha = 0.00380$ and futility boundary $\alpha_0 = 0.7$. From Table 6.1 we get the first stage rejection boundary $\alpha_1 = 0.0080$. Assume now that we observe at the first stage the p-value $p_1 = 0.015$. Then $\alpha_1 < p_1 \leq \alpha_0$ which means that we cannot reject H_0 at the interim analysis. As a consequence, we will continue the trial to the second stage. Assuming that we observe the p-value $p_2 = 0.02$ at the second stage we get for the product criterion $p_1 \cdot p_2 = 0.003$ which is below the critical value $c_\alpha = 0.00380$ and hence implies rejection of H_0.

Another choice of boundaries was suggested by Bauer and Röhmel (1995). They considered only designs without futility stopping ($\alpha_0 = 1$) but where α_1 is manually set to a larger value than the non-stochastic curtailment minimum c_α. The corresponding second stage critical value c which follows from level condition (6.4) is necessarily smaller than c_α but can easily be determined by the formula $c = -(\alpha - \alpha_1)/\log(\alpha_1)$. It is straightforward to extend this method to cases with $\alpha_0 < 1$. In these cases, level condition (6.4) leads to the second stage boundary $c = d_{\alpha,\alpha_0,\alpha_1}$ where

$$d_{\alpha,\alpha_0,\alpha_1} = (\alpha - \alpha_1)/\big(\log(\alpha_0) - \log(\alpha_1)\big) .$$

Simple algebra shows that the non-stochastic curtailment property $\alpha_1 \geq d_{\alpha,\alpha_0,\alpha_1}$ requires the constraint $\alpha_1 + \alpha_1\{\log(\alpha_0) - \log(\alpha_1)\} \geq \alpha$ for α, α_0, and α_1. As an example, assume again $\alpha = 0.025$ and $\alpha_0 = 0.7$. In order to increase the chance for a rejection at the interim analysis we choose now the larger $\alpha_1 = 0.01$. The corresponding second stage rejection boundary is then $d_{0.025,0.7,0.01} = (0.025 - 0.01)/(\log(0.7) - \log(0.01)) = 0.015/4.25 = 0.0035$.

It is convenient to consider the second stage local level, α_2, of Fisher's product test, i.e., the level α_2 defined by the identity $c_{\alpha_2} = d_{\alpha,\alpha_0,\alpha_1}$. This level is given by $\alpha_2 = 1 - F_{\chi^2_4}(-2\log(d_{\alpha,\alpha_0,\alpha_1}))$ where $F_{\chi^2_4}(\cdot)$ is the cumulative distribution function of the χ^2-distribution with 4 degrees of freedom. For instance, the boundary $d_{0.025,0.7,0.01} = 0.0035$ corresponds to the second stage level $\alpha_2 = 1 - F_{\chi^2_4}(-2\log(0.0035)) = 0.023$.

Following this idea, Bauer and Röhmel (1995) and Wassmer (1999b) considered another choice for α_1 and α_2 for pre-specified α and α_0. In analogy to Pocock's design for group sequential trials (see Chap. 2), they suggested to use the same local level at both stages $\alpha_1 = \alpha_2$. Accordingly, they define $\alpha_1 = \alpha^*$ and $c = c_{\alpha^*} = \exp(-\chi^2_{4,1-\alpha^*}/2)$ where α^* is determined by numerical root finding from level condition (6.4) for the given α and α_0. Some values of α^* for different α and α_0 can be found in Table 6.2.

To illustrate this method with an example we consider again a design at level $\alpha = 0.025$ with futility boundary $\alpha_0 = 0.7$. According to Table 6.2 we obtain the first and second stage levels $\alpha_1 = \alpha_2 = 0.0163$. This α_2 corresponds to the second stage boundary $c_{\alpha_2} = \exp(-\chi^2_{4,1-0.0163}/2) = 0.00231$. The first stage p-value $p_1 = 0.015$ is now below the first stage level $\alpha_1 = 0.0163$. Hence, we can reject H_0 at stage 1 and stop the trial.

Table 6.2 First and second stage rejection levels $\alpha^* = \alpha_1 = \alpha_2$ with Fisher's product test ($c = c_{\alpha^*} = \exp\{-\chi^2_{4,1-\alpha^*}/2\}$) for different significant levels α and different futility boundaries $\alpha_0 = 0.3, 0.4, \ldots, 1.00$

α_0		$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$
0.30	α_1	0.0373	0.0178	0.0068	0.0033
	c_{α_2}	0.00611	0.00255	0.00084	0.00037
0.40	α_1	0.0359	0.0173	0.0066	0.0032
	c_{α_2}	0.00585	0.00246	0.00082	0.00036
0.50	α_1	0.0349	0.0169	0.0065	0.0032
	c_{α_2}	0.00566	0.00240	0.00080	0.00036
0.60	α_1	0.0342	0.0166	0.0064	0.0032
	c_{α_2}	0.00552	0.00235	0.00079	0.00035
0.70	α_1	0.0336	0.0163	0.0063	0.0031
	c_{α_2}	0.00540	0.00231	0.00078	0.00035
0.80	α_1	0.0331	0.0161	0.0063	0.0031
	c_{α_2}	0.00531	0.00227	0.00077	0.00034
0.90	α_1	0.0327	0.0159	0.0062	0.0031
	c_{α_2}	0.00522	0.00225	0.00076	0.00034
1.00	α_1	0.0323	0.0158	0.0062	0.0030
	c_{α_2}	0.00515	0.00222	0.00075	0.00034

6.2.3 Weighted Fisher's Product Test

Fisher's product criterion assigns equal weights to the stage-wise p-values p_1 and p_2. This is reasonable if the two stages are expected to have approximately equal sample sizes. If the stage-wise sample sizes are rather different, then it appears more efficient to use the weighted Fisher's product test $C(p_1, p_2) = p_1 p_2^w$ for some positive $w \neq 1$ (Fisher 1932; Brannath et al. 2002). The larger the second stage sample size in comparison to the first stage sample size, the larger we would choose the weight w. Because of a similar non-stochastic curtailment phenomenon as for the equally weighted Fisher's product test we can only choose $\alpha_1 \geq c$. With the rejection and acceptance boundaries $c \leq \alpha_1 < \alpha < \alpha_0$ and weight $w > 0$ with $w \neq 1$ the level condition (6.1) becomes

$$\alpha_1 + \frac{c^{w^{-1}} (\alpha_0^{1-w^{-1}} - \alpha_1^{1-w^{-1}})}{1 - w^{-1}} = \alpha \tag{6.5}$$

because

$$\int_{\alpha_1}^{\alpha_0} \int_0^1 \mathbf{1}_{\{p_1 p_2^w \leq c\}} dp_2\, dp_1 = \int_{\alpha_1}^{\alpha_0} \frac{c^{w^{-1}}}{p_1^{w^{-1}}}\, dp_1 = \frac{c^{w^{-1}} (\alpha_0^{1-w^{-1}} - \alpha_1^{1-w^{-1}})}{1 - w^{-1}} .$$

Table 6.3 Second stage rejection boundaries $c_{\alpha;w}$ of the weighted Fisher's product test for different levels and different weights including the case of equal weights $w = 1.0$ from Sect. 6.2.2

w	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$
0.5	0.02532	0.01258	0.00503	0.00250
0.6	0.02098	0.01030	0.00405	0.00202
0.7	0.01709	0.00826	0.00319	0.00157
0.8	0.01378	0.00646	0.00244	0.00117
0.9	0.01100	0.00500	0.00184	0.000848
1.0	0.00870	0.00380	0.00131	0.000593
1.1	0.00686	0.00287	0.000931	0.000404
1.2	0.00537	0.00214	0.000650	0.000268
1.3	0.00418	0.00158	0.000447	0.000175
1.4	0.00324	0.00115	0.000304	0.000112
1.5	0.00248	0.000839	0.000204	0.000071

In the special case without futility stopping ($\alpha_0 = 1$) and with the smallest possible first stage rejection level $\alpha_1 = c$, the level condition becomes

$$c + \frac{c^{w^{-1}}(1 - c^{1-w^{-1}})}{1 - w^{-1}} = c + \frac{c^{w^{-1}} - c}{1 - w^{-1}} = \alpha \ . \tag{6.6}$$

Equation (6.6) can be solved numerically for c. Table 6.3 gives numerical examples. We denote the resulting c by $c_{\alpha;w}$ because it corresponds to the case where we assign maximum (or full) level to the second stage. This generalizes the rejection boundary c_α defined in (6.3) for the equally weighted Fisher's product test. Accordingly, we can define $c_{\alpha;1} = c_\alpha$.

Following the philosophy of Bauer and Köhne (1994), we can also account for an early rejection boundary $\alpha_0 < 1$ by keeping the second stage boundary $c_{\alpha,w}$ and adjusting α_1 to meet the general level condition (6.5). For example, with the weight $w = 1.5$, overall level $\alpha = 0.025$ and futility boundary $\alpha_0 = 0.7$ we obtain $\alpha_1 = 0.00622$.

An alternative approach is generalizing the method of Bauer and Röhmel (1995) and to fix $\alpha_1 < \alpha < \alpha_0$ and solve (6.3) for c (Brannath et al. 2002). This gives the second stage boundary

$$d_{\alpha,\alpha_0,\alpha_1;w} = \left(\frac{(1 - w^{-1})(\alpha - \alpha_1)}{\alpha_0^{1-w^{-1}} - \alpha_1^{1-w^{-1}}} \right)^w \ .$$

As an example, we fix the weight $w = 1.5$, $\alpha = 0.025$, $\alpha_0 = 0.7$, and $\alpha_1 = 0.001$ which leads to the second stage rejection boundary $d_{0.025,0.7,0.001;1.5} = 0.00064$.

6.2.4 Inverse Normal Combination Test

Bauer and Köhne (1994) already noted that there are many ways of choosing a combination test for applying the principle. Fisher's product criterion was explicitly used just as one example. Lehmacher and Wassmer (1999) considered the *weighted inverse normal combination function*

$$C(p_1, p_2) = 1 - \Phi\big(w_1\, \Phi^{-1}(1-p_1) + w_2\, \Phi^{-1}(1-p_2)\big)\ , \tag{6.7}$$

where w_1 and w_2 denote pre-specified positive weights such that $w_1^2 + w_2^2 = 1$ and Φ^{-1} denotes the inverse of the standard normal cdf $\Phi(\cdot)$. This combination function is known from meta-analysis as well (Hedges and Olkin 1985; Sonnemann 1991). A possible choice for the weights is $w_1 = w_2 = 1/\sqrt{2}$, however, if the sample sizes of the two stages are expected to be quite different, it is more efficient to choose unequal weights.

If the p-values are independent and uniformly distributed, then $Z_1 = \Phi^{-1}(1-p_1)$ and $Z_2 = \Phi^{-1}(1-p_2)$ are independent and standard normal. Hence, also

$$\tilde{Z}_2 = w_1 \cdot Z_1 + w_2 \cdot Z_2$$

is normally distributed with mean 0 and variance $w_1^2 + w_2^2 = 1$. As a consequence,

$$C(p_1, p_2) = 1 - \Phi\big(w_1\, \Phi^{-1}(1-p_1) + w_2\, \Phi^{-1}(1-p_2)\big) = 1 - \Phi(\tilde{z}_2)$$

is uniformly distributed, like the p-value of the usual z-test. Note that the standard normal distribution of $\tilde{Z}$ and uniform distribution of $C(p_1, p_2)$ holds independently from the adaptation rule, a property that is not satisfied for the usual z-test (Proschan and Hunsberger 1995).

Since $C(p_1, p_2)$ is uniformly distributed, we obtain a level α test when using the decision boundaries $\alpha_0 = 1$, $\alpha_1 = 0$, and $c = \alpha$. This mimics the usual z-test, although, the stage-wise p-values p_1 and p_2 themselves may not come from z-tests. In the case of an early rejection ($\alpha_1 > 0$) and/or futility stopping ($\alpha_0 < 1$) the decision boundaries must be specified to meet the level condition (6.1).

Note that Z_1 and $\tilde{Z}_2 = w_1 \cdot Z_1 + w_2 \cdot Z_2$ have the same bivariate distribution as the sequential test statistics of a group sequential test with two stages and information times $t_1 = \sqrt{w_1}$ and $t_2 = 1$; see Chap. 3 in Part I. Hence, we achieve Type I error control if we choose from the family of the boundaries introduced for group sequential designs. Due to this property, the rejection boundaries α_1, α_0, and c of the weighted inverse normal test can be chosen equal to the local levels of a two-stage group sequential test with information rate t_1 (Lehmacher and Wassmer 1999). So we could, for instance, use boundaries according to O'Brien and Fleming (1979) or Pocock (1977), or use boundaries that take into account early stopping in favor of H_0. Notably, this implies that we can use standard statistical software for group sequential designs to design an adaptive design with the weighted inverse normal combination function. In particular, the calculation of sample sizes is straightforward and can be performed as described in Part I of this book. It might be regarded as a decisive advantage of using the inverse normal method as the combination function that all proposals and results from the classical group sequential theory can be used.

In contrast to Fisher's combination test, there is no intrinsic early rejection boundary associated with the weighted inverse normal combination function. Although we cannot consider non-stochastic curtailment as for Fisher's combination test, we can consider stopping for futility if $p_1 > \alpha_0$ and can calculate the significance level α_1 for the first stage p-value such that the overall significance level is met. That is, we proceed similar to the Fisher's combination test design with full level α spent in the second stage. Numerically we can obtain α_1 such that

$$\alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_0^1 \mathbf{1}_{\{C(p_1,p_2)\le\alpha\}} dp_2\, dp_1 = \alpha\,.$$

With the use of the bivariate standard normal distribution cdf $F(\cdot,\cdot)$ with correlation $1/\sqrt{2}$ (see Sect. 1.2), this can be achieved by solving

$$\begin{aligned}
&\alpha_1 + P_{H_0}(u_{1-\alpha_1} < Z_1 \le u_{1-\alpha_0}, \tilde{Z}_2 > u_{1-\alpha}) \\
&\quad = \alpha_1 + \Phi(u_{1-\alpha_1}) - \Phi(u_{1-\alpha_0}) - F(u_{1-\alpha}, u_{1-\alpha_1}) + F(u_{1-\alpha}, u_{1-\alpha_0}) \qquad (6.8)\\
&\quad = \alpha_0 - F(u_{1-\alpha}, u_{1-\alpha_1}) + F(u_{1-\alpha}, u_{1-\alpha_0}) = \alpha
\end{aligned}$$

for α_1, where $u_{1-x} = \Phi^{-1}(1-x)$ refers to the $(1-x)$-quantile of the standard normal cdf. As an example, for $\alpha = 0.05$ and $\alpha_0 = 0.5$ one obtains $\alpha_1 = 0.0044$. So the trial can be stopped at interim with the rejection of H_0 if $p_1 \le 0.0044$, and the second stage can be performed at full level α. The condition is that one has to adhere to the stopping rule, i.e., to stop the trial for futility if p_1 exceeds α_0. Figure 6.3 provides a graphical illustration of the rejection region of the inverse normal combination function in comparison to Fisher's product test for this example and the comparison with the case where no stopping for futility is foreseen.

The comparison of the rejection regions between Fisher's product test and the inverse normal method in Fig. 6.3 shows that for large p_1 it is typically easier at the second stage to reject with Fisher's product test than with the inverse normal test. As a consequence the resulting first stage rejection boundary is larger for Fisher's product test than for the inverse normal method. Also, the influence of α_0 on the first stage boundary α_1 is higher for Fisher's combination test, for the inverse normal it is only very small. The reason is the small area under the curve for $p_1 \ge \alpha_0$ for the inverse normal method and the relatively large area for Fisher's combination test. So the gain for shifting the area where $p_1 > \alpha_0$ to the top where $p_1 \le \alpha_1$ is higher for Fisher's combination test. We will call this the "heavy-tailed" property of Fisher's combination test.

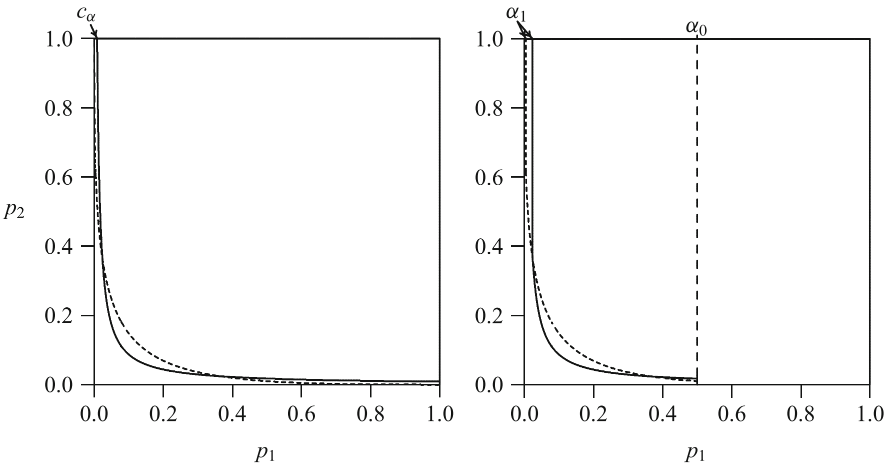

Fig. 6.3 Comparison of rejection regions of the adaptive two-stage designs according to Fisher's combination test (*solid line*) and inverse normal combination method (*dashed line*) for significance level $\alpha = 0.05$. The *left graph* shows the rejection region of the inverse normal combination function without early decision boundaries, i.e., with $\alpha_1 = 0$ and $\alpha_0 = 1$. Fisher's product test is with $\alpha_0 = 1$ and intrinsic first stage rejection boundary $\alpha_1 = c_\alpha = 0.0087$. In the *right graph*, $\alpha_0 = 0.5$ for both combination tests with the same second stage rejection boundary as in the *left panel* and α_1 such that the respective level condition is met (see text). For the inverse normal method, $\alpha_1 = 0.0044$; for Fisher's combination test, $\alpha_1 = 0.0233$

A consequence of this observation is that with the inverse normal method we will have less often contradicting first and second stage p-values when we reject H_0. The p-values appear contradictory if one p-value is large while the other one is small. Such results will cause difficulties in the interpretation of the results and may be taken as indication that there is a systematic difference between the two stages with regard to the efficacy of the study treatments. Such inhomogeneity of treatment effects is denoted as *treatment-stage interaction*. It can be caused by hidden systematic differences in the recruitment process or study conditions between the stages. The exclusion of hidden inhomogeneities is generally difficult and hence it is desirable to avoid results that (with rejection of H_0) indicate treatment-stage interactions.

As another example we illustrate the situation where a design according to Pocock (1977) is used for the inverse normal combination test. In this case,

$$\alpha^* + \int_{\alpha^*}^{\alpha_0} \int_0^1 \mathbf{1}_{\{C(p_1,p_2) \le \alpha^*\}} dp_2\, dp_1 = \alpha$$

is the condition for finding the rejection region for the first and the second stage of the trial. This is to reject H_0 at the first stage if $Z_1 \ge u_{1-\alpha^*}$ or reject H_0 at the second stage if $\tilde{Z}_2 \ge u_{1-\alpha^*}$ and thus describes a one-sided design with constant boundaries $u_1 = u_2 = u = u_{1-\alpha^*}$ taking into account a (binding) stopping for futility rule (see

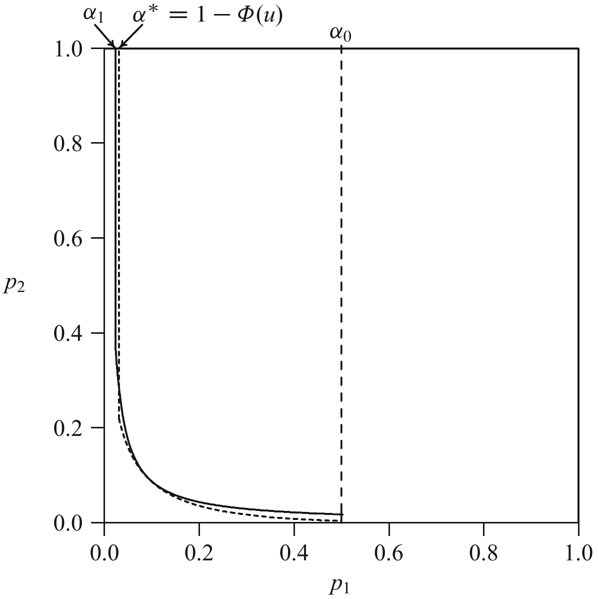

Fig. 6.4 Comparison of rejection regions of the adaptive two-stage designs according to Fisher's combination test (*solid line*) and inverse normal combination test (*dashed line*) for significance level $\alpha = 0.05$ and stopping for futility boundary $\alpha_0 = 0.50$; $\alpha_1 = 0.0233$ for Fisher's combination test, $u = 1.871$ for inverse normal combination test

Sect. 2.3). For $\alpha = 0.05$ and $\alpha_0 = 0.50$, one numerically finds $\alpha^* = 0.0307$ with $u = 1.871$. In Fig. 6.4 the rejection regions of this design are displayed together with the rejection region of a Bauer and Köhne (1994) design with $\alpha_1 = 0.0233$ and $c_\alpha = 0.0087$.

Figure 6.4 shows that the two decision regions yield quite similar rejection rules for Fisher's combination test and the inverse normal method, respectively. As above, due to the heavy tailed area of Fisher's combination test for $p_1 \geq \alpha_0$ it is more difficult for the inverse normal method that contradicting first and second stage p-values yield a significant test result at the end of the trial.

6.2.5 Sample Size Adaptations in Group Sequential Designs

The weighted inverse normal method can be used to introduce data-dependent sample size adaptations to group sequential designs. To illustrate this, let us consider a two-stage group sequential design with an experimental treatment and a control group. We assume a normally distributed endpoint with mean μ_1 under the control and μ_2 under the experimental treatment. The null hypothesis

$$H_0 : \mu_2 - \mu_1 = 0 \text{ (or } \leq 0)$$

is tested against

$$H_1 : \mu_2 - \mu_1 > 0 \ .$$

For simplicity, we assume that the variance σ^2 of the outcome is known and the same in both treatment groups. Furthermore, we assume that the sample sizes of the two treatment groups are balanced and equal to n_k at stage $k = 1, 2$.

We present the group sequential test in terms of the stage-wise z-test statistics (z-scores)

$$Z_k = \frac{\bar{X}_{k2} - \bar{X}_{k1}}{\sigma} \sqrt{\frac{n_k}{2}} \ ,$$

where n_k is the per group sample size of the cohort at stage k, and $\bar{X}_{kj}$ are the stage-wise averages at stage k in treatment group j, $j, k = 1, 2$. According to Chap. 3 a group sequential design for $H_0 : \mu_2 - \mu_1 = 0$ with first stage n_1 would use the test statistics Z_1 at stage 1 and stage 2 the cumulative z-score

$$\tilde{Z}_2 = w_1 \cdot Z_1 + w_2 \cdot Z_2 \ , \text{ where } w_k = \sqrt{\frac{n_k}{n_1 + n_2}} \ , \ k = 1, 2 \ .$$

The weighted z-score statistics $\tilde{Z}_2$ can be rewritten as

$$\tilde{Z}_2 = w_1 \cdot \Phi^{-1}(1 - p_1) + w_2 \cdot \Phi^{-1}(1 - p_2) \ ,$$

where $p_1 = 1 - \Phi(Z_1)$ and $p_2 = 1 - \Phi(Z_2)$ are the p-values from the stage-wise z-tests. Furthermore, $u^L \leq Z_1 < u_1$ is equivalent to $\alpha_1 < p_1 \leq \alpha_0$ with $\alpha_0 = 1 - \Phi(u^L)$ and $\alpha_1 = 1 - \Phi(u_1)$. Hence, the group sequential test with early rejection bound u_1 and lower stopping for futility bound u^L can be rewritten as (and is equivalent to) an inverse normal combination test. This observation provides the opportunity for adaptations of the sample sizes at the interim analysis of the two-stage group sequential test. Following the inverse normal combination method we keep using $\tilde{Z}_2 = w_1 \cdot Z_1 + w_2 \cdot Z_2$ at the second stage also when the sample size n_2 is changed to $\tilde{n}_2$ and the stage-wise test statistics Z_2 is adapted to the new sample size, i.e., $Z_2 = (\bar{X}_{22} - \bar{X}_{21})/\sigma \sqrt{\tilde{n}_2/2}$.

A group sequential test can be reformulated in terms of the weighted inverse normal method. In a sense, the reverse holds true as well. More precisely, the inverse normal method can always be reformulated in terms of the independent and standard normally distributed z-scores $Z_1 = \Phi^{-1}(1 - p_1)$ and $Z_2 = \Phi^{-1}(1 - p_2)$ computed from independent cohorts of patients. In this formulation we stop at stage 1 and reject H_0 if $Z_1 \geq u_1 = \Phi^{-1}(1-\alpha_1)$, stop for futility and retain H_0 if $Z_1 \leq u^L = \Phi^{-1}(1-\alpha_0)$, and otherwise continue to stage 2 where we reject H_0 if $\tilde{Z}_2 = w_1 \cdot Z_1 + w_2 \cdot Z_2 \geq u_2 = \Phi^{-1}(1-c)$. We will refer to this formulation of the weighted inverse normal method as the *two-stage weighted z-score test*, although, both formulations are equivalent. Cui et al. (1999) have independently proposed the weighted z-score test for adaptive sample size reassessments.

The crucial point in the above consideration is that the weights w_1 and w_2 for Z_1 and Z_2 remain fixed also when changing the second stage sample size. For this reason the cumulative statistic $\tilde{Z}_2$ has also be named *weighted z-score*. Note that after such a change the weights $w_j = \sqrt{n_1/(n_1+n_2)}$ do not reflect the actual stage 2 sample size $\tilde{n}_2$ and differ from the weights that are used in the classical z-test statistic which ignores the adaptive nature of the design. Hence, when adapting the sample size, we either over- or underweight Z_2 in $\tilde{Z}_2$ (compared to the classical z-test), depending on whether we increase or decrease the sample size. It can be shown that this miss-weighting is unavoidable in adaptive designs that equal classical group sequential designs in the non-adaptation case of staying with the pre-fixed sample sizes (Posch et al. 2003).

Lehmacher and Wassmer (1999) have shown that the loss in power due to over- or underweighting of the stage-wise z-scores is rather limited if extreme sample size adaptations are avoided. This can be illustrated by a comparison of the weighted z-score

$$\tilde{Z}_2 = w_1 \cdot Z_1 + w_2 \cdot Z_2$$

with the unweighted z-score

$$Z_2^\star = \frac{\sqrt{n_1}}{\sqrt{n_1+\tilde{n}_2}}\, Z_1 + \frac{\sqrt{\tilde{n}_2}}{\sqrt{n_1+\tilde{n}_2}}\, Z_2\ ,$$

which would have been used when the sample $\tilde{n}_2$ would have been the pre-planned one. Power is directly related to the non-centrality parameters of the test statistics. The non-centrality parameter of the weighted z-test is

$$\begin{aligned}\tilde{\vartheta} &= E(Z_2^\star) = w_1 E(Z_1) + w_2 E(Z_2)\\ &= (w_1\sqrt{n_1} + w_2\sqrt{\tilde{n}_2})\,\frac{\mu_2-\mu_1}{\sqrt{2}\sigma}\ ,\end{aligned}$$

and the non-centrality parameter of the unweighted z-score test is

$$\begin{aligned}\vartheta^* &= \sqrt{\frac{n_1}{n_1+\tilde{n}_2}}E(Z_1) + \sqrt{\frac{\tilde{n}_2}{n_1+\tilde{n}_2}}E(Z_2)\\ &= \sqrt{n_1+\tilde{n}_2}\,\frac{\mu_2-\mu_1}{\sqrt{2}\sigma}\ ,\end{aligned}$$

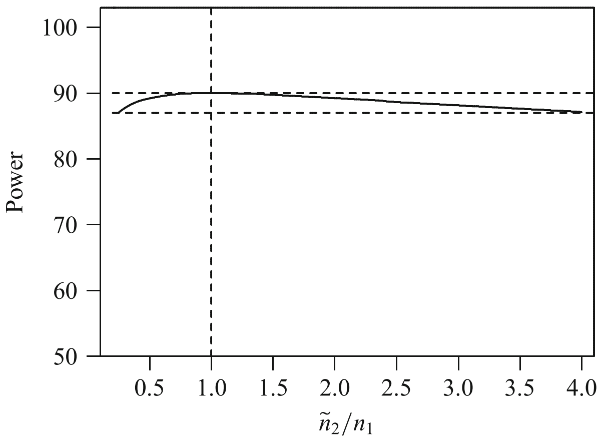

Fig. 6.5 Power of the weighted z-score test in comparison to the unweighted test. The *solid line* shows the power of the weighted z-score test when the power of the usual (unweighted) z-score test is 90 %. The *dashed horizontal lines* are at 90 and 87.5 %. No early stopping, $\alpha = 0.05$

and thus

$$\tilde{\vartheta} = \frac{w_1 + w_2\sqrt{\tilde{n}_2/n_1}}{\sqrt{1 + \tilde{n}_2/n_1}}\,\vartheta^* \le \vartheta^* \,.$$

Hence, we can illustrate the power loss which is due to using the statistic $\tilde{Z}_2$ instead of $Z_2^\star$ in dependence of $\tilde{n}_2/n_1$. Figure 6.5 provides a plot of the power with the weighted z-score test in dependence of $\tilde{n}_2/n_1$ for $w_1 = w_2 = 1/\sqrt{2}$, i.e., under the assumption of initially equal stage-wise sample sizes, and with ϑ^* such that the power of the unweighted z-score test is 90 %. Furthermore, $\alpha = 0.05$ and no early stopping is considered.

The figure shows that when $\tilde{n}_2$ is between $n_1/4$ and $4n_1$ the power loss is less than 3 % points. The same can be observed for all practically relevant significance and power values. That is, for not dramatic changes in the sample size the use of the weighted z-score test does not result in a relevant loss in power as compared to the test that is based on the usual global test statistic Z_2*. The use of this test statistic is thus reasonable taking into account the advantage of the possibility for redesigning the trial in a data-driven way.

One should note that in Fig. 6.5 the sample sizes are fixed and data independent also in the case $\tilde{n}_2 \neq n_2$. Of course, this is an unrealistic assumption for the adaptive design and hence the conclusion drawn from Fig. 6.5 should also be verified in examples where the sample sizes follow specific data-dependent rules.

The use of the inverse normal method or the weighted z-score test has been discussed and criticized a lot (for example, Burman and Sonesson 2006; Emerson 2006; Jennison and Turnbull 2003; Tsiatis and Mehta 2003). A large part of the criticism is concerned with the fact that the weighted z-score statistic is used instead of the optimum statistic Z_2^*. Certainly the use of $\tilde{Z}_2$ instead of Z_2^* might look curious, however, allowing only for small to moderate changes in the sample size weakens this criticism a lot and justifies the use of $\tilde{Z}_2$ in a two-stage adaptive design (Brannath et al. 2006a). One should also note that sample size reassessment is not the only application of the adaptive principle and thus the essential advantage or an alternative justification might become clear when discussing these applications (see Part III of this book).

A decisive (second) advantage of the inverse normal method over other combination testing methods is the following: If the weights are chosen according to the planned sample sizes and if no adaptation were performed, the inverse normal test statistic coincides with the test statistic Z_2^*. Hence, the inverse normal method can be regarded as a generalization of the classical group sequential methodology.

The choice of a futility boundary $\alpha_0 < 1$ has the advantage to increase power when the study has sufficient chance to proceed to the second stage, and if $\alpha_0 < 1$ is accounted for in (6.1) when choosing α_1 and c (Posch and Bauer 1999; Bretz et al. 2009a). The reason for this is that a decrease in α_0 leads to an increase in c and/or α_1 and thereby to an increase in the chance to reject. However, it must be noted that such choice leads to a *binding* futility boundary in the sense that the study *must* be stopped if $p_1 > \alpha_0$, because otherwise, the Type I error rate may be inflated.

Alternatively, we might set $\alpha_0 = 1$ in (6.1) for the determination of α_1 and c. By doing so, we can still stop and retain H_0 if p_1 is large, since stopping the trial with retainment of H_0 cannot inflate the Type I error rate. We call an α_0 that is not accounted for in determination of α_1 and c a *non-binding* futility boundary. A non-binding futility rule has the advantage to allow for some flexibility with regard to early retainment of H_0: the futility stopping can be an unforeseen act which needs not be pre-planned.

The disadvantage of a non-binding futility rule is that it inflates the Type II error rate, i.e., it will always deflate power if not accounted for in the power analysis. Therefore it appears advisable to pre-specify some minimal α_0 and to account for this boundary in the power analysis. If the trial is continued when $\alpha_1 < p_1 \leq \alpha_0$, then the anticipated Type II error rate is maintained as well. As a result, we determine Type I and Type II error rates under different assumptions with regard to the futility stopping rule.

In summary, we have seen that the weighted inverse normal combination method can be used to introduce data-driven sample size adaptations to group sequential designs. The attractive feature of this method is that without an adaptation we just follow the classical group sequential test. As a consequence, designing a trial might be performed in the same way as it is done for the non-adaptive case. Only if an adaptation is performed, this requires a change in the test statistic used at the second stage.

6.3 Conditional Error Function Approach

Proschan and Hunsberger (1995) introduced another way to specify the rejection region of an adaptive two-stage test design. They proposed to define such a design with the use of the so-called *conditional error function* thereby allowing for recalculating the sample size based on the observed effect. It turns out that this approach can be extended to a general principle called the *Conditional Rejection Probability (CRP)* principle as proposed by Müller and Schäfer (2004). Furthermore, there is a direct relationship between the conditional error function and the combination testing approach. In this section, we introduce the principles, describe the connections between them and discuss applications.

6.3.1 Proschan and Hunsberger's Method

Like in combination tests, one pre-specifies first stage rejection and acceptance levels $\alpha_1 < \alpha_0$ and in addition a non-increasing function $A(p_1)$ with values in the unique interval $[0, 1]$ where p_1 is the first stage p-value for H_0 computed from the first stage cohort. The function $A(p_1)$ is called the conditional error function of the design. The trial is stopped at stage 1 with rejection of H_0 if $p_1 \leq \alpha_1$ and with retainment of H_0 if $p_1 > \alpha_0$. If $\alpha_1 < p_1 \leq \alpha_0$ the trial is continued to stage 2 where H_0 is rejected if $p_2 \leq A(p_1)$. See Fig. 6.6 for a graphical illustration of the conditional error function approach.

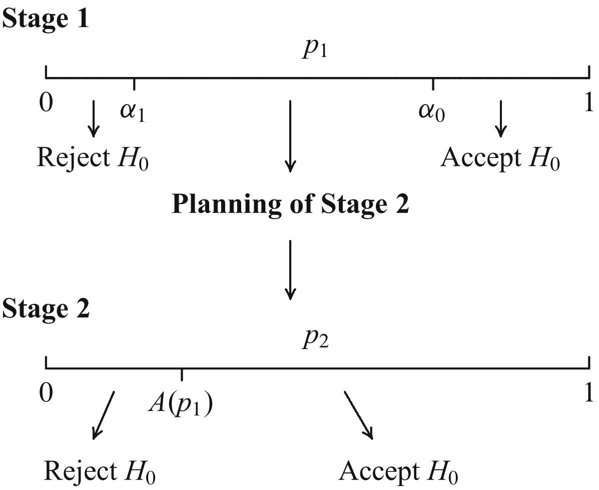

Fig. 6.6 Conditional error function approach. For the planning, fix first stage sample sizes, test, α_1, α_0, and the conditional error function $A(p_1)$ with $0 \leq A(p_1) \leq 1$. After stage 1, compute the p-value p_1 from the stage 1 data. Then, either stop or fix the design for stage 2 based on the data from stage 1. After stage 2, compute the p-value p_2 from the stage 2 data and reject H_0 if $p_2 \leq A(p_1)$

If p_2 is independent from p_1 and uniformly distributed, the conditional probability to reject H_0 given p_1 equals $A(p_1)$. If p_1 and p_2 are p-clud (as defined in Sect. 6.2.1) the conditional Type I error rate is smaller or equal to $A(p_1)$. Hence, $A(p_1)$ determines the conditional Type I error rate given the interim data. As a consequence, the overall Type I error rate is the expectation of $A(p_1)$ over the continuation region $\alpha_1 < p_1 \leq \alpha_0$ plus the probability to reject at stage 1. In order to meet the overall level α the conditional error function should therefore satisfy the level condition

$$\alpha_1 + \int_{\alpha_1}^{\alpha_0} A(p_1)dp_1 = \alpha \ . \tag{6.9}$$

Proschan and Hunsberger (1995) suggested the *circular conditional error function*

$$A(p_1) = \begin{cases} 1, & p_1 \leq c \\ 1 - \Phi\left(\sqrt{\{\Phi^{-1}(1-\alpha_1)\}^2 - \{\Phi^{-1}(1-p_1)\}^2}\right), & \alpha_1 < p_1 \leq \alpha_0 \\ 0, & p_1 > \alpha_0 \ , \end{cases} \tag{6.10}$$

where $\alpha_0 \leq 0.5$ is fixed and α_1 is determined numerically to meet condition (6.9). Examples of α_1 for $\alpha_0 = 0.1, 0.15, \ldots, 1.00$ and several α are given in Table 6.4 below.

The conditional error function (6.10) is denoted *circular* because $y = \sqrt{u^2 - x^2}$ is the equation of a circle and $A(p_1) = 1 - \Phi(\sqrt{u^2 - x^2})$ for $u = \Phi^{-1}(1 - \alpha_1)$ and $x = \Phi^{-1}(1 - p_1)$. Note the constraint that the futility boundary α_0 is below 0.5. One reason for this is that $A(p_1)$ would become increasing for $p_1 > 0.5$. We will come back to the circular conditional error function and its derivation later in Sect. 6.3.4.

By intuition, a conditional error function should be non-decreasing in p_1, because rejection of H_0 at stage 2 should be easier the stronger the first stage evidence against H_0. Note further that the early decision boundaries are included in the conditional error function. Obviously, $A(p_1) = 1$ implies that we reject for all p_2, and hence we

Table 6.4 Rejection level α_1 for the circular conditional error function (Proschan and Hunsberger test) at different significant levels α with different futility boundaries $\alpha_0 = 0.1, 0.15, \ldots, 0.50$

α_0	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$
0.10	0.03812	0.01641	0.00570	0.00262
0.15	0.03433	0.01512	0.00533	0.00247
0.20	0.03204	0.01428	0.00508	0.00236
0.25	0.03040	0.01366	0.00489	0.00228
0.30	0.02911	0.01316	0.00473	0.00221
0.40	0.02711	0.01235	0.00448	0.00210
0.50	0.02551	0.01170	0.00427	0.00201

can stop and reject H_0 already at stage 1. Similarly, $A(p_2) = 0$ implies that we retain for all p_2 and therefore stop and retain H_0 at stage 1.

We finally note that Proschan and Hunsberger (1995) originally defined conditional error functions as functions of a first stage z-score z_1 and the circular conditional error function as $A(z_1) = 1 - \Phi(\sqrt{u^2 - z_1^2})$ for $\Phi^{-1}(1-\alpha_0) \leq z_1 < u = \Phi^{-1}(1-c)$. However, this is just another way to formalize the same adaptive test. Since any z-score can be transformed to the p-value $p_1 = 1 - \Phi(z_1)$, we can always redefine a non-increasing conditional error function $A(z_1)$ to the non-decreasing conditional error function $\tilde{A}(p_1) = A\big(\Phi^{-1}(1-p_1)\big)$ in p_1.

6.3.2 Relationship Between Conditional Error Functions and Combination Tests

The combination testing and the conditional error function approach were proposed independently and did not refer to each other. However, there is a strong relationship between the two approaches. The rejection region of any combination test can be expressed in terms of a conditional error function $A(p_1)$ such that $C(p_1, p_2) \leq c$ is equivalent to $p_2 \leq A(p_1)$, namely by the function whose graph borders the rejection region in the (p_1, p_2)-plane, see, for example, Figs. 6.2 and 6.3. More formally, the conditional error function of a combination test is defined as

$$A(p_1) = \max\{y \in [0,1] : C(p_1, y) \leq c\}$$

for $\alpha_1 < p_1 \leq \alpha_0$, $A(p_1) = 1$ for $p_1 \leq \alpha_1$ and $A(p_1) = 0$ for $p_1 > \alpha_0$. The level condition (6.9) of the conditional error function is then equivalent to the level condition (6.1) of the combination test.

For instance, Fisher's product test is equivalent to $p_2 \leq A(p_1)$ with

$$A(p_1) = \begin{cases} 1, & \text{if } p_1 \leq \alpha_1 \\ c/p_1, & \text{if } \alpha_1 < p_1 \leq \alpha_0 \\ 0, & \text{if } p_1 > \alpha_0 \ . \end{cases}$$

This follows from the equivalence of $p_1 p_2 \leq c$ and $p_2 \leq c/p_1$ for $p_1 > 0$. Similarly, the rejection region of the weighted inverse normal test (or weighted z-score test) is given by the conditional error function

$$A(p_1) = \begin{cases} 1, & \text{if } p_1 \leq \alpha_1 \\ 1 - \Phi\left(\frac{u_{1-c} - w_1 \Phi^{-1}(1-p_1)}{w_2}\right), & \text{if } \alpha_1 < p_1 \leq \alpha_0 \\ 0, & \text{if } p_1 > \alpha_0 \ , \end{cases} \tag{6.11}$$

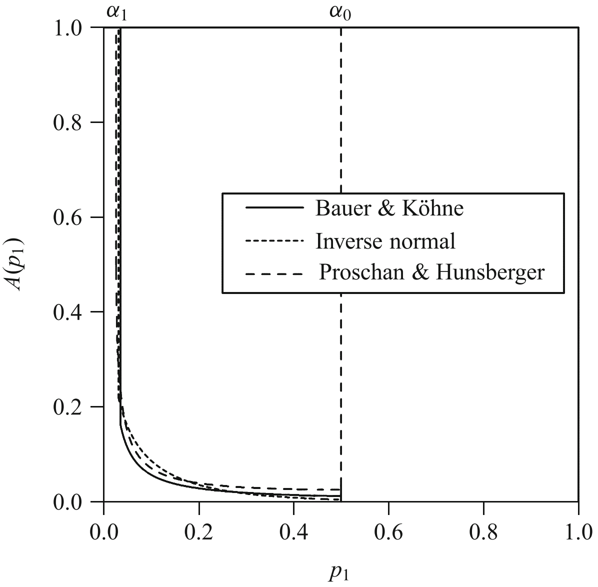

Fig. 6.7 Different conditional error functions. The figure shows the conditional error function and corresponding rejection region for the circular conditional error function, Fisher's product test, and the inverse normal combination function. All three conditional error functions are at level $\alpha = 0.05$ with $\alpha_0 = 0.5$ and $\alpha^* = \alpha_1 = \alpha_2$ such that the level condition is met. $\alpha^* = 0.0349$ for Bauer and Köhne, $\alpha^* = 0.0307$ for inverse normal, $\alpha^* = 0.0255$ for Proschan and Hunsberger

since $C(p_1, p_2) \leq c$ is equivalent to $p_2 \leq 1 - \Phi\big((u_{1-c} - w_1\Phi^{-1}(1-p_1))/w_2\big)$. In Proschan and Hunsberger (1995), this function was proposed as *linear conditional error function*.

Figure 6.7 provides a graphical illustration of the circular conditional error function together with the conditional error functions of Fisher's product test and the inverse normal method. For illustrative purposes we have fixed α and α_0 and determined $\alpha_1 = \alpha_2$ for each of the methods to meet the level condition. This is the only choice that has been suggested for all three methods. Note that different α_1 are required for the different conditional error functions to meet the level condition. Of course, the comparison of different conditional error functions will depend on the choice of α_0, α_1, and α_2.

As seen in Fig. 6.7 the circular conditional error function yields a heavy-tailed rejection region similarly to Fisher's product test. In this case ($\alpha_1 = \alpha_2$) this property is even more pronounced for the circular conditional error function. Generally, there is a tendency that these two procedures behave quite similarly and no large differences with respect to necessary maximum and expected sample size are expected (see Wassmer 1998).

We note that the relationship between a combination function and its conditional error function is the same as the relationship between a two-dimensional function and a single level curve of this function. Whereas every two-dimensional function has a unique level curve for a given value c, a single level curve does not uniquely determine the function. The strong relationship between conditional error functions and combination test has been notified by several authors (Posch and Bauer 1999; Wassmer 1999c; Proschan 2003; Vandemeulebroecke 2006).

We finally note that for any conditional error function $A(p_1)$ the combination function $C(p_1, p_2) = p_2 - A(p_1)$ with second stage boundary $c = 0$ leads to the same rejection region in the (p_1, p_2)-plane as the conditional error function $A(p_1)$. This combination function is, however, not the only one that gives the rejection region of the conditional error function, and there might be more natural combination functions associated with $A(p_1)$. Examples are the Fisher's product and the inverse normal combination test. The circular conditional error function is more naturally obtained from the combination function

$$C(p_1, p_2) = 1 - \Phi\big(\Phi^{-1}(1-p_1)^2 + \Phi^{-1}(1-p_2)^2\big) .$$

Optimal conditional error functions were derived in Brannath and Bauer (2004). Liu et al. (2002) and other authors consider more general conditional error functions which depend on more information from the interim data than just a single z-score or p-value. With a one-parameter family of distributions and for the sake of sample size reassessments it is to a large extend sufficient to consider conditional error functions that depend only on a single statistic. There are, however, examples where the conditional error function may depend on more than a single first stage statistic. Such examples occur in trials with nuisance parameters where the conditional error function can also depend on a first stage estimate of nuisance parameters (Posch et al. 2004; Timmesfeld et al. 2007; Gutjahr et al. 2011) or in adaptive trials with several hypotheses (König et al. 2008). The latter will be discussed later in Part III of the book. In the most extreme case the conditional error function could depend on all interim data. However, for the control of the overall Type I error rate the null distribution of the conditional error function must be known or be stochastically bounded by some known distribution. This implies that we need to know or estimate the null distribution of the arguments in the conditional error function.

6.3.3 The CRP Principle

Müller and Schäfer (2001, 2004) suggested to start with a conventional non-adaptive test design at level α, for example, a fixed sample size or group sequential design, and to use at the interim analysis its conditional Type I error rate as conditional error function when data-driven design adaptations will be performed. This approach has several advantages which will be discussed below.

To give a first illustration of the Müller and Schäfer method we consider a trial that starts with the conventional one-sample z-test. Let μ be the mean of independent and normally distributed responses X_i and assume that we plan to test $H_0 : \mu \leq 0$

with the conventional z-test at sample size n. The test decision function of the pre-planned z-test is $\varphi = \mathbf{1}_{\{Z \geq u_{1-\alpha}\}}$ where $Z = \sum_{i=1}^{n} X_i/\sqrt{n}$ and $u_{1-\alpha} = \Phi^{-1}(1-\alpha)$.

Assume that an interim analysis is performed after $n_1 < n$ observations, and it is decided there to change the sample size to $\tilde{n} > n_1$, $\tilde{n} \neq n$. Since the sample size was changed based on the interim data, we cannot exclude an inflation of the Type I error rate when testing H_0 with the usual z-test at the modified sample size $\tilde{n}$. With the method of Müller and Schäfer we first need to compute the conditional Type I error rate

$$A = A(\text{interim data}) = E_{H_0}(\varphi|\text{interim data})$$

of the pre-planned z-test with sample size n, where $E_{H_0}(\,\cdot \mid \text{interim data}\,)$ is the conditional expectation under the null hypothesis, H_0. If we denote by $x_1, \ldots, x_{n_1}$ the observations from the interim analysis, then simple algebra yields

$$A = P_{H_0}\left(\sum_{i=n_1+1}^{n} X_i/\sqrt{n-n_1} \geq u_{1-\alpha}\sqrt{n/(n-n_1)} - \sum_{i=1}^{n_1} x_i/\sqrt{n-n_1} \right) ,$$

where the probability is with respect to X_i, $i = n_1+1, \ldots, n$, and $x_1, \ldots, x_{n_1}$ are viewed as fixed numbers. Furthermore, $\sum_{i=n_1+1}^{n} X_i/\sqrt{n-n_1}$ is standard normal and independent from $x_1, \ldots, x_{n_1}$. This implies

$$A = 1 - \Phi\left(u_{1-\alpha}\sqrt{n/(n-n_1)} - \sum_{i=1}^{n_1} x_i/\sqrt{n-n_1} \right) ,$$

which can also be written as function of the first stage z-score $z_1 = \sum_{i=1}^{n_1} x_i/\sqrt{n_1}$, namely as

$$A(z_1) = 1 - \Phi\big(u_{1-\alpha}\sqrt{n/(n-n_1)} - z_1\sqrt{n_1/(n-n_1)}\big) . \tag{6.12}$$

In order to preserve the Type I error rate, we now have to apply a test $\tilde{\varphi}$ whose conditional Type I error rate is not larger than $A(z_1)$. The simplest way to preserve the conditional Type I error rate $A(z_1)$ is to compute a p-value p_2 for the independent second stage cohort and to reject H_0 if $p_2 \leq A(z_1)$. If, for given z_1, the p-value p_2 is uniformly distributed or has a conditional distribution that is stochastically larger than the uniform distribution, then

$$E_{H_0}(\tilde{\varphi}|\text{interim data}) = P_{H_0}(p_2 \leq A(z_1)|\text{interim data}) \leq A(z_1) .$$

In our example, p_2 could be the p-value of the z-test applied to the second stage data, namely $p_2 = 1 - \Phi(\tilde{Z}_2)$ where $\tilde{Z}_2 = \sum_{i=n_1+1}^{\tilde{n}} X_i/\sqrt{\tilde{n}-n_1}$ is the second stage z-score of the adapted trial.

The above considerations are not restricted to the one-sample z-test and can be applied to other z-tests as well. However, Type I error rate control is then typically achieved only asymptotically. Consider the null hypothesis $H_0 : \theta \leq 0$ where θ is an efficacy parameter for an experimental treatment compared to a control. We can think, for example, at a parallel group design with θ being the mean difference of some metric response variable or the log-odds ratio of a binary response. If in these (and many other) cases the pre-planned test is the score or Wald-test for H_0, then the pre-planned test is of the above form $\varphi = \mathbf{1}_{\{Z \geq u_{1-\alpha}\}}$ where Z is asymptotically standard normal.

Moreover, if Z_1 is the standardized score or Wald test statistic from the interim data, then also Z_1 is asymptotically normal and the asymptotic correlation between Z_1 and Z is $\mathrm{Corr}(Z_1, Z) = \sqrt{I_1/I}$ where I_1 and I are the amounts of information from the interim and pre-planned full data, respectively. This implies that for $\theta = 0$ the conditional distribution of Z given $Z_1 = z_1$ is the normal distribution with mean $\sqrt{I_1/I} z_1$ and variance $1 - I_1/I = (I - I_1)/I$. For the pre-planned test, this yields the conditional Type I error rate

$$A(z_1) = P_{H_0}(Z \geq u_{1-\alpha} | Z_1 = z_1) = 1 - \Phi\big(u_{1-\alpha}\sqrt{I/(I-I_1)} - z_1\sqrt{I_1/(I-I_1)}\big) \,. \tag{6.13}$$

For the design with adapted sample size we can again apply a test to the second stage data only (typically also a score or Wald test) which leads to a p-value p_2 that has a uniform or a stochastically larger conditional null distribution. Hence, the test $\tilde{\varphi} = \mathbf{1}_{\{p_2 \leq A(z_1)\}}$ has a conditional Type I error rate equal to or below $A(z_1)$.

For the z-test the Müller and Schäfer method leads to an adaptive test that is equivalent to the inverse normal or z-score method with $\alpha_1 = 0$ and $\alpha_0 = 1$. This follows from the identity of the conditional error functions (6.11) and (6.12) with $\alpha_1 = 0$, $\alpha_0 = 1$ and $w_1 = \sqrt{I_1/I}$, $w_2 = \sqrt{(I - I_1)/I}$. Moreover, when starting with a two-stage group sequential design for H_0 (instead of the z-test) then the same arguments as above show that the Müller and Schäfer method also leads to the conditional error function (6.11) now with $\alpha_1 = 1 - \Phi(u_1)$ and $\alpha_0 = 1 - \Phi(u^L)$ where u_1 and u^L are the group sequential rejection and acceptance boundaries for stage 1. However, for a group sequential test with more than two stages and sample size adaptations before the last interim analysis, the Müller and Schäfer method leads to a conditional error function which is more complex than (6.12) or (6.13) and needs numerical integration.

The above approach can be formulated in rather general terms and is called a principle, the CRP principle (Müller and Schäfer 2004). Let φ be the test decision function of the initial conventional non-adaptive test, i.e., $\varphi = 1$ if the initial test rejects and $\varphi = 0$ if it retains H_0. Assume that we perform an interim analysis after recruitment of a part of the initially anticipated sample. Assume further that we learn from the interim data and/or external information that we should change design features, for example, the sample size. As above, we compute the conditional rejection probability of the initial test given the interim data $A(\text{interim data}) = E_{H_0}(\,\varphi \mid \text{interim data}\,)$.

If we change the design, for example, increase or decrease the sample size, then we choose a new test $\tilde{\varphi}$ for the new design that has the property

$$\tilde{A} = E_{H_0}(\,\tilde{\varphi} \,|\, \text{interim data}\,) \leq A(\text{interim data}) \; . \tag{6.14}$$

Note that $\tilde{A}$ is like A the conditional rejection probability under H_0 given the interim data, however, now for the altered design and new test $\tilde{\varphi}$. Should we decide at the interim analysis that no change of the original design is required then we can follow the pre-specified design and use the original test φ which also satisfies (6.14) by definition of A.

Following this principle, the conditional Type I error rate will never exceed A(interim data), whether we alter or stay with the initial design. Hence, the overall Type I error rate will be bounded by

$$E_{H_0}\big(A(\text{interim data})\big) = E_{H_0}\big(E_{H_0}(\varphi|\text{interim data})\big) = E_{H_0}(\varphi) \leq \alpha \; ,$$

where the outer E_{H_0} denotes the unconditional expectation under H_0. This implies Type I error rate control at level α. A rigorous mathematical prove of Type I error rate control with the CRP principle can be found in Brannath et al. (2012).

The possibility to stay with the conventional design and conventional test procedure in the "no adaptation" case is the main advantage of the CRP principle. It allows us to plan a trial as usual based on classical arguments, although, the design may be changed in course of the trial. It also allows us to introduce design adaptations to trials for which no adaptations have been foreseen, as long as the conditional Type I error rate of the pre-planned test can be computed. Of course, such unscheduled design adaptations should be done only in exceptional cases and with great care, because they can question the confirmatory character of the trial and introduce complications that are beyond Type I error rate control. Nevertheless, unforeseen design adaptations are not that uncommon in clinical trials and the effect on the Type I error is often unclear such that an inflation cannot be ruled out when doing conventional tests. In this case, the application of the CRP principle, if possible, would improve the quality and validity of the trial because Type I error rate control is then out of question.

The CRP principle can be applied to any design, whether non-adaptive or adaptive, to permit additional adaptations not anticipated in advance. To apply the CRP principle we must, however, be able to compute (or at least estimate) the conditional Type I error rate of the initial test φ which can become difficult in complex designs or in the presence of nuisance parameters (Posch et al. 2004; Timmesfeld et al. 2007; Gutjahr et al. 2011). The CRP principle is also particularly helpful in the context of multi-stage group sequential trials or in adaptive trials with several hypotheses (König et al. 2008). For binary response, an application was provided by Englert and Kieser (2012, 2015).

As mentioned before, the CRP principle permits us to do adaptations at an unplanned interim analysis. This has the important implication that the interim sample size can be a random number. In this case we can apply the same method as if the interim sample size would have been known in advance. With the CRP method we need also not to be concerned about correlations between efficacy data and the interim sample size. The method even allows us to repeatedly look into the data and to decide at each look whether to change or not yet change the sample size. We know, if the sample size is not (yet) changed, then no adjustment is required. Hence the conditional Type I error rate of future interim analyses remains unchanged as well.

It is important to note that the CRP principle does not allow us to reject H_0 at an interim analysis where such rejection has not been foreseen. This is because early rejection requires the conditional Type I error rate to be equal to 1 which is impossible without a pre-specified interim test. If the interim sample size is random and independent from the interim efficacy estimate, then early rejection can be achieved by combining the error spending function method with the inverse normal or the CRP principle (Denne 2001). We can pre-specify a spending function and use the group sequential boundaries that result from the spending function and observed interim information fraction when applying the inverse normal or the Müller and Schäfer method. A similar approach is possible also with other combination tests.

6.3.4 Type I Error Maximization Method

We are now prepared to present the method of Proschan and Hunsberger (1995) which leads to the circular conditional error function. The method follows a general principle which we call the *Type I error rate maximization method* and which can be extended to other situations (for example, Graf and Bauer 2011; Graf et al. 2014). We note that this principle was already proposed in the "early days" of adaptive designs in papers such as Case et al. (1987) and Gugerli et al. (1993).

Before we present the derivation of the circular conditional error function, we consider a simplified scenario where we can choose only among two different total sample sizes $n < m$, for example, $n = 100$ and $m = 200$. The choice is made at an interim analysis with $n_1 < n$ observations, for example, $n_1 = 50$, and it can be based on all information accumulated so far. We furthermore assume that the way how we choose the sample size has not been pre-specified in advance such that the sample size adaptation remains unknown.

We ask the question whether we can adjust the rejection boundary u for the classical (unweighted) z-test such that the Type I error rate is preserved even though the sample size adaptation rule remains undetermined. A positive answer seems difficult because there is no way to determine the distribution of the unweighted z-test statistic when the sample size rule is unknown. However, the concept of the conditional error function allows us to determine the rule under which the Type I error rate is maximal. Adjusting the critical boundary such that the Type I error rate is preserved under this rule guarantees Type I error rate control with any other rule.

To determine the worst case rule that maximizes the Type I error rate we compute the conditional functions of the two z-tests with sample sizes n and m, respectively, and arbitrary rejection boundary u. From (6.12) we know that the conditional error

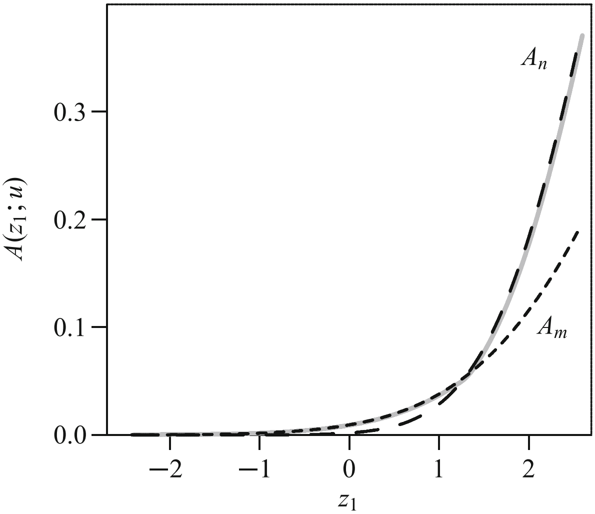

Fig. 6.8 Two conditional error functions A_n and A_m according to different samples sizes n and m. The *gray line* indicates the maximum of the two functions

functions are

$$A_n(z_1; u) = 1 - \Phi\big(u\sqrt{n/(n-n_1)} - z_1\sqrt{n_1/(n-n_1)}\big)$$

and

$$A_m(z_1; u) = 1 - \Phi\big(u\sqrt{m/(m-n_1)} - z_1\sqrt{n_1/(m-n_1)}\big) ,$$

where z_1 is the z-score computed from the interim data. Figure 6.8 provides a plot of the two conditional error functions in dependence of z_1 for $n = 100$, $m = 200$, $n_1 = 50$, and $u = 2$.

Since each of the conditional error functions equals the conditional probability to reject H_0 with the corresponding sample size, we maximize the Type I error rate by choosing for each z_1 the sample size with larger conditional error function. This maximizes the conditional and thereby also the unconditional Type I error rate. The resulting conditional rejection probability equals the maximum of the two conditional error functions

$$A_{\max}(z_1; u) = \max\{A_n(z_1; u), A_m(z_1; u)\} ,$$

which is indicated by the gray line in Fig. 6.8. Accordingly, the maximum Type I error rate is the probability under the gray line or the integral

$$\mathrm{Err}_{\max}(u) = \int_{-\infty}^{\infty} A_{\max}(z_1; u)\, \phi(z_1)\, dz_1 \ .$$

In order to preserve the Type I error rate, we now choose $\tilde{u}$ such that $\mathrm{Err}_{\max}(\tilde{u}) = \alpha$. For our numerical example in Fig. 6.8 and $\alpha = 0.025$ we obtain $\tilde{u} = 2.0537$. Hence, whatever sample size $\tilde{n} \in \{n, m\}$ we choose at interim (and whatever the reasons we have for our choice), the level is under control if we reject H_0 with $z^{(\tilde{n})} \geq 2.0537$ where $z^{(\tilde{n})}$ denotes the unweighted (classical) z-test statistic with sample size $\tilde{n}$.

Of course, the procedure of maximizing the Type I error rate leads to a conservative test if we follow another rule than the worst case rule which maximizes the Type I error rate. In practice, investigators rarely aim to maximize Type I error rates (even though their action must be expected to inflate them) and hence will usually not follow the worst case rule. Hence, using the unweighted z-tests with critical value $\tilde{u}$ such that $\mathrm{Err}_{\max}(\tilde{u}) = \alpha$ leads, in general, to a strictly conservative test.

However, using the conditional error function $A_{\max}(p_1)$ we can construct a test that exhausts the level: rejecting H_0 if $p_2 \leq A_{\max}(p_1)$ gives a test with Type I error rate

$$\int_0^1 \int_0^1 \mathbf{1}_{\{p_2 \leq A_{max}(p_1)\}} dp_1\, dp_2 = \int_0^1 A_{\max}(p_1) dp_1 = \mathrm{Err}_{\max}(\tilde{u}) = \alpha \ .$$

Hence, it appears more efficient to reject H_0 with the rejection rule $p_2 \leq A_{\max}(z_1; \tilde{u})$ than with the rule $z^{(\tilde{n})} \geq \tilde{u}$.

The conservatism of the usual z-test at level u is an indication for the inefficiency of the unweighted z-test (with adjusted critical boundary) in comparison to the adaptive test with rejection region $p_2 \leq A_{\max}(z_1; u)$. To confirm this suspicion note that with a fixed sample size n the z-test $z^{(n)} \geq u$ is equivalent to the adaptive test with rejection rule $p_2 \leq A_n(z_1)$. Hence, the z-test $z^{(\tilde{n})} \geq u$ is equivalent the adaptive test with rejection region $p_2 \leq A_{\tilde{n}}(z_1)$. Since

$$A_{\tilde{n}}(z_1) \leq \max\{A_n(z_1, u), A_m(z_1; u)\} = A_{\max}(z_1; u)$$

with $A_{\tilde{n}}(z_1; u) = A_{\max}(z_1; u)$ only if we choose the sample size with larger (conditional) Type I error rate, we obtain that the adaptive test with conditional error function $A_{\max}(z_1; u)$ rejects either as often or more often than the unweighted z-test. This shows that the adaptive test with conditional error function $A_{\max}(z_1; u)$ provides a uniform improvement of the usual z-test.

We finally note that the above observation has a remarkably fundamental consequence. It shows that the likelihood ratio test principle, which leads to the adjusted z-test $z^{(\tilde{n})}$ in the above testing situation, does not provide the most efficient test when sample sizes are changed in an unspecified way. It appears that in adaptive trials with an unspecified adaptation rule, tests based on the likelihood ratio principle can be uniformly improved by tests which follow the conditional invariance principle (see Sect. 6.1). In our example, this is the adaptive test with conditional error function $A_{\max}(z_1; u)$.

We are now coming back to the derivation of the circular conditional error function. Proschan and Hunsberger (1995) consider not only two but also any sample size $n \geq 0$. Moreover, they express the conditional error function $A_n(z_1; u)$ in terms of the ratio $(n - n_1)/n_1$. We prefer to use the reciprocal $R = n_1/(n - n_1)$ and then obtain for $A_n(z_1; u)$ the simple expression

$$A_R(z_1; u) = 1 - \Phi(u\sqrt{1+R} - z_1\sqrt{R}) \ .$$

The circular conditional error function is obtained by maximizing $A_R(z_1; u)$ with respect to $R \geq 0$ including the case $R = \infty$ which corresponds to $n = n_1$, i.e., the possibility of stopping the trial at interim. Maximizing with regard to all real positive R, even though for given n_1 only specific rational numbers (and infinity) are possible, has the advantage to obtain an upper bound that is independent from n_1. Now, maximizing $A_R(z_1; u)$ is equivalent to minimizing

$$f(R) = u\sqrt{1+R} - z_1\sqrt{R} \ .$$

Now,

$$f'(R) = \frac{1}{2}\left(\frac{u}{\sqrt{1+R}} - \frac{z_1}{\sqrt{R}}\right) \ .$$

If $z_1 \geq u$, then $f'(R) < 0$ for all $R \geq 0$ and therefore the minimum is attained for $R = \infty$. This gives the maximum conditional error function $A_{\max}(z_1; u) = 1 - \Phi(-\infty) = 1$ and means to stop the trial with rejection of H_0 at interim. It is intuitively clear that stopping (and rejecting H_0) is the optimal "fishing for significance" strategy when z_1 exceeds or meets the fixed rejection boundary u.

When $z_1 < 0$ then $f'(R) > 0$ for all $R \geq 0$ and hence the minimum of $f(R)$ is achieved for $R = 0$ in which case $f(R) = u$ and $A_{\max}(z_1, u) = 1 - \Phi(u)$. This conditional level is obtained by either increasing the second stage sample size to infinity (which is practically impossible) or discarding the first stage data and starting a new trial at level $1 - \Phi(u)$.

In the remaining cases, $0 \leq z_1 < u$, we have $f'(R) = 0$ for $R = z_1^2/(u^2 - z_1^2)$ giving the minimum

$$f(R) = u\sqrt{\frac{u^2}{u^2 - z_1^2}} - z_1\sqrt{\frac{z_1^2}{u^2 - z_1^2}} = \sqrt{u^2 - z_1^2} \ .$$

Therefore $A_{\max}(z_1;u) = 1 - \Phi(\sqrt{u^2 - z_1^2})$ for $0 \leq z_1 \leq u$. Summarizing, the maximum conditional Type I error rate is

$$A_{\max}(z_1;u) = \begin{cases} 1 - \Phi(u) & \text{if } z_1 \leq 0 \\ 1 - \Phi(\sqrt{u^2 - z_1^2}) & \text{if } 0 \leq z_1 < u \\ 1 & \text{if } z_1 \geq u\,. \end{cases}$$

Therefore, the maximum Type I error rate is

$$\begin{aligned} \mathrm{Err}_{\max}(u) &= \big(1 - \Phi(u)\big) \cdot \Phi(0) + \int_0^u \left(1 - \Phi\left(\sqrt{u^2 - z_1^2}\right)\right) \phi(z_1)\,dz_1 + 1 - \Phi(u) \\ &= \frac{3}{2}\big(1 - \Phi(u)\big) + \int_0^u \left(1 - \Phi(\sqrt{u^2 - z_1^2})\right) \phi(z_1)\,dz_1\,. \end{aligned}$$

The following observation leads to a simple analytical expression: If U and V are independent and standard normal, then $X^2 + Y^2$ is χ^2-distributed with 2 degrees of freedom which is the exponential distribution with rate parameter $\lambda = 1/2$. Therefore

$$\begin{aligned} \exp\left(-\frac{u^2}{2}\right) &= P(X^2 + Y^2 \geq u^2) \\ &= P(|X| \geq u) + \int_{-u}^{u} P(|Y| \geq \sqrt{u^2 - x^2})\phi(x)dx \\ &= 2\big(1 - \Phi(u)\big) + 4\int_0^u \big(1 - \Phi(\sqrt{u^2 - x^2})\big)\phi(x)dx\,. \end{aligned}$$

Hence

$$\mathrm{Err}_{\max}(u) = 1 - \Phi(u) + \frac{1}{4}\exp\left(-\frac{u^2}{2}\right)\,.$$

Plugging in, for instance, the critical boundary $u = 1.96$ of the z-test at level $\alpha = 0.025$ gives $\mathrm{Err}_{\max} = 0.056$. This is more than twice as large as the nominal level α.

Instead of starting a new trial when $z_1 < 0$, Proschan and Hunsberger assume stopping the trial for futility. For this reason they consider futility boundaries $\alpha_0 \leq 0.5$. With such a boundary α_0 the maximum Type I error rate is

$$\mathrm{Err}_{\max}(u, \alpha_0) = \int_{z_{\alpha_0}}^{u} \left(1 - \Phi\left(\sqrt{u^2 - z_1^2}\right)\right) \phi(z_1)\,dz_1 + 1 - \Phi(u)\,. \tag{6.15}$$

If $\alpha_0 = 0.5$ then

$$\mathrm{Err}_{\max}(u, 0.5) = \mathrm{Err}_{\max}(u) - \frac{1}{2}\big(1 - \Phi(u)\big) = \frac{1}{2}\big(1 - \Phi(u)\big) + \frac{1}{4}\exp\left(-\frac{u^2}{2}\right) .$$

For the unadjusted boundary $u = u_{1-\alpha}$ this still leads to an inflated level, for example, with $\alpha = 0.025$ we have $\mathrm{Err}(u_{1-\alpha}, 0.5) = 0.045$. Hence, in order to meet the nominal level α we need to increase the rejection boundary u to avoid a Type I error inflation with the unweighted z-test. The corresponding u can be obtained through $u = \Phi^{-1}(1 - c)$ from Table 6.4.

Expression (6.15) is identical to the Type I error rate of the adaptive test with circular conditional error function and boundaries $c = \alpha_1 = 1 - \Phi(u)$ and α_0. Like in the case with only two optional sample sizes this adaptive test is uniformly more powerful than the z-test with adjusted boundary u.

6.4 Two-Sided Adaptive Tests

Up to now we have only considered one-sided hypotheses. However, in practice, often two-sided hypotheses are considered, either for ethical reasons or because both sides of the alternative hypothesis are of scientific relevance. For an adaptive trial a test for a hypothesis $H_0 : \theta = 0$ against the two-sided alternative $H_1 : \theta \neq 0$ can be obtained by the application of a combination test to the stage-wise two-sided p-values. Since for $\theta = 0$ the two-sided p-values satisfy the p-clud condition (at least asymptotically), application of a combination test to these p-values provides Type I error rate control for H_0.

Using two-sided p-values can, however, lead to difficulties in the interpretation of the trial outcome when H_0 is rejected. In this case, we usually want to decide which side of the alternative hypothesis ($\theta > 0$ or $\theta < 0$) is true. The answer appears clear if the data of the two independent cohorts from the two stages consistently indicate the same side of the alternative. Often, the two-sided p-values are $p_k = 2\{1 - \Phi(|z_k|)\}$ where $z_k = \hat{\theta}_k/\mathrm{se}_k$ is the z-test statistic based on the maximum likelihood estimate $\hat{\theta}_k$ of θ from stage k, and se_k is an estimate of its standard deviation. When using t-tests instead of z-tests then the standard normal distribution function is replaced by a t-distribution, and the arguments below apply similarly.

Consistency between the results from the two stages means that the two z-score z_1 and z_2 from the two stages have the same sign. In this case rejection of H_0 can easily be interpreted in terms of the sign of θ. However, when the sign of z_1 and z_2 differs then rejection of the two-sided null hypothesis does not allow a claim on the sign of θ. A discussion of this issue can be found in Bauer and Köhne (1994). It can be avoided by retaining H_0 if the stage-wise data are inconsistent. This leads to a strictly conservative test and to some loss in power. If the stage-wise sample sizes are both sufficiently large, then conflicting directions will be rare under the alternative, and the loss in power will be small. However, if one (or both) of the stages is small, the loss in power can be of relevance. We refer to Bauer and Köhne (1994) for more details.

A better approach to deal with conflicting directions is to perform a separate adaptive test at level $\alpha/2$ for each of the two one-sided hypotheses

$$H_0^{(-)} : \theta \leq 0 \qquad \text{and} \qquad H_0^{(+)} : \theta \geq 0$$

(Wassmer 1999c; Müller and Schäfer 2001). Rejecting the two-sided hypothesis if one of the one-sided hypotheses is rejected provides a level α test for $H_0 : \theta = 0$. Logical consistency requires that we can never reject both one-sided null hypotheses. This is achieved by disjoint rejection regions for the two one-sided tests: with two disjoint one-sided tests we can always make a claim of the sign of θ when H_0 is rejected. Furthermore, futility stopping with retainment of $H_0 : \theta = 0$ must imply the simultaneous acceptance of both one-sided hypotheses. This implies constraints on the critical regions of the two one-sided tests.

Disjoint rejection regions can be best described in terms of their conditional error functions. Assume that the two one-sided adaptive tests have rejection rules

$$p_2^{(-)} \leq A^{(-)} \qquad \text{and} \qquad p_2^{(+)} \leq A^{(+)} ,$$

where $A^{(-)}$, $A^{(+)}$ are conditional error functions, and

$$p_2^{(-)} = 1 - \Phi(z_2) \qquad \text{and} \qquad p_2^{(+)} = 1 - \Phi(-z_2) = \Phi(z_2)$$

are the one-sided second stage p-values for $H_0^{(-)}$ and $H_0^{(+)}$, respectively. Since, $p_2^{(+)} = 1 - p_2^{(-)}$ we have that $p_2^{(+)} \leq A^{(+)}$ is equivalent to $p_2^{(-)} \geq 1 - A^{(+)}$. Hence, the rejection regions are disjoint if and only if $A^{(-)} < 1 - A^{(+)}$ for all interim outcomes. In plot(a) of Fig. 6.9 the rejection regions are not disjoint whereas in plot (b) they are.

If both conditional error functions are below 1, then $A^{(-)} < 1 - A^{(+)}$ is equivalent to $A^{(-)} + A^{(+)} < 1$, otherwise one conditional error function can be equal to 1 and the other equal to 0, i.e., $\max\{A^{(-)}, A^{(+)}\} = 1$ and $\min\{A^{(-)}, A^{(+)}\} = 0$. Hence, logical consistency is achieved if and only if at all interim sample points

$$A^{(-)} + A^{(+)} < 1 \quad \text{or} \quad \left(\min\{A^{(-)}, A^{(+)}\}, \max\{A^{(-)}, A^{(+)}\}\right) = (0, 1) .$$

There are several possibilities for the choice of logically consistent $A^{(-)}$ and $A^{(+)}$. One way is to require

$$\alpha_0^{(-)} = 1 - \alpha_1^{(-)} \qquad \text{and} \qquad \alpha_0^{(+)} = 1 - \alpha_1^{(+)} , \tag{6.16}$$

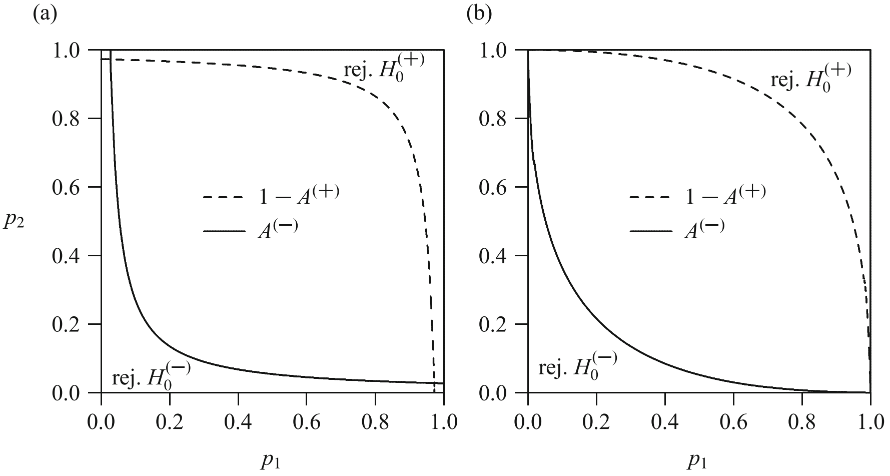

Fig. 6.9 Rejection regions of two one-sided combination tests for $H_0^{(-)}$ and $H_0^{(+)}$, for illustrative purposes, both at two-sided level $\alpha = 0.25$. Panel (**a**) is for Fisher's product test with $\alpha_1 = c_{\alpha/2} = \exp(-\chi^{-2}_{4,1-\alpha/2}/2) = 0.0271$ and $\alpha_0 = 1$, panel (**b**) is for the inverse normal combination test with $\alpha_1 = 0$ and $\alpha_0 = 1$. The choice in (**a**) does not provide logically consistent tests, since the rejection regions for the two hypotheses are not disjoint. In panel (**b**) the regions are disjoint and hence the tests are consistent

where $\alpha_1^{(+)}, \alpha_1^{(-)}$ and $\alpha_0^{(+)}, \alpha_0^{(-)}$ denote the early rejection and acceptance boundaries of the two one-sided combination tests. This requirement is illustrated in Fig. 6.10 for Fisher's combination test, two-sided $\alpha = 0.25$, and $\alpha_0^{(-)} = \alpha_0^{(+)}$, i.e., identical decision rules for $H_0^{(-)}$ and $H_0^{(+)}$.

Note that the requirement (6.16) has an only small influence on the critical values for reasonably chosen (two-sided) significance level α. This is due to the fact that the probability of conflicting decision is $c^2_{\alpha/2}$ in this case and thus very small for reasonable α.

For the inverse normal method and an early rejection boundary α_1 there is also an area where $p_1 < \alpha_1$ or $p_1 > 1 - \alpha_1$ such that (6.16) is a meaningful requirement. However, also in this case for reasonable significance level α and commonly used designs the effect of adjusting for this is negligibly small and corresponds to the fact that the critical levels of a group sequential design at one-sided level $\alpha/2$ are virtually identical to critical levels at two-sided level α (see Sect. 2.3).

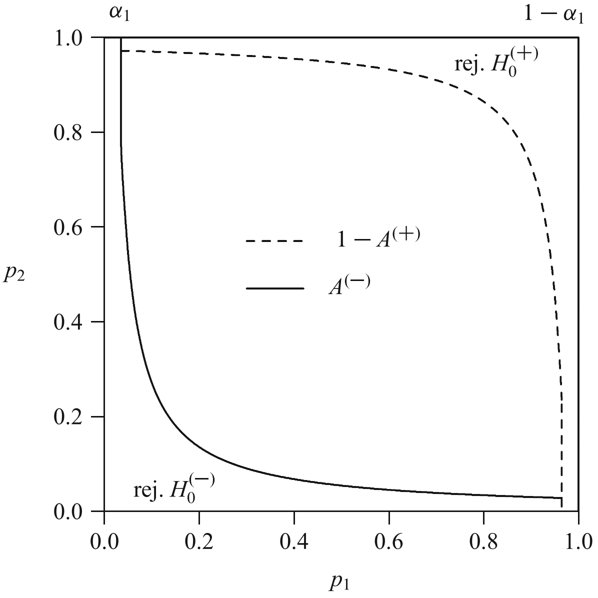

Fig. 6.10 Rejection regions of two one-sided Fisher's product tests for $H_0^{(-)}$ and $H_0^{(+)}$, two-sided $\alpha = 0.25$. It is illustrated with $c_{\alpha/2} = \exp(-\chi^{-2}_{4,1-\alpha/2}/2) = 0.0271$ and $\alpha_1 = 0.035$ and $\alpha_0 = 1 - \alpha_1$ such that the level condition is met

6.5 The Multi-Stage Case

In the seminal papers of Bauer and Köhne (1994) and Proschan and Hunsberger (1995) adaptive designs were introduced essentially for a design with two stages. Lehmacher and Wassmer (1999) and Müller and Schäfer (2001) considered the more general case of adaptive designs as a generalization of the classical group sequential designs (see also, Cui et al. 1999). On the one hand, the inverse normal combination testing principle is not restricted to two stages only, and the decision boundaries for the group sequential designs can also be used for more than two stages. On the other hand, the CRP principle can be used for any designs, including fixed sample and group sequential designs with an arbitrary number of stages. The general principle simply states that the conditional rejection probability can be calculated at *any* time point during the course of the trial. This can be done even iteratively and is not restricted to the two-stage case.

If one decides to use the inverse normal method, one can use *any* group sequential design that is defined for the sequence of overall test statistics $Z_1^*, \ldots, Z_K^*$, as described in Part I of this book. Any sequence of valid boundaries $u_1, \ldots, u_K$ can then be applied for the sequence of the inverse normal test statistics. At stage k, this is simply to use the weighted inverse normal combination function

$$C(p_1,\ldots,p_k) = 1 - \Phi\left(\frac{w_1\,\Phi^{-1}(1-p_1) + \cdots + w_k\,\Phi^{-1}(1-p_k)}{\sqrt{w_1^2 + \cdots + w_k^2}}\right) \tag{6.17}$$

for the adjusted significance levels $1 - \Phi(u_k)$, or, equivalently, the transformation

$$\tilde{Z}_k = \frac{w_1\,\Phi^{-1}(1-p_1) + \cdots + w_k\,\Phi^{-1}(1-p_k)}{\sqrt{w_1^2 + \cdots + w_k^2}} \tag{6.18}$$

for the original critical values u_k, $k = 1,\ldots,K$. In most cases, the weights will be according to the planned sample sizes $n_1,\ldots,n_K$ and we know from Chap. 3 of this book how to derive valid decision regions for (6.17) or (6.18). It is also possible to include futility stopping boundaries $u_1^0,\ldots,u_{K-1}^0$. So, for example, the designs according to Pampallona and Tsiatis (1994) are valid choices, too.

We note that it is even possible to use critical boundaries that are defined through the use of an α-spending function where the maximum number of stages, K, is pre-fixed. Clearly, this is not the application the α-spending function approach was intended for originally, but it defines a valid set of boundaries. Finally, problems as described in Sect. 6.4 for the definition of overall two-sided tests do not occur for the inverse normal combination function as it was described for Fisher's combination test. At least, there is no such practically relevant effect (see the remark in the last paragraph in this section). Hence, it is even possible to define corresponding overall two-sided tests which might be preferable from a regulator's point of view.

The application of Fisher's combination test principle to a three-stage adaptive design was described in Bauer and Köhne (1994) and Bauer and Röhmel (1995). Principally, there is no difficulty performing an adaptive design with Fisher's combination test for more than two stages. As for the two-stage case, all information may be used to plan the subsequent stages if the decision is based on the combination of the separate p-values. Considering the general case, let a study be performed with at most K stages. Applying Fisher's combination test, the study stops with the rejection of H_0 at stage k if $p_1 p_2 \ldots p_k \le c_{\alpha_k} := \exp(-\chi^2_{2k,1-\alpha_k}/2)$, where $2k$ refers to the degrees of freedom (df) of the χ^2-distribution. The study stops with the acceptance of H_0 at stage k if $p_k \ge \alpha_0^{(k)}$, $k \in \{1,\ldots,K-1\}$, where $\alpha_0^{(k)}$ are preassigned values which are not necessarily constant in k. The Type I error rate at stage k is given by

$$P_k = \int_{\alpha_1}^{\alpha_0^{(1)}} \int_{c_{\alpha_2}/p_1}^{\alpha_0^{(2)}} \int_{c_{\alpha_3}/(p_1 p_2)}^{\alpha_0^{(3)}} \cdots \int_{c_{\alpha_{k-1}}/(p_1 p_2 \cdots p_{k-2})}^{\alpha_0^{(k-1)}} \cdot \int_0^{c_{\alpha_k}/(p_1 p_2 \cdots p_{k-1})} dp_k\, dp_{k-1} \cdots dp_1\ , \tag{6.19}$$

and the overall Type I error rate of the procedure is $\sum_{k=1}^{K} P_k$. There exists a closed form solution of (6.19) that is provided in Wassmer (1999b). This formula can be used to find the "local" significance levels α_k, such that the procedure meets the overall level α. Wassmer (1999b) considered some possible choices of α_k and $\alpha_0^{(k)}$. We note that in these proposals the futility bounds are defined for the stage-wise p-values and not—as for the group sequential designs—for the overall test statistic at stage k.

Finally, the *recursive combination tests* were introduced by Brannath et al. (2002) for generalizing the combination testing principle of Bauer and Köhne (1994) using Fisher's combination test. The idea is simple and not restricted to Fisher's combination test only. Consider a two-stage combination test with rejection rule (6.3). For example, if $\alpha = 0.025$, H_0 is rejected at the second stage if

$$p_1 p_2 \leq 0.0038 \ .$$

Now suppose the first stage p-value is, say, 0.08, yielding no rejection of H_0 at the first interim analysis. For the second stage this means to reject H_0 if

$$p_2 \leq 0.0038/0.08 = 0.0475 \ .$$

In other words, the significance level to be used for the remainder of the trial is 4.75 %. The remainder of the trial needs not to consist of one stage only, it can also be, for example, a two-stage trial at level 4.75 %. This means that one can "introduce" an additional interim stage, and this can be done even iteratively. Note that this is equivalent to calculating the conditional Type I error rate when using the two-stage Fisher's combination test as initial test.

Chapter 7
Decision Tools for Adaptive Designs

Adaptive designs incorporate interim analyses where the interim data are used to decide whether to terminate or to continue the trial, and, if the trial is continued, how to design the remainder of the trial. This raises the important question of how to make the decisions on potential design modifications. On the one hand, such decisions require suitable methods and statistics to summarize the relevant information from the interim data, and there is often a priori or external information which one would like to combine with the information from the interim data. On the other hand, we need rules or at least guidance on how to use this information for the interim decisions, for example, a guidance when a sample size modification should take place.

In this chapter we remain focused on designs with a single null hypothesis and adaptations of the sample sizes. Note that a change in sample size includes the possibility of stopping the trial at the interim analysis by setting the second stage sample size equal to zero.

7.1 Conditional Power

In order to decide whether to stop the trial or to change the pre-specified trial design at an interim analysis, it appears natural to ask for the chance of a success if the trial is continued as pre-planned. This chance is best quantified by the conditional probability for a rejection of the null hypothesis given the observed interim data and the test design. This quantity is called *conditional power*. By conditioning on the interim result, one restricts attention to those future outcomes that are possible under the observed interim data and disregards those that are impossible. This would not be the case when recomputing the unconditional power at the interim analysis which is the power based on a reestimated effect size. With the unconditional power we

G. Wassmer, W. Brannath, *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*, Springer Series in Pharmaceutical Statistics,
DOI 10.1007/978-3-319-32562-0_7

would also include those outcomes that are impossible under the observed interim data and thereby disregard important information from the interim analysis.

The true conditional power depends (like the unconditional power) on the true parameter value. Since the true parameter value is unknown we need to make a specification. As we will see below, the choice of the alternative is crucial for the performance of the adaptive design. Since the choice of the parameter value in the conditional power calculation is still controversial, we will discuss several possibilities. Before this we will show how one can compute the conditional power for combination tests.

7.1.1 The Conditional Power of Combination Tests

For the computation of the conditional power with a combination test we need to know the distribution of the first and second stage p-values for parameter values of the alternative hypothesis. This requires assumptions on the null and alternative hypothesis as well as the test statistics used for the first and second stage p-values. We therefore focus here on specific but commonly encountered situations.

We assume, like in Sect. 6.2.5, that $H_0 : \theta = 0$ is tested against the one-sided alternative $H_1 : \theta > 0$ where $\theta = \mu_2 - \mu_1$ and μ_1, μ_2 refer to the means of independent and normally distributed responses (for example, in a control and a treatment group) with common variance σ^2, and where larger values of θ correspond to a more beneficial experimental treatment. We furthermore assume that the first and second stage p-values are of the form

$$p_k = 1 - \Phi(\hat{\theta}_k \sqrt{I_k}) \ , \quad k = 1, 2 \ , \tag{7.1}$$

where $\hat{\theta}_k = \bar{X}_{k2} - \bar{X}_{k1}$ is the mean difference and $I_k = n_k/(2\sigma^2)$ the information (i.e., the reciprocal of the variance of $\hat{\theta}_k$) from the stage k data. Therefore,

$$Z_k = \hat{\theta}_k \sqrt{I_k} \sim N(\theta \sqrt{I_k}, 1) \quad \text{for all } \theta \ .$$

Consider now an adaptive two-stage design with conditional error function $A(p_1)$. This means that we reject H_0 at the second stage if $p_2 \leq A(p_1)$. We know from Sect. 6.3 that every combination test has such a conditional error function. According to the above assumptions, we have that H_0 is rejected if and only if

$$\Phi^{-1}(1 - p_2) = Z_2 \geq \Phi^{-1}\big(1 - A(p_1)\big) \ .$$

Since the conditional power is the conditional probability to reject H_0, and $\hat{\theta}_2$ is independent from the first stage data, we obtain the formula

$$\mathrm{CP}_\theta = P_\theta\big(Z_2 \geq \Phi^{-1}(1 - A(p_1))\big) = 1 - \Phi\big(\Phi^{-1}(1 - A(p_1)) - \theta \sqrt{I_2}\big) \tag{7.2}$$

for the conditional power CP_θ. In this formula p_1 is considered as a fixed number (not as a random variable) because we condition on the interim data. Note that we need only consider the conditional power and conditional error function for $\alpha_1 < p_1 \leq \alpha_0$.

From (7.2) we see that for given p_1 the conditional power increases with increasing θ. Under the alternative, $\theta > 0$, the conditional power is increasing in $I_2 = n_2/(2\sigma^2)$, converging to 1 when n_2 becomes infinity. We can alternatively write the conditional power in terms of the interim estimate $\hat{\theta}_1$ instead of p_1 by plugging in expression (7.1). This yields

$$\begin{aligned}\mathrm{CP}_\theta &= 1 - \Phi\left(\Phi^{-1}\big(1 - A(1 - \Phi(\hat{\theta}_1\sqrt{I_1}))\big) - \theta\sqrt{I_2}\right) \\ &= \Phi\left(\Phi^{-1}\big(A(1 - \Phi(\hat{\theta}_1\sqrt{I_1}))\big) + \theta\sqrt{I_2}\right) .\end{aligned} \tag{7.3}$$

It can be seen from this and formula (7.2) that for any given θ the conditional power is decreasing in p_1 and increasing in $\hat{\theta}_1$ because $A(\cdot)$ is decreasing in its argument. Hence, the conditional power increases with increasing evidence for the alternative hypothesis, and it decreases with increasing evidence for the null hypothesis.

In practice, we will have to replace the information σ^2 by an interim estimate $\hat{\sigma}_1^2$ in the computation of the p-value p_1 and the conditional power CP_θ. For sufficiently large sample sizes this will provide a reasonable approximation to the unknown true conditional power.

We finally note that formulas (7.1)–(7.3) hold in a much broader context for suitable I_k. For instance, with unequal per group sample sizes n_{jk} at stage k and treatment group j they are valid for

$$\frac{n_{1k}\, n_{2k}}{(n_{1k} + n_{2k})\,\sigma^2}$$

Another example is time to event data for which the formulas hold asymptotically.

7.1.2 Conditional Power with Fisher's Product Test and Inverse Normal Method

Formula (7.3) becomes more explicit for specific combination tests. For instance, with Fisher's product test we can insert $A(p_1) = c/p_1$ where c is the critical value for the second stage test, see Sects. 6.2.2 and 6.3.2. This yields

$$\mathrm{CP}_\theta = 1 - \Phi\left(\Phi^{-1}\left(1 - \frac{c}{1 - \Phi(\hat{\theta}_1\sqrt{I_1})}\right) - \theta\sqrt{I_2}\right) .$$

According to (6.11), for the inverse normal combination test we obtain

$$\Phi^{-1}(1 - A(p_1)) = \frac{u_{1-c} - w_1 \Phi^{-1}(1 - p_1)}{w_2} = \frac{u_{1-c} - w_1 \hat{\theta}_1 \sqrt{I_1}}{w_2} ,$$

and thus

$$\mathrm{CP}_\theta = 1 - \Phi\Big(\frac{u_{1-c} - w_1 \hat{\theta}_1 \sqrt{I_1}}{w_2} - \theta \sqrt{I_2}\Big) .$$

Note that $I_1 = n_1/(2\sigma^2)$ and $I_2 = n_2/(2\sigma^2)$ depend on the actual sample sizes n_1 and n_2 whereas w_1 and w_2 are fixed numbers.

Figure 7.1 compares the conditional power for Fisher's product test and the inverse normal combination test at (one-sided) level $\alpha = 0.05$ in a two-stage design with $\alpha_0 = 0.5$ and $w_1 = w_2 = 1/\sqrt{2}$. For Fisher's combination test the full level α is located in the second stage which yields $\alpha_1 = 0.0233$ (see Table 6.1). The same α_1 and α_0 is used for the inverse normal combination test. With the use of the

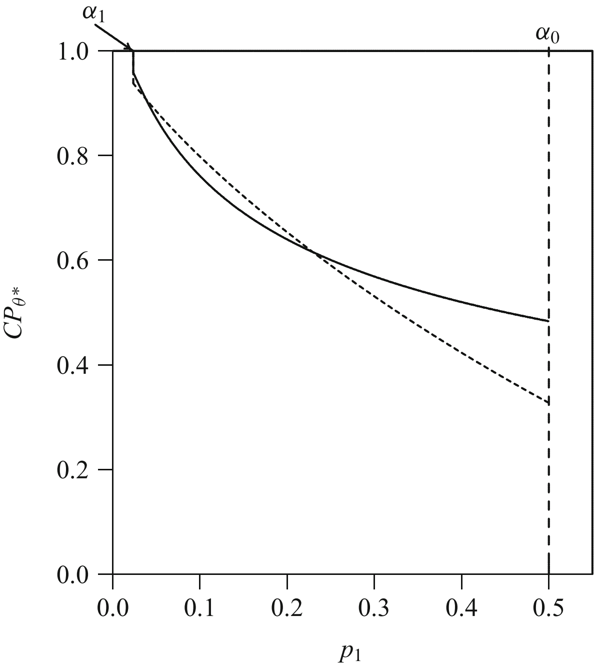

Fig. 7.1 Comparison of the conditional power of two-stage designs according to Fisher's product test (*solid line*) and inverse normal combination method (*dashed line*) at one-sided level $\alpha = 0.05$ in a two-stage design with $\alpha_0 = 0.5$, $\alpha_1 = 0.0233$, and $I_1 = I_2$. The conditional power is computed for the parameter value θ^* for which a classical z-test with information $I_1 + I_2$ would have power 90 %

bivariate standard normal cdf [equivalent to (6.8)] and a numerical search we obtain $u_{1-c} = 1.779$. The conditional power is computed for values of $n_1 = n_2$, σ, and θ for which a classical z-test with total sample size $n_1 + n_2$ would have power 90 % (this is for the "shift" $\theta\sqrt{I_2} = (1.645 + 1.282)/\sqrt{2} = 2.069$). The figure shows that with identical first stage levels, Fisher's product test has a larger conditional power than the inverse normal test if the first stage p-value is large and a smaller conditional power if the p-values are small. Note that this corresponds to the heavy tailed property of Fisher's combination test described earlier.

7.2 Futility Stopping Based on Conditional Power

If the conditional power is inadequately small, then one could stop the trial for futility. This has been suggested by Halperin et al. (1982) and Lan et al. (1982) in the context of group sequential trials (for a review of related methods, see Lachin 2005). Due to the monotonicity of the conditional power in p_1, stopping the trial if the conditional power is below a specific threshold cp_0 is equivalent to stopping the trial if p_1 is above some level α_0.

Formula (7.2) shows that the thresholds cp_0 and α_0 are related by the identities

$$\mathrm{cp}_0 = 1 - \Phi\big(\Phi^{-1}(1 - A(\alpha_0)) - \theta\sqrt{I_2}\big)$$

and

$$\alpha_0 = A^{-1}\big(\Phi(\Phi^{-1}(\mathrm{cp}_0) - \theta\sqrt{I_2})\big)\ .$$

From Fig. 7.1 explicit values can be obtained if the effect size is assumed to be equal to its original estimate. For example, for the inverse normal method and the design described in the last section, stopping for futility if $p \geq 0.50$ corresponds to stopping the trial if the conditional power falls below 33 %, for Fisher's combination test the boundary is around 48 %. Especially the latter case might be considered as problematic because a relatively large power dictates the stopping of the trial without the rejection of the null hypothesis. A problematic issue here is that the effect size is set equal to the originally assumed one. It is clear that this assumption might be inappropriate and—for example—the conditional power should be considered for a range of plausible parameter values. This more complex consideration might also be used to perform a sample size recalculation for the second stage of the trial. We consider this and related issues in the following sections.

7.3 Sample Size Modification Based on Conditional Power

An alternative to stopping the trial is to increase the conditional power by increasing the number of observations during the forthcoming stages. Often a combination of these strategies is appropriate where we stop the trial if an extreme outcome is observed at the interim analysis (with rejection or acceptance of the null hypothesis) like in group sequential designs, and otherwise recalculate the second stage sample size to achieve a sufficiently large conditional power (see Proschan and Hunsberger 1995; Lan and Trost 1997; Cui et al. 1999; Lehmacher and Wassmer 1999; Chi and Liu 1999; Posch and Bauer 2000; Denne 2001; Liu and Chi 2001; Shun et al. 2001; Friede and Kieser 2001).

Mehta and Pocock (2011) proposed the *promizing zone approach* which means to increase the sample size only for "promising" interim effects. They and Chen et al. (2004) showed that one can still use a conventional analysis if a suitable rule based on conditional power is applied. Jennison and Turnbull (2015) showed that this rule is inefficient and can be improved, but the inverse normal method instead of the conventional analysis needs to be used for the general control of the Type I error rate, see also Emerson et al. (2011) and Glimm (2012). Interestingly, the optimum rule is more symmetric and kind of inverse U shaped, with a maximum increase in sample size around the midpoint of the continuation region. We will discuss the issue of assessing overall design characteristics a bit more at the end of Sect. 7.4.

To illustrate how to choose the second stage sample size based on conditional power, like in the previous section we assume that we test the one-sided null hypothesis $H_0 : \theta \leq 0$ for the parameter θ and that the stage-wise p-values are given by $p_k = 1 - \Phi(\hat{\theta}_k \sqrt{I_k})$ for the stage-wise estimates $\hat{\theta}_k \sim N(\theta, 1/I_k), k = 1, 2$. Furthermore, we assume that we aim on a target conditional power, cp, at some alternative $\theta = \theta_a > 0$. Typical values for cp are 0.80 or 0.90. The conditional power cp can then be achieved by the choice of the second stage information I_2: choose I_2 such that the equation

$$\mathrm{CP}_{\theta_a} = 1 - \Phi\big(\Phi^{-1}(1 - A(p_1)) - \theta_a \sqrt{I_2}\big) = \mathrm{cp}$$

is satisfied. This yields the formula

$$I_2 = \frac{\big(\Phi^{-1}(\mathrm{cp}) + \Phi^{-1}(1 - A(p_1))\big)^2}{\theta_a^2}, \tag{7.4}$$

and choosing I_2 by formula (7.4) guarantees that H_0 is rejected at the end of the second stage with a probability of at least cp if $\theta \geq \theta_a$.

Note that a sample size reassessment with formula (7.4) is reasonable only if the anticipated conditional power cp is larger than the conditional error function $A(p_1)$. Hence, if we use the same target value cp for all $\alpha_1 < p_1 \leq \alpha_0$, then we should have $\mathrm{cp} \geq A(\alpha_1)$. Note also that (7.4) is just the well-known sample size formula of the z-test with the overall significance level replaced by the conditional error $A(p_1)$.

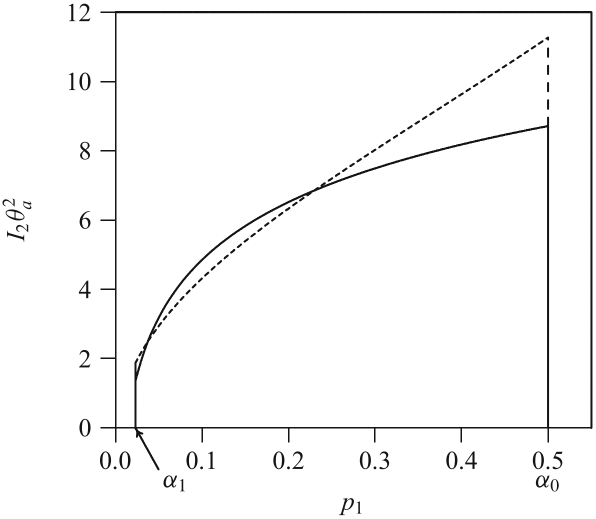

Fig. 7.2 Second stage shift $I_2\theta_a^2$ with Fisher's product test (*solid line*) and inverse normal combination method (*dashed line*) at one-sided level $\alpha = 0.05$ for conditional power cp $= 0.8$ for the designs as in Fig. 7.1

Figure 7.2 shows the squared shift $I_2\theta_a^2 = \left(\Phi^{-1}(\text{cp}) + \Phi^{-1}(1 - A(p_1))\right)^2$ for the designs from Fig. 7.1. This quantity can easily be used to perform a sample size calculation for the second stage based on conditional power at given parameter value θ_a. One can see that the maximum sample size appears for the largest first stage p-value with which the trial continues to the second stage, i.e., at $p_1 = \alpha_0$, and that the maximum is larger with the inverse normal method than with Fisher's product test. However, the inverse normal method provides smaller sample sizes for small first stage p-values.

In the common situation where $\theta = \mu_2 - \mu_1$ for means μ_1 and μ_2 of normally distributed responses of two balanced treatment groups with a common variance σ, we have that $I_2 = n_2/(2\sigma^2)$ where n_2 is per group second stage sample size. In this case formula (7.4) leads to the following reassessment formula for n_2:

$$n_2 = 2\sigma^2 \frac{\left(\Phi^{-1}(\text{cp}) + \Phi^{-1}(1 - A(p_1))\right)^2}{\theta_a^2}. \tag{7.5}$$

In practice, the determination of the second stage sample size from (7.4) requires the estimation of nuisance parameters. For example, in (7.5) we need to estimate σ. If the interim sample size is sufficiently large, we can estimate the nuisance parameters from the interim data and thereby obtain a sample size with which the target conditional power is achieved at least approximately. Alternatively, we could use a one-sided confidence limit for the nuisance parameters such that the target conditional power is achieved with some pre-specified probability (Kieser and Wassmer 1997). In general, however, this would lead to larger sample sizes.

7.4 On the Parameter Value Used in the Conditional Power and Sample Size Calculation

An important question is which parameter value should be used in the conditional power and sample size calculation. We will consider here three different possibilities for the parameter estimate and the Bayesian concept of predictive power.

7.4.1 Using a Minimal Clinically Relevant Effect Size

Sometimes one can identify the smallest effect size $\theta_{\min}$ that is considered clinically relevant or worthwhile to be identified in the clinical trial. Such minimal effect size is then used in the initial sample size calculation where we then aim to guarantee that the unconditional power is above some target value (for example, 90 %) for all $\theta \geq \theta_{\min}$.

An overall sample size calculation will in general not provide control of the conditional power: at the interim analysis the conditional probability of rejecting H_0 given the interim data may well be below this target value, although, the targeted power is achieved overall. This can be seen from Fig. 7.1 where the conditional power function of the inverse normal test (slashed line) also represents the conditional power of the usual z-test with unconditional power 90 %. The interim analysis was assumed to be at 50 % of the total sample size.

The concept of a minimal clinically relevant alternative can also be applied at the interim analysis in the conditional power and corresponding sample size calculation. In this case we compute (7.2) and (7.4) with some minimal value $\theta_a = \theta_{\min}$ for θ and thereby assure that the conditional probability of rejection H_0 at the end of the second stage is at least *cp* whenever $\theta \geq \theta_{\min}$.

The minimal value $\theta_{\min}$ can be determined at the time of the interim analysis using the interim data and all other current information and expert knowledge available so far. In some instances we may even use (7.4) with the initially pre-specified effect size. This could be the case, for instance, if the data do not provide (sufficient) evidence for a change of the a priori assumptions, but investigators aim on a sufficiently high change to reject H_0 after the interim analysis. As argued above, this chance is best quantified by the conditional power and reaching a specific conditional power will often require an increase in sample size, although, the assumption on the effect size remains constant.

The use of the pre-specified $\theta_{\min}$ may also be appropriate if the interim analysis is done for other reasons than reestimation of the sample size. Examples are the verification of safety issues or the selection of treatment arms or subpopulations. The latter examples require multiple testing and will be discussed later in Part III.

7.4.2 *Using the Interim Estimate*

A natural and often made suggestion is to use the interim estimate $\hat{\theta}_1$ for θ_a in the calculation of conditional power and second stage sample size. Since conditional power is defined only for positive θ_a we should use $\hat{\theta}_1^+ = \max\{\hat{\theta}_1, 0\}$ instead of $\hat{\theta}_1$. In practice, we need to modify sample size formula (7.4) when using $\theta_a = \hat{\theta}_1^+$ because otherwise we could obtain an infeasible large (or even infinite) sample size. Too small second stage sample sizes may also be problematic. Hence, a reasonable modification of (7.4) is

$$\hat{I}_2 = \max\left\{I_{2,\min}, \min\left\{I_{2,\max}, \frac{\left(\Phi^{-1}(\mathrm{cp}) + \Phi^{-1}(1 - A(p_1))\right)^2}{(\hat{\theta}_1^+)^2}\right\}\right\}, \tag{7.6}$$

where $I_{2,\max}$ and $I_{2,\min}$ are minimum and maximum information numbers.

One can see from this formula that $\hat{I}_2$ and hence the required second stage sample size becomes large if $I_{2,\max}$ is large and $\hat{\theta}_1^+$ is close to 0. Since values of $\hat{\theta}_1$ close to 0 are likely under the null hypothesis (or for θ close to 0), one may then obtain large expected sample sizes. This unfavorable property has been noted by Jennison and Turnbull (2003) and Tsiatis and Mehta (2003). It has caused strong concerns against adaptive sample size adjustments. In practice, we should therefore be careful in our choice of $I_{2,\max}$.

Since (7.6) produces large sample sizes for large p_1, too, large second stage sample sizes may also be avoided by a suitable choice of α_0. Another way to avoid too large second sample sizes is to impose a lower bound $\theta_{a,\min}$ for the effect size used in the conditional power and sample size calculation. Lastly, a possibility for a reduction of the excess sample size is to use a biased corrected estimate instead of the maximum likelihood estimate in the sample size recalculation formula (Coburger and Wassmer 2003). This is because a bias correction in tendency avoids extreme and thus small values of $\hat{\theta}_1^+$, at least under reasonable assumptions.

Bauer and König (2006) discuss the probability distribution of the conditional power at an interim analysis when using the interim estimate. Using this distribution, a median unbiased estimate for the unknown true conditional power can be obtained. However, the distribution is markedly skewed to the left more than to the right. Therefore there is a tendency for an underestimation of the conditional power in the mean. As a consequence we overestimate the required sample size in average. This provides another explanation for the observed inefficiency of designs with adaptive sample size adaptations. Using the initial effect size produces more stable sample sizes, however, also leads to a generally systematic miss estimation, because the true effect size will rarely equal the initially assumed one. Hence, neither estimation method seem to work perfectly.

7.4.3 Using the Bayesian Posterior Mean

Often some prior information is available which provides us with some initial effect size estimate. Bayesian methods enable us to combine such priori information with the interim data and can therefore be helpful for mid-trial sample size recalculations. More specifically, specify a normal prior distribution π for θ and use the posteriori mean for the conditional power and sample size calculation. Using the posterior mean then provides a compromise between the two previous approaches of using the initial and interim estimate.

If θ_0 is the mean and $1/I_0$ the variance of the normal prior density $\pi_0(\theta)$, then the posterior mean is the weighted sum (for example, Berger 1985)

$$\hat{\theta}_{\pi_0} = \theta_0 \frac{I_0}{I_0 + I_1} + \hat{\theta}_1 \frac{I_1}{I_0 + I_1} \ . \tag{7.7}$$

Clearly, the smaller the variance of $\pi_0(\theta)$ the larger is I_0 and the stronger is $\hat{\theta}_{\pi_0}$ influenced by the prior mean θ_0. The prior mean could be the initial efficacy estimate and I_0 the information of θ in the pilot data, often the sample size, which lead to the estimate θ_0. When there are differences between the pilot and study population (which is often the case), a smaller I_0 and also a smaller θ_0 could be more reasonable.

Using the posterior mean we obtain the conditional power estimate

$$\begin{aligned} \mathrm{CP}_{\hat{\theta}_{\pi_0}} &= 1 - \Phi\big(\Phi^{-1}(1 - A(p_1)) - \hat{\theta}_{\pi_0}\sqrt{I_2}\big) \\ &= 1 - \Phi\Big(\Phi^{-1}(1 - A(p_1)) - \frac{\theta_0 I_0 + \hat{\theta}_1 I_1}{I_0 + I_1}\sqrt{I_2}\Big) \end{aligned} \tag{7.8}$$

and the sample size recalculation formula

$$\hat{I}_2 = \max\left\{ I_{2,\min}, \min\left\{ I_{2,\max}, \frac{\big(\Phi^{-1}(\mathrm{cp}) + \Phi^{-1}(1 - A(p_1))\big)^2}{(\hat{\theta}^+_{\pi_0})^2} \right\}\right\} \ ,$$

where $\hat{\theta}^+_{\pi_0} = \max\{\hat{\theta}_{\pi_0}, 0\}$.

7.4.4 Bayesian Predictive Power

In the latter section we used the Bayesian posterior mean to account for prior knowledge in the conditional power calculation. This approach does not utilize the variance of the posterior distribution, although, the variance reflects our current uncertainty with regard to the treatment effect. The posterior variance can be utilized by adopting the Bayesian point of view more consequently and computing the Bayesian predictive power (Spiegelhalter et al. 1986; Dmitrienko and Wang 2006; Wang 2007; Lan et al. 2009). For a review of this approach, see also Dallow and Fina (2011).

To detail the concept of *predictive power* assume again a prior distribution with density $\pi_0(\theta)$. Given the interim data the density of the posterior distribution for θ is

$$\pi_1(\theta|\hat{\theta}_1) = \frac{\pi_0(\theta)\,\phi((\hat{\theta}_1-\theta)\sqrt{I_1})}{\int_{-\infty}^{\infty}\pi_0(\theta')\phi((\hat{\theta}_1-\theta')\sqrt{I_1})\,d\theta'}\ ,$$

where I_1 is the first stage information, i.e., the reciprocal of the standard deviation of the estimate $\hat{\theta}_1$. The *predictive power* is defined as the average of the conditional power with respect to the posterior distribution for θ. It is given by

$$\mathrm{PP}_{\pi_0}(\hat{\theta}_1) = \int_{-\infty}^{\infty}\mathrm{CP}_{\theta}(\hat{\theta}_1)\,\pi_1(\theta|\hat{\theta}_1)\,d\theta\ .$$

Plugging in

$$\mathrm{CP}_{\theta}(\hat{\theta}_1) = \int_{-\infty}^{\infty}\mathbf{1}_{(z_2\geq\Phi^{-1}\{1-A(p_1))}\,\phi(z_2-\theta\sqrt{I_2})\,dz_2\ ,$$

we obtain the double integral

$$\begin{aligned}\mathrm{PP}_{\pi_0}(\hat{\theta}_1) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathbf{1}_{\{z_2\geq\Phi^{-1}(1-A(p_1))\}}\,\phi(z_2-\theta\sqrt{I_2})\,\pi_1(\theta|\hat{\theta}_1)\,dz_2\,d\theta\\ &= \int_{-\infty}^{\infty}\mathbf{1}_{\{z_2\geq\Phi^{-1}(1-A(p_1))\}}\left(\int_{-\infty}^{\infty}\phi(z_2-\theta\sqrt{I_2})\,\pi_1(\theta|\hat{\theta}_1)\,d\theta\right)dz_2\\ &= \int_{\Phi^{-1}(1-A(p_1))}^{\infty}f(z_2|\hat{\theta}_1)\,dz_2\ ,\end{aligned}$$

where

$$f(z_2|\hat{\theta}_1) = \int_{-\infty}^{\infty}\phi(z_2-\theta\sqrt{I_2})\,\pi_1(\theta|\hat{\theta}_1)\,d\theta$$

is the density of the second stage z-score $z_2 = \hat{\theta}_2\sqrt{I_2}$ given that the parameter θ is a random variable with density $\pi_1(\theta|\hat{\theta}_1)$.

The density $f(z_2|\hat{\theta}_1)$ has been called *predictive distribution* (Berger 1985). It equals the convolution between the standard normal density and the posterior density of $\theta\sqrt{I_2}$ because z_2 is the sum of $\theta\sqrt{I_2}$ and the standard normally distributed random variable $z_2 - \theta\sqrt{I_2}$. Using a normal prior with mean θ_0 and variance $1/I_0$ the posterior distribution $\pi_1(\theta|\hat{\theta})$ of θ is normal with mean $\hat{\theta}_{\pi_0}$ defined in (7.7) and variance $(I_0 + I_1)^{-1}$ (Berger 1985). Therefore, $\theta\sqrt{I_2}$ is normal with mean $\hat{\theta}_{\pi_0}\sqrt{I_2}$ and variance $I_2/(I_0 + I_1)$ and $f(z_2|\hat{\theta}_1)$ is the normal density with mean $\hat{\theta}_{\pi_0}\sqrt{I_2}$ and variance $1 + I_2/(I_0 + I_1) = (I_0 + I_1 + I_2)/(I_0 + I_1)$. Hence, the predictive power becomes

$$\mathrm{PP}_{\pi_0}(\hat{\theta}_1) = 1 - \Phi\left(\sqrt{\frac{I_0 + I_1}{I_0 + I_1 + I_2}}\left(\Phi^{-1}(1 - A(p_1)) - \hat{\theta}_{\pi_0}\sqrt{I_2}\right)\right) .$$

It is interesting to note that the predictive power is different to the conditional power (7.8) with the posterior mean $\hat{\theta}_{\pi_0}$ plugged in for θ_a. The two expressions differ by the factor $\sqrt{(I_0 + I_1)/(I_0 + I_1 + I_2)}$ in the argument of the function $1 - \Phi(\cdot)$ that is not present in the conditional power and is always smaller than 1. As a consequence the argument is shrinked to 0 and thereby the predictive power is always closer to 0.5 than the conditional power.

Setting $I_0 = 0$ we obtain the predictive power from the (improper) flat prior, namely

$$\mathrm{PP}_{\pi_0}(\hat{\theta}_1) = 1 - \Phi\left(\sqrt{\frac{I_1}{I_1 + I_2}}\left(\Phi^{-1}(1 - A(p_1)) - \hat{\theta}_1\sqrt{I_2}\right)\right) .$$

A comparison to the conditional power with plug-in estimate $\theta_a = \hat{\theta}_1$ shows that the predictive power is always closer to 0.5 than this type of conditional power. The arguments are the same as in previously discussed general case. Note that the shrinkage is now stronger since $(I_0 + I_1)/(I_0 + I_1 + I_2)$ decreases for decreasing I_0. Hence, using the full Bayesian framework results in a shrinkage towards 0.5 (see Spiegelhalter et al. 2004; Proschan et al. 2006).

The predictive power can be used for the reassessment of the second stage sample size n_2, or more general the information I_2. It appears natural to reassess I_2 such that $\mathrm{PP}_{\pi_0}(\hat{\theta}_1)$ reaches some target value pp, for example, pp $= 0.8$. One can expect this reassessment method to be more robust with regard to the alternatives.

Sample size reassessment with the predictive power is more complex than with conditional power since $\mathrm{PP}_{\pi_0}(\hat{\theta}_1)$ is not always monotone in I_2 and it is bounded by some number below 1. This can be best seen from the expression

$$\mathrm{PP}_{\pi_0}(\hat{\theta}_1) = \Phi\left(\sqrt{1 - v_2}\,\Phi^{-1}(A(p_1)) + \sqrt{v_2}\hat{\theta}_{\pi_0}\sqrt{I_0 + I_1}\right) , \tag{7.9}$$

where $v_2 = I_2/(I_0 + I_1 + I_2)$ is increasing in I_2. If $\Phi^{-1}(A(p_1)) > 0$ and $\hat{\theta}_{\pi_0} > 0$, then we obtain from Cauchy's inequality that

$$\max_{I_2 \geq 0} \mathrm{PP}_{\pi_0}(\hat{\theta}_1) = \Phi\Big(\sqrt{\big(\Phi^{-1}(A(p_1))\big)^2 + \big(\hat{\theta}_{\pi_0}\big)^2 (I_0 + I_1)}\Big) ,$$

and the maximum is attained for v_2 equal to

$$v_{2,\max} = \frac{\big(\hat{\theta}_{\pi_0}\big)^2 (I_0 + I_1)}{\big(\hat{\theta}_{\pi_0}\big)^2 (I_0 + I_1) + \big(\Phi^{-1}(A(p_1))\big)^2} .$$

In all other cases we have

$$\max_{I_2 \geq 0} \mathrm{PP}_{\pi_0}(\hat{\theta}_1) = \Phi\big(\max\{\Phi^{-1}(A(p_1)), \hat{\theta}_{\pi_0}\sqrt{I_0 + I_1}\}\big) .$$

If $\max_{I_2 \geq 0} \mathrm{PP}_{\pi_0}(\hat{\theta}_1) < \mathrm{pp}$, then there is no possibility to reach the target predictive power and stopping for futility is most reasonable. This is, for example, the case when $\max\{\Phi^{-1}(A(p_1)), \hat{\theta}_{\pi_0}\sqrt{I_0 + I_1}\} < 0$ and $\mathrm{pp} > 0.5$. Therefore, the case $\Phi^{-1}(A(p_1)) < 0$ and $\hat{\theta}_{\pi_0} < 0$ provides no solution for v_2.

If $\Phi^{-1}(A(p_1)) > 0$, $\hat{\theta}_{\pi_0} > 0$, and (7.9) is larger than *pp* for $v_{2,\max}$, then we obtain the target I_2 by searching for $v_2 \in (0, v_{2,\max}]$ such that $\mathrm{PP}_{\pi_0}(\hat{\theta}_1)$ equals *pp*. Since (7.9) is increasing on $[0, v_{2,\max}]$ there is only one such v_2.

If $\max_{I_2 \geq 0} \mathrm{PP}_{\pi_0}(\hat{\theta}_1) > \mathrm{pp}$ and $\hat{\theta}_{\pi_0}\sqrt{I_0 + I_1} > 0 \geq \Phi^{-1}(A(p_1))$, then $\mathrm{PP}_{\pi_0}(\hat{\theta}_1)$ is increasing in I_2 and the maximum is attained for $v_2 = 1$. Hence, we also find a unique $I_2 > 0$ for which $\mathrm{PP}_{\pi_0}(\hat{\theta}_1) = \mathrm{pp}$.

If $\hat{\theta}_{\pi_0}\sqrt{I_0 + I_1} \leq 0 < \Phi^{-1}(A(p_1))$, then the maximum is attained for $I_2 = 0$ and $\mathrm{PP}_{\pi_0}(\hat{\theta}_1)$ is decreasing in I_2. If in addition $\max_{I_2 \geq 0} \mathrm{PP}_{\pi_0}(\hat{\theta}_1) > \mathrm{pp}$, then we will find some $I_2 > 0$ for which the target predictive power is attained. However, this I_2 may be too small for practical purposes.

Fortunately, $\hat{\theta}_{\pi_0}\sqrt{I_0 + I_1} \leq 0 < \Phi^{-1}(A(p_1))$ is a rare event that in most examples is even impossible. For instance, if $I_0 = 0$ and $A(p_1)$ is the conditional error function of the inverse normal method with weights w_1, w_2, then $\hat{\theta}_{\pi_0}\sqrt{I_0 + I_1} = \hat{\theta}_1\sqrt{I_1} = z_1 \leq 0$ implies $\Phi^{-1}(A(p_1)) = z_1(w_1/w_2) - u_{1-\alpha_2}/w_2 \leq 0$ for all $\alpha_2 \leq 0.5$. Moreover, it appears reasonable to stop the trial for futility if $\hat{\theta}_{\pi_0} \leq 0$ because $\hat{\theta}_{\pi_0}$ is the posteriori mean of the treatment effect which represents our "expectation" based on our prior knowledge and the interim data.

The reason for the non-monotonicity and boundedness of $\mathrm{PP}_{\pi_0}(\hat{\theta}_1)$ in I_2 is that not only the posterior mean but also the posterior variance of the non-centrality parameter $\theta\sqrt{I_2}$ increases with increasing I_2. Non-monotonicity occurs if $\Phi^{-1}(A(p_1)) > 0$ and $\hat{\theta}_{\pi_0} > 0$. Note that $\Phi^{-1}(A(p_1)) > 0$ implies that $A(p_1) > 0.5$ which corresponds to a large observed effect in the first stage. This usually requires only small additional information to reach significance at the second stage. On the

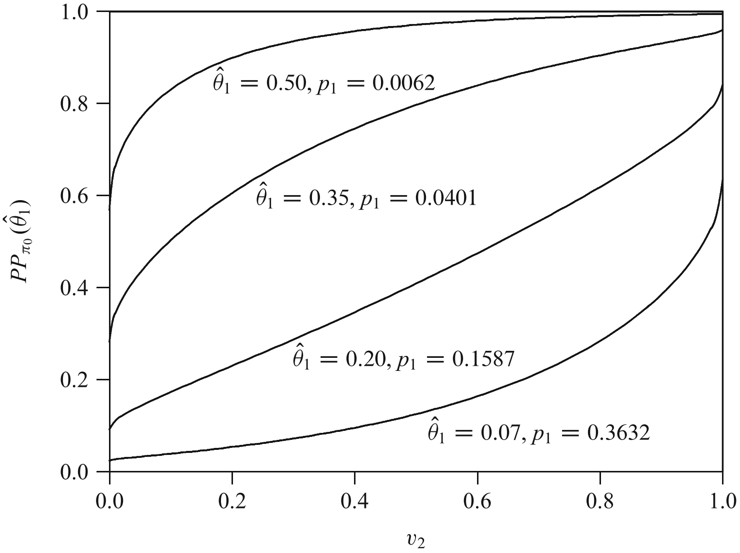

Fig. 7.3 Predictive power for flat prior, i.e., $v_2 = I_2/(I_1 + I_2)$ and inverse normal combination test at one-sided level $\alpha = 0.05$ and no futility stop. The four lines refer to different observations at interim with $I_1 = 25$

other hand, $v_{2,\max}$ is typically almost 1 such that the non-monotonicity does not cause problems in practical situations.

In Fig. 7.3, we illustrate the predictive power calculation for a flat prior, i.e., $I_0 = 0$. The graph shows $\text{PP}_{\pi_0}(\hat{\theta}_1)$ in dependence of $v_2 = I_2/(I_1 + I_2)$ for different effect sizes $\hat{\theta}_1$ at interim with $I_1 = 25$. The calculation of $A(p_1)$ assumes the inverse normal combination test with equal weights $w_1 = w_2$. A one-sided significance level $\alpha = 0.05$ and no early stopping rules are assumed, i.e., $u_{1-c} = 1.645$.

The graph shows that a predictive power of, say, 80 % can be achieved for all situations except for the one with the smallest effect and $p_1 = 0.3632$. For $p_1 = 0.1587$, however, this requires a huge sample size increase because v_2 is near to 1. The situation with $p_1 = 0.0401$ achieves predictive power around 80 % for $v_2 = 0.5$, i.e., $I_1 = I_2$. Note that only the situation with $p_1 = 0.0062$ yields $\Phi^{-1}(A(p_1)) > 0$ and indeed a local maximum is found at $v_2 = 0.995$. This situation, however, actually requires a sample size decrease for reaching 80 % predictive power such that the local maximum is of no concern. We finally note that this is a bit different in the non-adaptive case, i.e., in the situation where the usual overall test statistic is used and the weights are not fixed. Here, non-monotonicity is likely to happen for small (and reasonable) v_2 such that this might cause real problems. Dallow and Fina (2011) illustrate this and warn against the misuse of predictive power.

7.4.5 Discussion

We believe that conditional power is a useful tool for decision making at the interim analysis. The effect size used for the conditional power calculation, however, should not always be considered a realistic estimate of the true effect size. It should rather be triggered by the question of which effect sizes are considered worth the future efforts of rejecting the null hypothesis at the end of the trial. Such worthy effect sizes may be the same as at the beginning of the trial but may also change due to new evidence on clinical relevance, safety, and cost from in and outside the trial.

A reasonable alternative to computing a single conditional power is to draw a graph of the conditional power in dependency of the effect size parameter and discuss the trial perspective based on this graph and the current opinion on clinical relevance, benefits, and costs. Such a graph is comparable to the operating characteristics of a significance test but is with conditional (rather than unconditional) rejection probabilities. Hence, we call it the conditional operating characteristics.

Adding to the graph of the conditional operating characteristics a plot of the interim data's likelihood function, i.e., of the density of the interim data in dependence of the effect parameter, allows for a judgment of which effect sizes θ_a appear reasonable (and which unreasonable) for the given interim data. This is illustrated in Fig. 7.4 for the inverse normal design used for Fig. 7.1. The likelihood

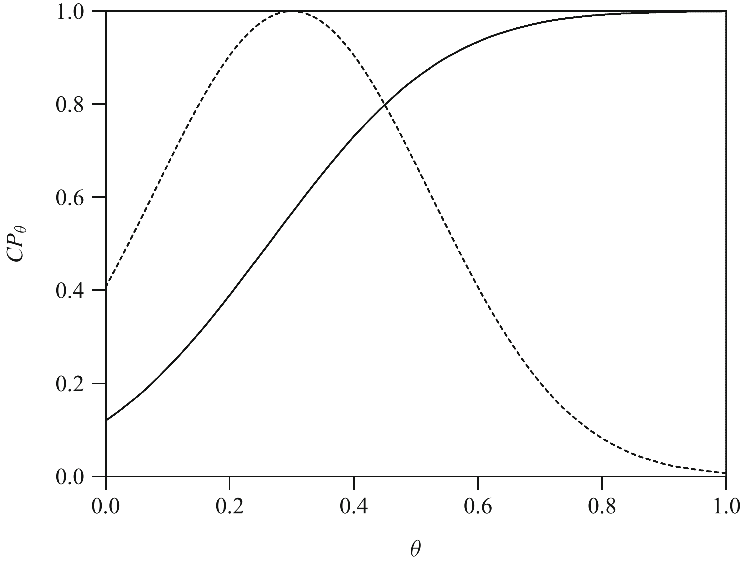

Fig. 7.4 Conditional power (*solid line*) and likelihood function (*dashed line*) for the inverse normal combination test design as in Fig. 7.1 (one-sided level $\alpha = 0.05$, $\alpha_0 = 0.5$, and $u_{1-c} = 1.779$). $\hat{\theta}_1 = 0.3$ is the interim observation at $I_1 = 20$, $I_1 = I_2$ is assumed for the conditional power calculation

function is arbitrarily scaled to have its maximum value equal to 1 which is achieved for $\theta = \hat{\theta}_1 = 0.3$. The conditional power for $\theta = 0.3$ is 56.6 % which is near to the predictive power with flat prior of 54.7 % (note the small shrinkage to 50 %). The plot shows, however, that it is not perfectly clear to increase the sample size because conditional power of 80 %, say, is achieved for parameter values with quite high likelihood.

We already noted that the process of the data-driven sample size recalculation was criticized due to the inefficiency of the resulting adaptive design as compared to a conventional group sequential design (Jennison and Turnbull 2003; Tsiatis and Mehta 2003). The use of conditional power was also criticized elsewhere (Bartroff and Lai 2008; Jennison and Turnbull 2006, 2015; Levin et al. 2013) yielding different sample size reestimation rules or even the rejection of data-driven sample size adjustments in general. The crucial point is the evaluation of overall design characteristics. Indeed, one might often find a classical group sequential designs that beats an adaptive design in terms of expected sample at given overall power curve and maximum sample size.

In response to this, first of all, we think that the evaluation of unconditional design characteristics like overall power and expected sample size remains important for adaptive designs. Particularly, in many cases extensive simulations are needed for evaluating these quantities because for many practical sample size reestimation rules no analytical formulae are available. Generally, if only a sample size increase is foreseen, there is an increase in overall power but this comes with the price of an increased overall sample size. In an adaptive design the different choices of recalculation rules (including limits for the maximum sample size) need to be rigorously evaluated and compared.

It is important to understand, however, that the actually selected sample size needs not to adhere to a pre-specified rule when using the combination testing or the CRP principle. This kind of flexibility can be regarded as an advantage in principle, too, not only because there might be unexpected facts arising at interim stages, and the adaptive design methodology allows for an adequate reaction on it. So the trial might start with a small sample size and there is the *possibility* to increase the sample size if at an interim stage it turns out that the originally planned sample size would yield an underpowered trial. The assessment of the conditional power can in this case be regarded as some kind of "updated" overall power that can be calculated under a range of effect size, and interim information on the effect size is summarized in the likelihood function. Particularly, the inverse normal combination test has the advantage that, if no adaptations were made and if the weights are chosen according to the planned sample sizes, the same statistic is used as if a common group sequential test was performed.

We even think that a sample size decrease might be reasonable, as shown in the following case study.

7.5 A Case Study

The following example that was also reported in Bauer et al. (2016) illustrates the implementation of an adaptive group sequential design with sample size reestimation in a Phase III clinical trial MUSEC (MUltiple Sclerosis and Extract of Cannabis, Trial Registration Number NCT00552604; for details see Zajicek et al. 2012) that investigated a standardized oral cannabis extract (CE) for the symptomatic relief of muscle stiffness and pain in adult patients with stable multiple sclerosis (MS) and ongoing troublesome muscle stiffness. The primary outcome measure was a 11 point category rating scale (CRS) measuring patient reported change in muscle stiffness from baseline to 12 weeks of treatment.

The pre-planned sample size calculations were based on the observed proportion of subjects with relief from muscle stiffness (0–3 categories on the CRS) in the CE and placebo arms in a previously conducted study on cannabinoids in MS: 0.42 and 0.27, respectively. A Fisher exact test for comparing such two proportions with 5 % significance level and power 80 % requires 170 evaluable subjects per arm. Adjusting for a dropout rate of 15 %, the pre-planned total sample size was 400 subjects.

An unblinded interim analysis was planned after the first 200 subjects had completed the 12 weeks treatment. An early stopping for superiority using the O'Brien and Fleming boundary was considered as well as an unblinded sample size reestimation procedure based on conditional power considerations for the second stage. The adjustment for these adaptations was implemented using the inverse normal p-value combination method with equal weights. At the time of the interim analysis, 101 subjects randomized to CE arm and 97 subjects to placebo had finished their 12 weeks treatment. The numbers of subjects with relief from muscle stiffness in the CE and placebo arms were 27 and 12, respectively. The first stage one-sided p-value was 0.0055. Early rejection was almost reached considering the first stage adjusted significance level of the O'Brien-Fleming design being 0.0026. At the time of the interim analysis 250 subjects had already been randomized and the conditional power calculations (using the pre-planned or the observed effect as true effect) for a reduced total of 300 subjects still achieved values above 90 %. Therefore, the iDMC made the recommendation to reduce the patient number from 400 to 300. Note that by sticking to the original plan the analogous conditional power calculations revealed values above 97 %, hence also very small (irrelevant) observed effect differences in a large second stage would have caused a rejection at the final statistical analysis.

The study continued enrolling new subjects and the final analysis was conducted when 143 subjects in the CE and 134 in the placebo arm completed their treatment. This was slightly below the planned target number. Overall, the rate of relief from muscle stiffness after 12 weeks was almost twice as high with CE than with placebo, 0.294 vs 0.157, the stage 2 rates were 0.357 vs. 0.243. This yielded an inverse normal test statistic of 2.573 exceeding the critical boundary 1.977 of the final analysis. Hence, the difference was statistically significant.

There was an increase in the control rate from the first to the second stage whereas the effect size slightly decreased from stage 1 to stage 2. Due to the small patient number in stage 2 this was considered as due to chance, a test for difference between the stages was not significant for both treatment groups. Also, it could be at least partly be explained by more patients with severe disease in stage 1. Nevertheless, this kind of treatment-stage interactions might be problematic since it indicates a possible time trend or even a bias introduced by the experimenter. It is not a specific problem of confirmatory adaptive designs but appears to happen when performing interim analysis.

An issue arises from decreasing rather than increasing the sample size. From the guidelines and current practice, a sample size increase seems to be appropriate, a decrease is usually discouraged. However, if early rejection was considered a valid option in the first place, adding additional data in a second stage should allow both options—an increase, but also a decrease of the planned total sample size. But we understand that statisticians must be rather brave to reduce the sample size because in case that a rejection cannot be reached in the final analysis she or he could be blamed for having reduced the sample size.

The study shows that a decrease in sample size might be a reasonable option. The decrease was additionally justified by the fact that safety was not of a major concern, so that there was no demand for a larger safety sample. From a company's perspective, the smaller necessary patient cohort seems to be attractive mainly due to the reduction in costs and time. Note also that this might be regarded as an alternative to adaptively adding interim analyses or pre-planning group sequential trials with more stages from the beginning. In these cases, a similar reduction in patients might have been achieved, however, with the cost of doing more interim stages. We also note that the conditional error approach allows to add an additional interim look and so it would have been possible to stick to the original sample size 400, but to add an additional look at 300 patients. However, this was not foreseen in the study protocol and regulators required to pre-specify the types of adaptation to be performed.

Chapter 8
Estimation and p-Values for Two-Stage Adaptive Designs

With unblinded sample size adaptations the usual estimate such as the overall treatment difference can be biased and the usual confidence intervals may not have correct coverage probabilities. Thus, when providing naïve (unadjusted) point estimates and/or confidence intervals in journals or reports, one must be aware of the poor behavior of these quantities. In adaptive designs, this is even more problematic because, for example, the coverage probability of a fixed sample confidence interval can decrease dramatically, like the Type I error rate increases as described in Sect. 6.3.4. In this chapter we discuss the construction of valid confidence intervals that have correct coverage probabilities. We will also discuss point estimates for adaptive designs that account for potential estimation bias. Specifically, we start by showing how to construct overall p-values for adaptive designs. In a sense, this is the generalization of the approaches discussed in Chap. 4 for the adaptive case. The focus of this chapter is on adaptive designs with a single interim analysis. This can be easily extended for the multi-stage case.

8.1 Overall p-Values for Adaptive Two-Stage Designs

It has become common standard to report an overall p-value at the end of a clinical trial. A p-value provides a quantitative measure of how plausible the null hypothesis is for the given trial data. The smaller the p-value the less plausible the null hypothesis. The p-value is usually defined as the largest significance level under which the null hypothesis must be retained. Hence, a p-value smaller than or equal to the nominal level α implies that we can reject the null hypothesis at level α.

G. Wassmer, W. Brannath, *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*, Springer Series in Pharmaceutical Statistics,
DOI 10.1007/978-3-319-32562-0_8

A p-value larger than α implies that we must retain H_0. If the p-value is only slightly larger than α (say 0.0255 where $\alpha = 0.025$), then the data still provide some evidence against the null hypothesis, although, the evidence is not compelling enough to formally reject H_0. Under specific circumstances such evidence may be sufficient to confirm efficacy of a new treatment. Hence, reporting p-values besides the decision on H_0 can be of high value.

The computation of a p-value requires the definition of a significance test for each significance level $0 \leq \nu \leq 1$. In a fixed sample size design such significance tests are most naturally defined via the sufficient test statistic T and a family of critical values u_ν such that $\{T \geq u_\nu\}$ provides the critical region of a level ν test. As we have seen in Chap. 4 there is no similarly natural family of hypothesis tests for group sequential designs. The reason is that the sufficient statistic in a group sequential design is two-dimensional, namely involves the data-dependent sample size and the overall z-score. Consequently, there exists more than one single construction method for p-values. We have discussed, among others, two main suggestions, namely p-values based on the stage-wise ordering and the repeated p-values. Both suggestions have merits and drawbacks. As we will see in this section both concepts can be extended to adaptive designs.

8.1.1 Stage-Wise Ordering and Related p-Values

In Sect. 4.1 we have introduced the stage-wise ordering for group sequential designs. With this ordering, rejection of H_0 at an early interim analysis is considered to provide more evidence against H_0 than rejection or acceptance of H_0 at a later stage. A sample point with acceptance of H_0 at an early stage is considered to provide less evidence against H_0 than a sample point where the trial stops at a later stage. Within stages, sample points are ordered according to the overall sufficient test statistic.

Stage-Wise Ordering for Two-Stage Combination Tests

We can introduce a similar ordering on the outcome space of a two-stage combination test. Like in Sect. 6.2.1 we denote by α_1 and α_0 the first stage rejection and acceptance levels, by $C(p_1, p_2)$ the combination function and by c the critical value for $C(p_1, p_2)$. Recall that p_1 and p_2 are the first and second stage p-values, and that the combination test rejects H_0 if either $p_1 \leq \alpha_1$ at stage 1, or $\alpha_1 < p_1 \leq \alpha_0$ and $C(p_1, p_2) \leq c$ at stage 2. In all other cases, i.e., when $p_1 > \alpha_0$, or $\alpha_1 < p_1 \leq \alpha_0$ and $C(p_1, p_2) > c$, the test retains H_0. Recall further that the combination test stops at the interim analysis when either $p_1 \leq \alpha_1$ or $p_1 > \alpha_0$, otherwise, it continues with the second stage; see Fig. 6.1. The level α of this combination test is given by formula (6.1) in Sect. 6.2.1.

The stage-wise ordering would order two different trial outcomes x and x' according to the first stage p-values $p_1,\ p'_1$ whenever both outcomes imply that the trial stops at the interim analysis. If both outcomes imply that the trial continues to the second stage, then the trial outcomes x and x' are ordered according to the combination functions $C(p_1, p_2)$ and $C(p'_1, p'_2)$. In both cases, lower values provide more evidence against H_0 than larger values. If the trial continues to the second stage for the sample point x and stops at the interim analysis for the other sample point x', then the ordering depends on p'_1: if $p'_1 \leq \alpha_1$, then x' is considered to provide more evidence against H_0 than x, if $p'_1 > \alpha_0$, then x' is thought of providing less evidence against H_0 than x.

The stage-wise ordering can alternatively be summarized as follows: for two sample points x and x', the second, x', provides more evidence against H_0 than the first, x, by short $x' \succeq x$, if either of the following three cases applies:

1. either $p_1 \leq \alpha_1$ or $p_1 > \alpha_0$, and $p'_1 \leq p_1$,
2. $\alpha_1 < p_1 \leq \alpha_0$ and $p'_1 \leq \alpha_1$,
3. $\alpha_1 < p_1 \leq \alpha_0$ and $\alpha_1 < p'_1 \leq \alpha_0$ and $C(p'_1, p'_2) \leq C(p_1, p_2)$.

In all other cases $x \succ x'$. Note that cases 1. to 3. are exclusive.

Overall Exact *p*-Value Based on the Stage-Wise Ordering

The stage-wise ordering can be used to define an overall p-value, denoted by $Q(p_1, p_2)$, that can be reported at the end of the trial. This p-value is defined as the probability under H_0 to observe a similar or more extreme outcome x' than the outcome x observed in our trial. This probability can be computed by consideration of the cases 1. to 3. and computation of the probability with respect to the outcome x' while keeping x fix. Under the assumption of independent and uniformly distributed p-values p'_1 and p'_2 this yields

$$Q(p_1, p_2) = \begin{cases} p_1 & \text{if } p_1 \leq \alpha_1 \text{ or } p_1 > \alpha_0 \\ \alpha_1 + \int_{\alpha_1}^{\alpha_0} \int_0^1 \mathbf{1}_{\{C(x,y) \leq C(p_1,p_2)\}} dy\, dx & \text{if } \alpha_1 < p_1 \leq \alpha_0 \ . \end{cases} \tag{8.1}$$

In more detail, the first line follows from case 1. and the fact that under the uniform distribution the probability of $p'_1 \leq p_1$ is at most p_1. The second line follows from adding the probabilities of cases 2. and 3. and assuming independent and uniformly distributed p-values.

Like in a group sequential design, the overall p-value equals the first stage p-value p_1 whenever the trial stops at the interim analysis. Hence, at the first stage the overall p-value is independent from the combination function. If the trial continues to the second stage, then expression (8.1) equals the Type I error rate of a combination test with interim levels α_1, α_0, combination function $C(x, y)$ and second stage critical value equal to $C(p_1, p_2)$. As a consequence, the p-value (8.1) could also be determined by the nested family of adaptive significance tests that have rejection region $R_\nu = \{p_1 \leq \nu\}$ for $\nu \leq \alpha_1$ and $\nu > \alpha_0$, and rejection region $R_\nu = \{p_1 \leq \alpha_1\} \cup \{\alpha_1 < p_1 \leq \alpha_0, C(p_1, p_2) \leq c_\nu\}$ for $\alpha_1 < \nu \leq \alpha_0$ where c_ν solves level condition (6.1) with α replaced by ν.

It is shown in Brannath et al. (2002) that under the p-clud assumption (6.2) the p-value (8.1) is in distribution larger or equal to the uniform distribution, i.e., $P_{H_0}(Q(p_1, p_2) \leq \nu) \leq \nu$ for all $0 \leq \nu \leq 1$. Hence, q is a valid p-value of the adaptive combination test. Moreover, if the p-values are independent and follow exactly the uniform distribution, i.e., are not strictly conservative, then also q follows the uniform distribution exactly, i.e., $P_{H_0}(q \leq \nu) = \nu$ for all ν.

Overall Exact p-Value for Fisher's Product Test

With Fisher's product combination function, $C(p_1, p_2) = p_1 p_2$, formula (8.1) can be made explicit and becomes

$$Q(p_1, p_2) = \begin{cases} p_1 & p_1 \leq \alpha_1 \text{ or } p_1 > \alpha_0 \\ \alpha_1 + p_1 p_2 \left(\log(\alpha_0) - \log(\alpha_1)\right) & \alpha_1 < p_1 \leq \alpha_0 \text{ and } p_1 p_2 \leq \alpha_1 \\ p_1 p_2 + p_1 p_2 \left(\log(\alpha_0) - \log(p_1 p_2)\right) & \alpha_1 < p_1 \leq \alpha_0 \text{ and } p_1 p_2 > \alpha_1 \end{cases} \tag{8.2}$$

(see Brannath et al. 2002). Formula (8.2) can be verified in the same way as we have verified formula (6.4) in Sect. 6.2.2, and from the identity

$$\int_{\alpha_1}^{\alpha_0} \int_0^1 \mathbf{1}_{\{xy \leq p_1 p_2\}} dy\, dx = p_1 p_2 - \alpha_1 + p_1 p_2 \left(\log(\alpha_0) - \log(p_1 p_2)\right)$$

for $p_1 p_2 > \alpha_1$. The second line in (8.2) follows from the non-stochastic curtailment property of Fisher's product test, namely the fact that

$$\{(x, y) \in [0, 1]^2 : x \leq p_1 p_2\} \subseteq \{(x, y) \in [0, 1]^2 : xy \leq p_1 p_2\} \,.$$

Figure 8.1 shows a plot of the p-value function for Fisher's combination test, for illustrative reasons with $\alpha_1 = 0.2$ and $\alpha_0 = 0.6$. As already mentioned the p-value equals p_1 if the trial stops at the first stage. For $\alpha_1 < p_1 \leq \alpha_0$ and $0 \leq p_2 \leq 1$ we have $\alpha_1 < Q(p_1, p_2) \leq \alpha_0$ and $Q(p_1, 0) = \alpha_1$, and we have $Q(\alpha_0, 1) = \alpha_0$. This means that when proceeding to the second stage the overall p-value can

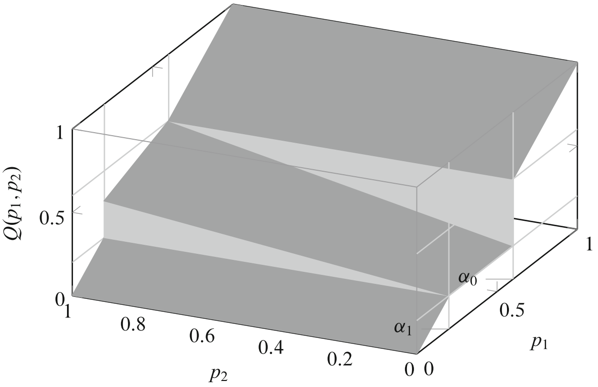

Fig. 8.1 Overall exact p-value for Fisher' product test with $\alpha_1 = 0.2$ and $\alpha_0 = 0.6$

never fall below α_1 and is never larger than α_0. This property (which holds for other combination functions as well) may be questioned if at the second stage a considerably larger sample size is used. However, it is a favorable property that the p-value equals p_1 in case of stopping at the first stage and the previous unfavorable property mentioned first is a consequence.

Overall Exact *p*-Value for Inverse Normal Method

If we use the inverse normal combination function (6.7) or z-score combination method (see Sect. 6.2) with weights w_1 and w_2, then the p-value (8.1) equals the stage-wise ordering p-value of a group sequential test discussed in Sect. 4.1, with interim information fraction $t_1 = w_1^2$ and sequential z-scores z_1 and $\tilde{z}_2 = w_1 z_1 + w_2 z_2$ where z_1 and z_2 are the stage-wise z-scores. In particular, if the weights are according to the pre-planned information fraction t_1 ($w_1 = \sqrt{t_1}$ and $w_2 = \sqrt{1 - t_1}$) and the actual information fraction is as pre-planned, then the overall exact p-value (8.1) coincides with the stage-wise ordering p-value of the pre-planned group sequential design (see Sect. 4.1.2).

8.1.2 Repeated p-Values for Two-Stage Combination Tests

Repeated p-values for group sequential tests were introduced in Sect. 4.1.2. Such p-values are not directly defined via a sample space ordering. Instead they are defined via a family of sequential boundaries for different significance levels $0 \leq \nu \leq 1$. Of course, given only the boundaries at the fixed level α such family is not uniquely defined.

If the group sequential boundaries at level α are of a common type, like Wang and Tsiatis boundaries for a specific Δ, or the boundaries are determined from a specific family of spending functions (for example, O'Brien and Fleming type spending functions), then it appears natural to use group sequential boundaries at different significance levels ν of the same type; see Sect. 4.1 for details. Due to the close relationship to group sequential tests, this approach can directly be extended to the inverse normal or weighted z-score combination method. We will first illustrate such a type of p-value with an example. Afterwards, we discuss how to extend the repeated p-value method for other combination tests.

Repeated p-Value for Inverse Normal Combination Tests

To illustrate the above mentioned construction method for repeated p-values of an inverse normal combination test, assume that we use the inverse normal combination function $C(p_1,p_2) = 1 - \Phi\{\sqrt{0.5}\,\Phi^{-1}(1-p_1) + \sqrt{0.5}\,\Phi^{-1}(1-p_2)\}$, without a binding futility rule ($\alpha_0 = 1$), and with first and second stage rejection levels according to the Wang and Tsiatis family

$$\alpha_1 = 1 - \Phi(c_{WT}(\alpha,\Delta)) \quad \text{and} \quad c = 1 - \Phi(c_{WT}(\alpha,\Delta)\cdot 2^{\Delta-0.5})\ ,$$

where Δ is fixed and $c_{WT}(\alpha,\Delta)$ is such that the corresponding group sequential test has Type I error rate α. Recall, that these boundaries will meet the level condition (6.1) for the combination test. Using the same type of rejection boundaries (and combination function) for all significance levels $0 \le \nu \le 1$, we let

$$\alpha_{1,\nu} = 1 - \Phi(c_{WT}(\nu,\Delta)) \quad \text{and} \quad c_\nu = 1 - \Phi(c_{WT}(\nu,\Delta)\cdot 2^{\Delta-0.5})\ ,$$

where $c_{WT}(\nu,\Delta)$ is such that level condition (6.1) is satisfied with α replaced by ν. Obviously, $c_{WT}(\nu,\Delta)$ is decreasing in ν. Hence, the supremum of significance levels for which H_0 is retained, is obtained by solving in ν the equation $p_1 = 1 - \Phi(c_{WT}(\nu,\Delta))$ if the trial stops at the first stage, and the equation $C(p_1,p_2) = 1 - \Phi(c_{WT}(\nu,\Delta)\cdot 2^{\Delta-0.5})$ if the trial continues to the second stage (see Sect. 4.1.2 where a direct calculation of the overall repeated p-value is described).

Construction of Repeated p-Values for General Combination Tests

To construct a repeated p-value for a combination test with combination function $C(x,y)$, we need a family of first stage rejection levels $\alpha_{1,\nu} < \nu$, $\nu \in [0,1]$, a family of first stage acceptance levels $\alpha_{0,\nu} > \nu$, $\nu \in [0,1]$, and a family of second stage critical values c_ν such that

$$\alpha_{1,\nu} + \int_{\alpha_{1,\nu}}^{\alpha_{0,\nu}} \mathbf{1}_{\{C(x,y)\le c_\nu\}}\, dy\, dx = \nu\ . \tag{8.3}$$

Given the levels $\alpha_{1,\nu}$, $\alpha_{0,\nu}$, and c_ν, the repeated p-value is defined as the largest ν for which the trial outcome is known to be outside the rejection region

$$R_\nu = \{p_1 \leq \alpha_{1,\nu}\} \cup \{p_1 \leq \alpha_{0,\nu}, C(p_1, p_2) \leq c_\nu\} \,. \tag{8.4}$$

If the trial stops at the first stage, then the repeated p-value can be computed as the supremum of those ν where $p_1 > \alpha_{1,\nu}$. We denote this first stage repeated p-value by q_1. If $\alpha_{1,\nu}$ is strictly increasing and continuous in ν, then q_1 is the unique root of $p_1 - \alpha_{1,\nu}$ which can easily be computed by numeric root finding. If $\alpha_{1,\nu}$ fails to be increasing in ν, then q_1 must be determined by a grid search which is more elaborative and requires continuity of $\alpha_{1,\nu}$ in ν.

If the trial continues to the second stage, then the p-value is determined as the supremum of ν where $p_1 > \alpha_{0,\nu}$ or $C(p_1, p_2) > c_\nu$. If $\alpha_{0,\nu}$ and c_ν are both strictly increasing and continuous in ν, then we can compute the unique root ν_0 of $p_1 - \alpha_{0,\nu}$ and the unique root ν_2 of $C(p_1, p_2) - c_\nu$ and define the p-value as the maximum $q_2 = \max\{\nu_0, \nu_2\}$. If $\alpha_{0,\nu}$ or c_ν fail to be increasing (but both are continuous in ν), then ν_0 and/or ν_2 can be determined by a grid search.

By continuity of the left side of (8.3) in $\alpha_{1,\nu}$, $\alpha_{0,\nu}$, and c_ν, all three boundaries will be continuous in ν whenever two of them are continuous. However, monotonicity in ν is not implied in the same way and may be difficult to achieve for all three quantities simultaneously. That such choice can be possible has been seen in the example for the inverse normal method and will be illustrated later for Fisher's product test.

Note that if $\nu_0 > \nu_2$, then the second stage p-value q_2 equals ν_0 which depends only on the first stage p-value p_1. This may be viewed as a disadvantage of repeated p-values because the second stage data are not utilized in this case. However, the larger $\alpha_{0,\nu}$, $\nu \in [0, 1]$, the less frequent we will have $\nu_0 > \nu_2$, and q_2 will equal ν_2 (which depends on both stage-wise p-values) in most cases. Moreover, without a binding futility rule ($\alpha_0 = 1$) we can choose $\alpha_{0,\nu} = 1$ for all $0 \leq \nu \leq 1$, and then the second stage repeated p-value always equals ν_2.

Monitoring Property of Repeated p-Values

A remarkable feature of the repeated p-value is that it provides p-values for both stages, and that both p-values, q_1 and q_2, can be reported independently from the stopping rule. More precisely, we can compute (and report) q_1 at the interim analysis and then freely decide whether we want to stop or continue the trial. If we decide to stop, then we report q_1 as final p-value, if we decide to continue, then we can compute and report q_2 at the end of the second stage as final p-value. The decision of whether to stop or continue can be made independently of any pre-specified stopping rule, and it can be based on all data or any other information from in and outside the trial. In summary, if $S \in \{1, 2\}$ denotes the stage at which the trial was stopped, then q_S is a valid p-value independently of how and for which reasons the trial was terminated at stage S.

The above flexibility with regard to the stopping rule follows from the fact that by (8.3) the critical regions R_ν defined in (8.4) have Type I error rate less than or equal to ν also in the (unrealistic) case where we always continue to the second stage and reject H_0 at level ν whenever the first or second stage data fall into R_ν. Obviously, always going to stage 2 is the "stopping rule" which maximizes the Type I error rate for all $\nu \in [0, 1]$ simultaneously. Hence, under this and any other stopping rule $S \in \{1, 2\}$, the H_0-probability that $q_S > \nu$ is at most ν.

By the same arguments it can be easily shown that reporting the minimum $\min\{q_1, q_2\}$ instead of q_2 at stage 2, also gives a valid p-value. However, in practice, one would not like to let q_1 overrule q_2, since the latter is based on more data than the first. Hence, one would always report q_2 at stage 2, even if q_1 is smaller.

The price for the flexibility with regard to the stopping rule is that the p-value q_S and even the less conservative p-value $\min\{q_1, q_2\}$ is strictly conservative whenever we do not follow the maximal rule of always going to stage 2. This is the case, in particular, if we follow the pre-specified stopping rule $S = 1 + \mathbf{1}_{\{\alpha_1 < p_1 \leq \alpha_0\}}$. Its strict conservatism is often viewed as a disadvantage of repeated p-values. The p-value based on the stage-wise ordering (see Sect. 8.1.1) does not have this property and is exact under the pre-specified stopping rule. However, the latter p-value does not allow us to deviate from the pre-specified stopping rule and does not have the monitoring property. One can show (see, for example, Brannath et al. 2003) that the conservatism of the repeated p-value cannot be removed without restrictions on the stopping rule. Hence the conservatism of the repeated confidence interval is an inevitable price that has to be paid for the flexibility with regard to the stopping rule.

Fisher's Product Test

Repeated p-values for Fisher's product test have not yet been considered in full generality. As a consequence, this section will contain some new suggestions. The question is how to choose $\alpha_{1,\nu}$, $\alpha_{0,\nu}$, and c_ν for $\nu \neq \alpha$ for which computations remain simple.

Assume at first that $\alpha_0 = 1$ and $\alpha_1 = c = \exp(-\chi^2_{4,1-\alpha}/2)$ where, as in Sect. 6.2.2, $\chi^2_{4,1-\alpha}$ is the $(1-\alpha)$-quantile of the χ^2-distribution with 4 df. In this case, a natural choice for the $\alpha_{i,\nu}$'s are the continuous and non-decreasing boundaries $\alpha_{0,\nu} = 1$ and $\alpha_{1,\nu} = c_\nu = \exp(-\chi^2_{4,1-\nu}/2)$ for all $\nu \in [0, 1]$. The resulting first and second stage p-values are

$$q_1 = 1 - F_{\chi^2_4}\big(-2\log(p_1)\big) \quad \text{and} \quad q_2 = 1 - F_{\chi^2_4}\big(-2\log(p_1 p_2)\big)\,, \tag{8.5}$$

where $F_{\chi^2_4}(\cdot)$ is the distribution function of the χ^2-distribution with $df = 4$. Similar p-values are used in meta-analyses. Note that q_2 is equal to $p_1 p_2 - p_1 p_2 \log(p_1 p_2)$ because $P(X \geq x) = \exp(-x/2)(1 + x/2)$ if X is χ^2-distributed with $df = 4$ (see Johnson and Kotz 1970; p.173). Therefore, if $p_1 p_2 > \alpha_1$ the repeated p-value of the second stage is equal to the overall exact p-value from (8.2).

If $\alpha_1 > c$ and/or $\alpha_0 < 1$, then there are several possibilities of extending the first stage levels to families $\alpha_{1,\nu} < \nu < \alpha_{0,\nu}$, $\nu \in [0,1]$. Often the first stage levels α_1 and α_0 are chosen according to some specific type of group sequential boundaries, for example, O'Brien and Fleming or other boundaries of the Wang and Tsiatis family, and the second stage level is then determined as

$$c = \frac{\alpha - \alpha_1}{\log(\alpha_0) - \log(\alpha_1)}$$

to meet level condition (6.4). As discussed in Sect. 6.2.2, the boundaries should satisfy the non-stochastic curtailment property $c \leq \alpha_1$. The same type of first stage boundaries could now be applied to levels ν for which

$$c_\nu = \frac{\nu - \alpha_{1,\nu}}{\log(\alpha_{0,\nu}) - \log(\alpha_{1,\nu})} \leq \alpha_{1,\nu} \ .$$

However, if $c_\nu > \alpha_{1,\nu}$, then we cannot use $\alpha_{1,\nu}$ as first stage rejection level because of the non-stochastic curtailment phenomenon. Instead, we can use the common first and second stage rejection boundary $c_\nu = \alpha_{1,\nu}$ which solves

$$c_\nu \cdot \big(1 + \log(\alpha_{0,\nu}) - \log(c_\nu)\big) = \nu \ . \tag{8.6}$$

The left side of Eq. (8.6) is the level of Fisher's combination test where first and second stage rejection boundaries are both equal to c_ν. The boundary c_ν can be determined by numerical root finding. Since the resulting first and second stage rejection boundaries may fail to be increasing in ν (but are continuous), a grid search must be applied to determine q_1 and q_2.

There is another method to extend α_1 and α_0 to families $\alpha_{1,\nu} < \nu < \alpha_{0,\nu}$ which can also be applied if the first stage levels are not chosen according to a specific type of group sequential boundaries and which guarantees increasing boundaries $\alpha_{0,\nu}$, $\alpha_{1,\nu}$, c_ν that satisfy the non-stochastic curtailment constraint $c_\nu \leq \alpha_{1,\nu}$ for all ν. The method is to fix the ratios $\eta_1 = \alpha_1/\alpha$ and $\eta_0 = \alpha_0/\alpha$ and to apply $\alpha_{1,\nu} = \nu\eta_1$, $\alpha_{0,\nu} = \nu\eta_0$ and

$$c_\nu = \frac{\nu - \alpha_{1,\nu}}{\log(\alpha_{0,\nu}) - \log(\alpha_{1,\nu})} = \frac{\nu(1-\eta_1)}{\log(\eta_0) - \log(\eta_1)} = \frac{\nu c}{\alpha}$$

for all ν where $\alpha_{0,\nu} = \nu\eta_0 < 1$. Obviously, the condition $\alpha_{0,\nu} < 1$ is equivalent to $\nu < \varpi_0 = \eta_0^{-1}$. Since $c < \alpha_1$ implies $c_\nu = \nu c/\alpha \leq \nu\alpha_1/\alpha = \alpha_{1,\nu}$, the above boundaries will automatically meet the non-stochastic curtailment constraint. Hence, level condition (8.3) is satisfied for all $\nu < \varpi_0$. For $\nu \geq \varpi_0$ we apply $\alpha_{0,\nu} = 1$, $\alpha_{1,\nu} = \nu\eta_1$, and $c_\nu = \nu(1-\eta_1)/\big(-\log(\nu\eta_1)\big)$ for all ν where the non-stochastic curtailment constraint $c_\nu < \alpha_{1,\nu} = \nu\eta_1$ is satisfied. The non-stochastic curtailment constraint can easily seen to be equivalent to $\nu < \varpi_1 = \eta_1^{-1}e^{(1-\eta_1^{-1})}$. Finally, for $\nu \geq \varpi_1$ we let $\alpha_{0,\nu} = 1$ and $\alpha_{1,\nu} = c_\nu = \exp(-\chi^2_{4,1-\nu}/2)$. In summary, we can use the boundaries

$$\alpha_{1,\nu} = \begin{cases} \nu\eta_1 & \text{if } \nu < \varpi_1 \\ \exp(-\chi^2_{4,\nu}/2) & \text{if } \nu \geq \varpi_1 \end{cases}, \qquad \alpha_{0,\nu} = \begin{cases} \nu\eta_0 & \text{if } \nu < \varpi_0 \\ 1 & \text{if } \nu \geq \varpi_0 \end{cases}$$

and

$$c_\nu = \begin{cases} \nu c/\alpha & \text{if } \nu < \varpi_0 \\ (1-\eta_1)\,\nu/\big(-\log(\eta_1\nu)\big) & \text{if } \varpi_0 \leq \nu < \varpi_1 \\ \exp(-\chi^2_{4,1-\nu}/2) & \text{if } \nu \geq \varpi_1 . \end{cases}$$

These boundaries are continuous and strictly increasing in ν by construction. If $\alpha_0 = 1$, then we can use the same boundaries but now with $\varpi_0 = 0$.

Since the above boundaries are increasing and continuous in ν, the repeated p-values can be determined by solving the equations $p_1 = \alpha_{1,\nu}$ and $C(p_1, p_2) = c_\nu$. This gives the first stage repeated p-value

$$q_1 = \begin{cases} p_1\eta_1^{-1} & \text{if } p_1 < \varpi_1\eta_1 \\ 1 - \chi^2_4\big(-2\log(p_1)\big) & \text{if } p_1 \geq \varpi_1\eta_1 . \end{cases}$$

If we let the function $h : x \in [0,\infty) \mapsto h(x) \in [0,1)$ be the inverse of the strictly increasing function $\nu \in [0,1) \mapsto \nu/\big(-\log(\nu)\big) \in [0,\infty)$, then the second stage repeated p-value becomes

$$q_2 = \begin{cases} \max\{p_1p_2\alpha/c, p_1\eta_0^{-1}\} & \text{if } p_1p_2 < \varpi_0 c/\alpha \\ \eta_1^{-1}h\big(p_1p_2\,/(\eta_1^{-1}-1)\big) & \text{if } \varpi_0 c/\alpha \leq p_1p_2 < \exp(-\chi^2_{4,1-\varpi_1}/2) \\ 1 - \chi^2_4\big(-2\log(p_1p_2)\big) & \text{if } p_1p_2 \geq \exp(-\chi^2_{4,1-\varpi_1}/2) . \end{cases}$$

The maximum in the first line of q_2 follows from the second stage rejection rule $p_1 \leq \alpha_{0,\nu}$ and $C(p_1, p_2) \leq c_\nu$ for levels $\nu < \varpi_0 = \alpha/\alpha_0$. If $\alpha_0 = 1$ and we use the above boundaries with $\varpi_0 = 0$, then we get the same repeated p-values but with $\varpi_0 = 0$ and hence without the first line in q_2.

8.1.3 Numerical Examples

We illustrate the different methods to calculate an overall p-value in a two-stage adaptive design by giving some numerical examples. We start by considering the inverse normal combination case. Suppose the critical values were chosen according to the Wang and Tsiatis class with $\Delta = 0.25$. With a one-sided significance level $\alpha = 0.025$ the constant $c_{\text{WT}} = 2.4239$ from Table 2.4 can be used because it is numerically identical to the two-sided constant for $\alpha = 0.05$. This yields the critical levels

$$\alpha_1 = 1 - \Phi(2.4239) = 0.00768 \quad \text{and}$$
$$c = 1 - \Phi(2.4239 \cdot 2^{-0.25}) = 1 - \Phi(2.0382) = 0.0208$$

for the test with inverse normal combination function and equal weights.

If the first stage yields a p-value that allows the rejection of H_0, i.e., $p_1 < 0.00768$, then the overall p-value that is based on the stage-wise ordering is p_1 itself.

If $p_1 > 0.00768$ and the inverse normal test statistic $C(p_1, p_2)$ yields the value, say, 0.023 (for example, $p_1 = 0.06$ and $p_2 = 0.1026$), with the use of the bivariate standard normal cdf $F(\cdot,\cdot)$ with correlation $1/\sqrt{2}$ (see Sect. 1.2) the overall exact p-value

$$0.00768 + P_{H_0}(p_1 > 0.00768, C(p_1, p_2) \leq 0.023)$$
$$= 1 - F(\Phi^{-1}(1 - 0.00768), \Phi^{-1}(1 - 0.023)) = 0.0271$$

is obtained.

The repeated p-value for the first stage is the smallest significance level for which the test allows the rejection of the null hypothesis at a given stage. For example, if p_1 is equal to the boundary 0.00768, the overall p-value is equal to α or, if the first stage p-value p_1 is 0.008, the overall p-value is 0.0258, and considerably larger than the first stage p-value itself. The value 0.0258 is obtained from the calculation of

$$1 - F(\Phi^{-1}(1 - 0.008), \Phi^{-1}(1 - 0.008) \cdot 2^{-0.25})$$

(see Sect. 4.1.2 and Sect. 8.1.2). The first stage repeated p-value is remarkably different to the p-value that is based on the stage-wise ordering. For the second stage the difference is smaller. The same second stage result as above produces the overall repeated p-value 0.0278 which is only slightly larger than 0.0271. This is obtained from

$$1 - F(\Phi^{-1}(1 - 0.023) \cdot 2^{0.25}, \Phi^{-1}(1 - 0.023)) .$$

Note that both types of overall p-value are consistent with the test decision, i.e., the overall p-value is smaller than α if and only if the test allows the rejection of H_0 at a given stage. The repeated p-value has the advantage that it can be calculated at both stages and also in the case that the first stage did not produce a significant test result.

For Fisher's combination test the calculations are as follows. We only consider the simplest case with $\alpha_0 = 1$ and $\alpha_1 = c = \exp(-\chi^2_{4,1-\alpha}/2)$. For $\alpha = 0.025$ this yields $\alpha_1 = c = 0.0038$.

If the first stage test result allows the rejection of H_0, i.e., $p_1 < 0.0038$, then the overall p-value that is based on the stage-wise ordering is p_1 itself.

If $p_1 > 0.0038$ and, as above, $p_1 = 0.06$ and $p_2 = 0.1026$, the Fisher's combination function $C(p_1, p_2)$ is $p_1 p_2 = 0.00616 > 0.0038 = \alpha_1$ and formula (8.2) yields the overall p-value

$$p_1p_2 - p_1p_2 \log(p_1p_2) = 0.0374 \, .$$

Using (8.5) the repeated p-values for Fisher's combination test for $p_1 = 0.06$ and $p_2 = 0.1026$ are

$$q_1 = 0.2288 \quad \text{and} \quad q_2 = 0.0374 \, .$$

Again, both can be interpreted as the smallest significance levels for which the test rejects H_0 at a given stage. As already mentioned, in this case ($p_1p_2 > c$) the exact overall and repeated overall p-values at stage 2 coincide. If $p_1p_2 \leq c$ the exact p-value that is based on the stage-wise ordering is always larger than the repeated p-value. This is because $x - x\log(x) \leq \alpha_1 - x\log(\alpha_1)$ if $x \leq \alpha_1$.

8.2 Adaptive Confidence Intervals

According to the ICH guideline E9 on statistical principles for clinical trials, estimates and confidence intervals should be provided in addition to p-values at least for the primary efficacy parameters. The usual confidence intervals which ignore the sequential and adaptive nature of the trial will in general be invalid and hence cannot be applied. Like for group sequential designs, however, valid confidence intervals can be constructed by applying a suitable family of adaptive hypothesis tests to all possible parameter values. Again, there is no unique or most natural family of such hypothesis tests, and hence several construction methods are possible (see Chap. 4).

We will discuss two approaches for the construction of one-sided confidence intervals for two-stage combination tests. These approaches follow similar methods for group sequential designs as described in Chap. 4. The first method provides confidence intervals with exact coverage probabilities (under specific assumptions) that can be reported at the end of the adaptive trial when following the pre-specified stopping rule. The second method yields repeated confidence intervals which can be reported at each stage of the trial independently of when and for which reason the trial has been stopped.

To discuss the construction of confidence intervals we need to make additional assumptions which are frequently satisfied. We assume a real valued efficacy parameter θ such that $\theta > 0$ if the treatment meets the target goal (typically efficacy or non-inferiority) and the null hypothesis $H_0 : \theta \leq 0$ is the one of primary interest. A typical example is $\theta = \mu_2 - \mu_1$ where μ_1 and μ_2 are the means of normally distributed responses in a control and treatment group, respectively, where the goal is to verify that $\mu_2 > \mu_1$, and the primary null hypothesis is $H_0 : \mu_2 \leq \mu_1$. If the goal is to show non-inferiority with non-inferiority margin $\theta_0 < 0$, i.e., to test the null hypothesis $H_0 : \mu_2 - \mu_1 \leq \theta_0$, then we define $\theta = \mu_2 - \mu_1 - \theta_0$ such that the null hypothesis can again be written as $H_0 : \theta \leq 0$.

We will construct confidence intervals by consideration of all null hypotheses $H_0^\delta : \theta \leq \delta$ with $-\infty < \delta < \infty$ exploiting the duality between confidence intervals and hypothesis tests. In most parts of this section, we will assume that each H_0^δ (including H_0) is tested using one-sided and stage-wise p-values $p_{k,\delta} = 1 - \Phi((\hat{\theta}_k - \delta)/\mathrm{se}_k)$, where $k = 1, 2$ indicate the stage, $\hat{\theta}_k$ is an estimate of θ from stage k, and se_k is the standard error of $\hat{\theta}_k$ or an estimate thereof. The estimate $\hat{\theta}_k$ and eventually also se_k are assumed to be computed from independent cohorts of patients. Note that the p-values $p_{k,\delta}$ are non-decreasing in δ at every sample point, i.e., $p_{k,\delta} \leq p_{k,\delta'}$ for all $\delta \leq \delta'$ and all trial outcomes. The latter property of $p_{1,\delta}$ and $p_{2,\delta}$ is the crucial one, and most of the constructions below will hold when this property is satisfied.

8.2.1 *Exact Confidence Bounds for Combination Tests*

In Sect. 4.1 we have seen how the stage-wise ordering can be used to define confidence intervals for group sequential trials that have exact coverage probability and are consistent with the final test decision. In this section, we use a slight modification of the stage-wise ordering for combination tests (see Sect. 8.1.1) to define a consistent and exact lower confidence bound $l_{k,\theta}^e$ which can be reported at the end of the confirmatory adaptive trial.

As mentioned before, the construction of $l_{k,\theta}^e$ will be based on stage-wise p-values $p_{k,\delta}$. The confidence intervals $(l_{k,\theta}^e; \infty)$ will be *"exact"* in the sense that the coverage probability of $(l_{k,\theta}^e; \infty)$ is exactly equal to $1 - \alpha$ if under $\theta = \delta$ the p-values $p_{1,\delta}$, $p_{2,\delta}$ are independent and uniformly distributed. The latter property will usually not be satisfied for finite samples, however, in many examples it is valid asymptotically when the number of observations (of both stages) converges to infinity. In this case the p-values $p_{1,\delta}$, $p_{2,\delta}$ will be approximately independent and uniformly distributed when the sample sizes are large which is usually the case in confirmatory trials. Furthermore, we will focus on confidence intervals that are consistent with the combination test in the sense that $l_{k,\theta}^e > 0$ if and only if the combination test rejects H_0 at stage k.

As before, we consider the null hypothesis $H_0 : \theta \leq 0$ and an adaptive design with combination function $C(p_1, p_2)$ and decision boundaries α_1, α_0, and c meeting level condition (6.1). Recall that the stage-wise ordering from Sect. 8.1.1 considers a sample point x' to provide more evidence against H_0 than the sample point x if

1. $p_1 \leq \alpha_1$ or $p_1 > \alpha_0$, and $p_1' \leq p_1$, or
2. $\alpha_1 < p_1 \leq \alpha_0$ and $p_1' \leq \alpha_1$, or
3. $\alpha_1 < p_1 \leq \alpha_0$ and $\alpha_1 < p_1' \leq \alpha_0$ and $C(p_1', p_2') \leq C(p_1, p_2)$.

Brannath et al. (2003) used a modification of this ordering for H_0^δ with $\delta \neq 0$. The ordering "$\succeq_\delta$" for H_0^δ is defined like the ordering "$\succeq$" for H_0 but with

3. replaced by the condition

4. $\alpha_1 < p_1 \leq \alpha_0$ and $\alpha_1 < p_1' \leq \alpha_0$ and $C(p'_{1,\delta}, p'_{2,\delta}) \leq C(p_{1,\delta}, p_{2,\delta})$. Note that this means to replace p_k', p_k by $p'_{k,\delta}, p_{k,\delta}$ in the combination function but to keep p_k', p_k in all other conditions. The reason for the replacement is that it is natural to apply the combination function $C(x, y)$ to $p_{k,\delta}$ instead of p_k when testing H_0^δ because $p_{k,\delta}$ is known to be (asymptotically) uniformly distributed under $\theta = \delta$. The reason for keeping p_k' and p_k in all other conditions is to achieve exact coverage.

It is interesting to note that 3. and 4. are equivalent for the weighted z-score or inverse normal combination method when applied to stage-wise z-scores $z_{k,\delta} = (\hat{\theta}_k - \delta)/\mathrm{se}_k$ or p-values $p_{k,\delta} = 1 - \Phi(z_{k,\delta})$. This is because

$$\begin{aligned} \tilde{z}_{2,\delta} &= w_1 z_{1,\delta} + w_2 z_{2,\delta} = w_1 z_1 + w_2 z_2 - \left(\frac{w_1}{\mathrm{se}_1} + \frac{w_2}{\mathrm{se}_2}\right)\delta \\ &= \tilde{z}_2 - \left(\frac{w_1}{\mathrm{se}_1} + \frac{w_2}{\mathrm{se}_2}\right)\delta\ , \end{aligned} \tag{8.7}$$

and hence the z-scores $\tilde{z}_2$ and $\tilde{z}_{2,\delta}$ (and corresponding inverse normal combination functions) imply the same second stage ordering. As a consequence, all orderings "$\succeq_\delta$" are identical to the original ordering "$\succeq$" from Sect. 8.1.1. Another important consequence is that when the weights w_k are chosen according to pre-planned sample sizes (or information numbers) and the samples sizes are as pre-planned, then the exact confidence interval of the combination test will coincide with the exact confidence interval of the group sequential test based on the stage-wise ordering (see Sect. 4.1).

Let us now turn back to general combination tests. Let $p_{k,\delta} = 1 - \Phi((\hat{\theta}_k - \delta)/\mathrm{se}_k)$ and define

$$\alpha_{1,\delta} = 1 - \Phi(\Phi^{-1}(1 - \alpha_1) - \delta/\mathrm{se}_1) \quad \text{and} \quad \alpha_{0,\delta} = 1 - \Phi(\Phi^{-1}(1 - \alpha_0) - \delta/\mathrm{se}_1)\ .$$

One can easily show that then 1., 2. and 4. are equivalent to

1.′ $p_{1,\delta} \leq \alpha_{1,\delta}$ or $p_{1,\delta} > \alpha_{0,\delta}$, and $p'_{1,\delta} \leq p_{1,\delta}$,
2.′ $\alpha_{1,\delta} < p_{1,\delta} \leq \alpha_{0,\delta}$ and $p'_{1,\delta} \leq \alpha_{1,\delta}$,
4.′ $\alpha_{1,\delta} < p_{1,\delta} \leq \alpha_{0,\delta}$ and $\alpha_{1,\delta} < p'_{1,\delta} \leq \alpha_{0,\delta}$ and $C(p'_{1,\delta}, p'_{2,\delta}) \leq C(p_{1,\delta}, p_{2,\delta})$,

whereby all conditions are now formulated in terms of $p_{k,\delta}$. This shows that $P_\delta(x' \succeq_\delta x) = Q_\delta(p_{1,\delta}, p_{2,\delta})$ for the p-value function

$$Q_\delta(p_{1,\delta}, p_{2,\delta}) = \begin{cases} p_{1,\delta} & \text{if } p_1 \leq \alpha_1 \text{ or } p_1 > \alpha_0 \\ \alpha_{1,\delta} + \int_{\alpha_{1,\delta}}^{\alpha_{0,\delta}} \int_0^1 \mathbf{1}_{\{C(x,y) \leq C(p_{1,\delta}, p_{2,\delta})\}}\, dy\, dx & \text{if } \alpha_1 < p_1 \leq \alpha_0\ . \end{cases} \tag{8.8}$$

In the expressions $P_\delta(x' \succeq_\delta x)$, x denotes the observed sample point, and x' a random variable that represents the distribution of the data under $\theta = \delta$. Correspondingly, the observed p-values $p_{1,\delta}$, $p_{2,\delta}$ for H_0^δ are considered as fix numbers, whereas the

p-values $p'_{1,\delta}, p'_{2,\delta}$ of the hypothetical sample point x' are independent and uniformly distributed random variables.

Note that the right side of (8.8) equals the overall exact p-value (8.1) of the combination test for H_0^δ with combination function $C(x, y)$ and early rejection and acceptance boundaries $\alpha_{1,\delta}$ and $\alpha_{0,\delta}$, applied to the stage-wise p-values $p_{1,\delta}$ and $p_{0,\delta}$. Note that all $\alpha_{1,\delta}$ and $\alpha_{0,\delta}$ are (or can be viewed as) fixed numbers if se_1 is fixed or independent from $p_{1,\delta}$. In this case, we do apply a combination test if we reject H_0^δ at level α when $P_\delta(x' \succeq_\delta x) \le \alpha$. Consequently, the region $\{\delta : P_\delta(x' \succeq_\delta x) > \alpha\}$ is a $(1-\alpha)$-confidence region for the parameter θ. The phrase "adaptive" means that the coverage probability is at least $1-\alpha$ for every adaptation rule where the second stage sample size is chosen based on the (fully unblinded) interim data or any other information from the interim analysis.

The independence of p_1 and se_1 may not hold exactly but is typically satisfied approximately for sufficiently large interim sample sizes, and then the coverage probability is expected to be close to $1-\alpha$. A thorough investigation of the impact of a dependence between p_k and se_k on the coverage probability has, however, not been undertaken yet.

Since $\alpha_{1,\delta}$ is increasing in δ, and

$$\alpha_{1,\delta} + \int_{\alpha_{1,\delta}}^{\alpha_{0,\delta}} \int_0^1 \mathbf{1}_{\{C(x,y) \le C(p_{1,\delta}, p_{2,\delta})\}}\, dy\, dx = \iint_{R_\delta} dy\, dx$$

for increasing regions $R_\delta = \{x \le \alpha_{1,\delta}\} \cup \{x \le \alpha_{0,\delta},\ C(x,y) \le C(p_{1,\delta}, p_{2,\delta})\}$, the right side of (8.8) is increasing in δ at all sample points x. Hence, the adaptive confidence region $\{\delta : P_\delta(x' \succeq_\delta x) > \alpha\}$ is a one-sided interval $(l^e_{2,\theta}; \infty)$ where the lower boundary $l^e_{2,\theta}$ is the unique solution of the equation $P_\delta\{x' \succeq_\delta x\} = \alpha$ in δ. This solution can be determined by numerical integration, to compute the right side of (8.8), and numerical root finding.

If the trial stops at the first stage ($p_1 \le \alpha_1$ or $p_1 > \alpha_0$), then the lower confidence bound $l^e_{1,\theta}$ is the solution of $p_{1,\delta} = \alpha$ which is the usual $(1-\alpha)$-lower confidence bound computed from the first stage sample, namely $l^e_{1,\theta} = \hat{\theta}_1 - \Phi^{-1}(1-\alpha)\,\mathrm{se}_1$.

An important point to be noticed is that the lower confidence bound $l^e_{k,\theta}$ can be computed only if the experimenter adheres to the pre-specified stopping rule, i.e., he stops the trial at the interim analysis if and only if either $p_1 \le \alpha_1$ or $p_1 > \alpha_0$. This is because the order relation "$x' \succ_\delta x$" remains undefined for sample points x that violate the stopping rule.

Let us finally consider the special case where $\alpha_1 = 0, \alpha_0 = 1$ and the trial always continues to the second stage. In this case, we can set $\alpha_{1,\delta} = 0$ and $\alpha_{0,\delta} = 1$ for all δ and then get the overall p-values $Q_\delta(p_{1,\delta}, p_{2,\delta}) = \int_0^1 \mathbf{1}_{\{C(x,y) \le C(p_{1,\delta}, p_{2,\delta})\}} dy\, dx$. This together with level condition (6.1) implies that $Q_\delta(p_{1,\delta}, p_{2,\delta}) = \alpha$ is equivalent to $C(p_{1,\delta}, p_{2,\delta}) = c$. Hence, solving the latter equality, which does not require numerical integration, is a simpler alternative to obtain the exact lower bound $l^e_{2,\theta}$ when $\alpha_1 = 0$ and $\alpha_1 = 1$.

8.2.2 Repeated Confidence Bounds for Combination Tests

Lehmacher and Wassmer (1999) and Brannath et al. (2002) adopted the repeated confidence interval approach to combination tests. Recall from Sect. 4.1 that repeated confidence intervals are obtained by applying the same group sequential boundaries to all hypotheses H_0^δ but to the shifted sequential z-statistics $\tilde{z}_{k,\delta} = \tilde{z}_k - \delta\sqrt{I_k}$. A similar idea can be applied to combination tests, namely to apply the same first stage rejection and acceptance levels α_1, α_0, combination function $C(x, y)$ and critical value c, to all H_0^δ and corresponding shifted p-values $p_{k,\delta} = 1 - \Phi((\hat{\theta}_k - \delta)/\mathrm{se}_k)$.

Following this approach, we reject H_0^δ at the first stage if $p_{1,\delta} \leq \alpha_1$. This gives the one-sided first stage repeated confidence interval

$$(l_{1,\theta}^r; \infty) \quad \text{with} \quad l_{1,\theta}^r = \hat{\theta}_1 - \Phi^{-1}(1 - \alpha_1)\,\mathrm{se}_1 \ .$$

Note that $l_{1,\theta}^r$ is like the usual lower confidence bound, but at adjusted level $1 - \alpha_1$ and computed from the first stage data only.

At the second stage we reject H_0^δ if $p_{1,\delta} \leq \alpha_0$ and $C(p_{1,\delta}, p_{2,\delta}) \leq c$. If $\alpha_0 < 1$, then solving $p_{1,\delta} = \alpha_0$ gives the lower bound

$$\tilde{l}_{0,\theta} = \hat{\theta}_1 - \Phi^{-1}(1 - \alpha_0)\mathrm{se}_1 \ . \tag{8.9}$$

Since $p_{k,\delta}$ is increasing in δ, and $C(x, y)$ is increasing in both arguments x and y, the equation $C(p_{1,\delta}, p_{2,\delta}) = c$ in δ has a unique solution which we denote by $\tilde{l}_{2,\theta}$, and which can easily be determined by numerical root finding if no analytical solution is available. Hence, the second stage repeated confidence set, which consists of all δ where $p_{1,\delta} < \alpha_0$ or $C(p_{1,\delta}, p_{2,\delta}) > c$, is the interval

$$(l_{2,\theta}^r; \infty) \quad \text{where} \quad l_{2,\theta}^r = \begin{cases} \tilde{l}_{2,\theta} & \text{if } \alpha_0 = 1 \\ \min\{\tilde{l}_{0,\theta}, \tilde{l}_{2,\theta}\} & \text{if } \alpha_0 < 1 \ . \end{cases}$$

Note that when $\alpha_1 = 0$, $\alpha_0 = 1$ and the trial is never stopped at stage 1, then the repeated confidence bound l_θ^r coincides with the solution $\tilde{l}_{2,\theta}$ of $C(p_{1,\delta}, p_{2,\delta}) = c$, and also with the exact lower confidence bound l_θ^e from the previous subsection, which is then also given by $\tilde{l}_{2,\theta}$. Hence, the repeated confidence interval has exact coverage probability when early stopping is not an option.

Repeated Confidence Interval for Inverse Normal Method

The solution $\tilde{l}_{2,\theta}^r$ of $C(p_{1,\delta}, p_{2,\delta}) = c$ becomes explicit if the combination function is of inverse normal or weighted z-score type,

$$C(p_1, p_2) = 1 - \Phi\big(w_1\Phi^{-1}(1 - p_1) + w_2\Phi^{-1}(1 - p_2)\big) \ ,$$

and the shifted p-values are of the common form

$$p_{k,\delta} = 1 - \Phi\left(\frac{\hat{\theta}_k - \delta}{\mathrm{se}_k}\right) \ , \ k = 1, 2 \ ,$$

where $\hat{\theta}_k$ is an estimates of θ, and se_k is the standard error estimate of $\hat{\theta}_k$, both computed from the data of the individuals recruited for stage k. By simple algebra one can see that

$$\tilde{l}^r_{2,\theta} = \tilde{\theta} - \Phi^{-1}(1-c) / \left(\frac{w_1}{\mathrm{se}_1} + \frac{w_2}{\mathrm{se}_2}\right) \ \text{ with } \ \tilde{\theta} = \left(\hat{\theta}_1 \frac{w_1}{\mathrm{se}_1} + \hat{\theta}_2 \frac{w_2}{\mathrm{se}_2}\right) / \left(\frac{w_1}{\mathrm{se}_1} + \frac{w_2}{\mathrm{se}_2}\right) . \tag{8.10}$$

An interesting special case of (8.10) is obtained when $\alpha_1 = 0$ and $\alpha_0 = 1$, i.e., no rejection or acceptance is possible at the interim analysis. In this case $\Phi^{-1}(1-c)$ equals the $(1-\alpha)100\,\%$-percentile $u_{1-\alpha}$ of the standard normal distribution, since $z = w_1 \Phi^{-1}(1-p_1) + w_2 \Phi^{-1}(1-p_2)$ is standard normally distributed for independent and uniformly distributed p-values p_1, p_2. Hence the lower confidence bound (8.10) becomes the standard lower confidence bound $\tilde{l}^r_{2,\theta} = \tilde{\theta} - u_{1-\alpha} \mathrm{se}_{\tilde{\theta}}$ of the normally distributed estimate $\tilde{\theta}$ that has standard error $\mathrm{se}_{\tilde{\theta}} = (w_1/\mathrm{se}_1 + w_2/\mathrm{se}_2)^{-1}$. Recall that this lower bound is also exact (when $\alpha_1 = 0$ and $\alpha_0 = 1$), i.e., it provides an interval with exact coverage probability. We will later discuss the properties of $\tilde{\theta}$ as point estimate for θ. This computation of the RCI can also be extended to the case where an early rejection is possible. Simply replace the $(1-\alpha)100\,\%$-percentile $u_{1-\alpha}$ by the corresponding adjusted critical level (Lehmacher and Wassmer 1999).

Monitoring Property

The repeated confidence interval has the same monitoring properties as its group sequential counter part. This means that we can report at each stage k the confidence interval $(l^r_{k,\theta}; \infty)$, irrespective of the stage and the reason the trial will or has been stopped. Like for group sequential tests, this follows from the correct coverage probability of the intersection of the first and second stage interval.

8.2.3 Two-Sided Confidence Intervals

One-sided intervals that provide a lower efficacy bound will be sufficient for most confirmatory clinical trials. However, there might be cases where two-sided intervals are required. Examples are (among others) equivalence trials or post-marketing superiority trials where two proved treatments are compared with regard to efficacy. In such trials the major goal is to test a hypothesis $H_0 : \theta = 0$ against the two-sided alternative $H_1 : \theta \neq 0$. It is also quite natural and usual to provide a finite interval $(l; u)$ with a lower and an upper bound l and u, respectively, in a report or an article. If a one-sided test at level α is performed, a consistent two-sided confidence interval would be a $(1-2\alpha)100\,\%$ interval. Particularly, it is widely accepted to test at a one-sided significance level 2.5 % but to provide two-sided 95 % confidence intervals.

A straightforward way to derive these intervals is to consider an adaptive two-stage design as two one-sided adaptive tests for

$$H_0^{(-)} : \theta \leq 0 \qquad \text{and} \qquad H_0^{(+)} : \theta \geq 0$$

(see Sect. 6.4). The calculation of the lower bound is performed as described above in this section, whereas the calculation of the upper bound is achieved through the calculation of the p-value for $H_0^{(+)}$ and applying the same principles as described for the p-values for $H_0^{(-)}$. We will illustrate this by a numerical example in the following section.

8.2.4 Numerical Examples

We illustrate the described ways to calculate exact confidence intervals for the inverse normal case. As above, we use the critical values according to the Wang and Tsiatis class with $\Delta = 0.25$ and a one-sided significance level $\alpha = 0.025$. These are given by $\alpha_1 = 0.00768$ and $c = 0.0208$ (or $u_1 = 2.4239$ and $u_2 = 2.0382$) for the test with inverse normal combination function and equal weights. We consider testing the mean μ of normally distributed observations with the variance, σ^2, assumed to be known. We assume that the study was planned with sample size $n_1 = n_2 = 20$ per stage of the trial for testing $H_0 : \mu < 0$, i.e., $w_1 = w_2 = 1/\sqrt{2}$. For testing

$$H_0^{\delta} : \mu \leq \delta \quad \text{against} \quad H_1^{\delta} : \mu > \delta\ ,$$

the shifted p-values

$$p_{k,\delta} = 1 - \Phi\Big(\frac{\bar{x}_k - \delta}{\sigma}\sqrt{\tilde{n}_k}\Big)\ ,\ k = 1, 2\ ,$$

where $\tilde{n}_k$ are the realized sample sizes per stage, are exactly uniformly distributed under H_0^{δ}. Note that always $\tilde{n}_1 = n_1$ and thus the first stage sample size should be unchanged.

If the trial stops at the first stage with the rejection of the null hypothesis, i.e., $p_{1,0} \leq 0.00768$, the lower confidence bound is the solution of $p_{1,l_{1,\delta}^e} = \alpha$ and is given by

$$l^e_{1,\delta} = \bar{x}_1 - \Phi^{-1}(1-\alpha)\frac{\sigma}{\sqrt{\tilde{n}_1}} ,$$

which is the usual $(1-\alpha)100\,\%$ lower confidence bound.

If the trial proceeds to the second stage, from (8.7) and (8.8) with $\alpha_{0,\delta} = 1$ we have to solve in δ the equation

$$\begin{aligned}
&\alpha_{1,\delta} + \int_{\alpha_{1,\delta}}^{1}\int_{0}^{1} \mathbf{1}_{\{C(x,y)\le C(p_{1,\delta},p_{2,\delta})\}} dy\, dx \\
&= P(Z_1 \ge u_1 - \frac{\delta}{\sigma}\sqrt{\tilde{n}_1}) + P\left(Z_1 < u_1 - \frac{\delta}{\sigma}\sqrt{\tilde{n}_1}, \tilde{Z}_2 \ge \tilde{z}_2 - \frac{\delta}{\sigma}\frac{\sqrt{\tilde{n}_1}+\sqrt{\tilde{n}_2}}{\sqrt{2}}\right) \\
&= 1 - F\left(u_1 - \frac{\delta}{\sigma}\sqrt{\tilde{n}_1}, \tilde{z}_2 - \frac{\delta}{\sigma}\frac{\sqrt{\tilde{n}_1}+\sqrt{\tilde{n}_2}}{\sqrt{2}}\right) = \alpha , \qquad (8.11)
\end{aligned}$$

where $\tilde{z}_2 = (\Phi^{-1}(1-p_{1,0}) + \Phi^{-1}(1-p_{2,0}))/\sqrt{2}$ and $F(\cdot,\cdot)$ is the bivariate standard normal cdf with correlation $1/\sqrt{2}$.

For example, with $\bar{x}_1 = 0.32, \bar{x}_2 = 0.35, \sigma^2 = 1$ and $\tilde{n}_1 = 20$ and $\tilde{n}_2 = 20$ (i.e., no sample size change, and the observed sample sizes are exactly the planned ones) one finds $p_{1,0} = 0.0762, p_{2,0} = 0.0588$, and $\tilde{z}_2 = 2.1187$. Solving (8.11) yields $l^e_{2,\delta} = 0.0102$. A corresponding upper bound is found by solving

$$F\left(u_1 - \frac{\delta}{\sigma}\sqrt{\tilde{n}_1}, \tilde{z}_2 - \frac{\delta}{\sigma}\frac{\sqrt{\tilde{n}_1}+\sqrt{\tilde{n}_2}}{\sqrt{2}}\right) = \alpha$$

for δ which yields the value 0.641. So the two-sided 95 % confidence interval that is based on the stage-wise ordering is given by (0.0102; 0.641).

In this case, the calculation is identical to the calculation of a confidence interval that is based on the stage-wise ordering in the classical group sequential design (see Sect. 4.1.1). This is because the sample sizes are equal to the planned ones. Importantly, the calculation can be extended here to the adaptive case where the sample size for the second stage can be recalculated in a data-driven way. For example, assume that the second stage sample size was changed to, say, $\tilde{n}_2 = 60$. The same observed mean values for the first and second stage as above yield the 95 % confidence interval (0.0738; 0.5650) which is different to the confidence interval that is calculated under the assumption that $\tilde{n}_1$ and $\tilde{n}_2$ were the originally assumed sample sizes.

We now consider the calculation of the repeated confidence intervals. Since $\mathrm{se}_k = \sigma/\sqrt{n_k}$, from (8.10) the formulae for the lower repeated confidence bounds are

$$l^r_{1,\delta} = \bar{x}_1 - u_1\frac{\sigma}{\sqrt{\tilde{n}_1}} \quad \text{and}$$

$$l^r_{2,\delta} = \frac{\sqrt{\tilde{n}_1}\bar{x}_1 + \sqrt{\tilde{n}_2}\bar{x}_2}{\sqrt{\tilde{n}_1}+\sqrt{\tilde{n}_2}} - u_2\frac{\sqrt{2}\sigma}{\sqrt{\tilde{n}_1}+\sqrt{\tilde{n}_2}} ,$$

where, as above, $\tilde{n}_k$ refer to the realized sample sizes at stage k, and $u_2 = \Phi^{-1}(1-c)$. For $\tilde{n}_1 = \tilde{n}_2 = 20$, this yields

$$l^r_{1,\delta} = -0.222 \quad \text{and} \quad l^r_{2,\delta} = 0.0127 \,,$$

whereas for $\tilde{n}_1 = 20$ and $\tilde{n}_2 = 60$ we obtain

$$l^r_{1,\delta} = -0.222 \quad \text{and} \quad l^r_{2,\delta} = 0.103 \,, \text{ respectively} \,.$$

The corresponding upper bounds are simply

$$u^r_{1,\delta} = \bar{x}_1 + u_1 \frac{\sigma}{\sqrt{\tilde{n}_1}} \quad \text{and}$$
$$u^r_{2,\delta} = \frac{\sqrt{\tilde{n}_1}\bar{x}_1 + \sqrt{\tilde{n}_2}\bar{x}_2}{\sqrt{\tilde{n}_1} + \sqrt{\tilde{n}_2}} + u_2 \frac{\sqrt{2}\sigma}{\sqrt{\tilde{n}_1} + \sqrt{\tilde{n}_2}} \,,$$

and for $\tilde{n}_1 = 20$, $\tilde{n}_2 = 20$ and $\tilde{n}_2 = 60$ given by

$$u^r_{1,\delta} = 0.862 \,,$$
$$u^r_{2,\delta} = 0.657 \quad \text{and} \quad u^r_{2,\delta} = 0.575 \,, \text{ respectively} \,.$$

So the first stage repeated confidence interval is considerably larger than the exact confidence interval from above. For the second stage, the lower bounds are larger for the repeated confidence intervals, for $\tilde{n}_2 = 20$ the length of the repeated confidence interval is slightly larger, too, but for $\tilde{n}_2 = 60$ the length of the repeated confidence interval is slightly smaller than the exact confidence interval.

We finally consider the case, where a (binding) futility bound was included for the first stage. Suppose it was decided to stop the trial for futility if $p_1 > 0.30$. The critical values are found to be $u_1 = 2.4006$ and $u_2 = 2.0187$. For $\tilde{n}_1 = 20$ and $\tilde{n}_2 = 20$ the two-sided repeated confidence intervals are

$$(-0.217;\ 0.856) \quad \text{and} \quad (0.0158;\ 0.654)$$

for the first and the second stage, respectively. That is, both are smaller than those where no stopping for futility was considered. Note that the second stage lower bound is bounded by (8.9) which is

$$\tilde{l}_{\delta,0} = \bar{x}_1 - \Phi^{-1}(1-\alpha_0)\frac{\sigma}{\sqrt{\tilde{n}_1}} = 0.203 \,.$$

Specifically, there is a bound for the second stage mean value such that at given $\tilde{n}_2$ the lower bound of the repeated confidence interval does not change any more. In our case, this is the value 0.724 and the interval for $\bar{x}_2 = 0.724$ is (0.203; 0.841). For $\bar{x}_2 = 2$, say, the interval is (0.203; 1.48) and so only the upper bound changes for $\bar{x}_2 > 0.724$. This is the "price" for using somewhat smaller critical values u_1 and u_2 for the calculation of the confidence intervals at the two stages of the trial. Note that this might become problematic if the critical bound that is adjusted for the futility stop becomes smaller than the usual critical level derived for the fixed sample case. In this case the RCI at the end of the trial would be smaller than the usual $(1-\alpha)100\,\%$ fixed sample confidence interval.

8.3 Point Estimation in Adaptive Designs

Because of the stopping rule and the data-driven sample size in an adaptive design, we must expect that the usual maximum likelihood estimate (ML-estimate) is biased and might therefore be an inappropriate estimate. Furthermore, the ML-estimate is not necessarily contained in an adaptive confidence interval like the ones discussed in the previous section. For these reasons, several alternative estimates have been suggested in the literature. We will review in this section the most important suggestions and compare them to the ML-estimate with regard to bias and mean square error (MSE).

In most of the literature (including regulatory guidelines) large emphasis is put on bias of point estimates, in particular, when dealing with sequential or adaptive designs. "Bias" is typically quantified as deviation of the mean or median of the estimate from the true parameter value. Bias is an important concept as it describes the systematic deviation of the estimate from the truth parameter value. Variance is another important quantity which also effects estimation precision, often more severe than bias. Hence, a "good" point estimate should not only have a reasonably small bias but also a reasonably small variance. A mean (or median) unbiased estimate with a large variance is similarly inappropriate as an estimate with a large bias. For instance, the mean of the first stage data alone, which is computed from a sample with pre-specified sample size, is mean and median unbiased, but is hardly acceptable as final estimate, because a large part of the data is disregarded and hence the variance is unreasonably large.

Most methods to remove bias will lead to an increase in variance (compared to the ML-estimate). Hence, bias reduction methods must carefully be assessed with regard to their overall precision, as, for example, quantified by the MSE, which is the sum of variance and square of mean bias. For these reasons, we will focus on the MSE (or its square root) in our comparison and its conclusions (see also the remark at the end of Sect. 4.2), but will also consider bias.

8.3.1 Maximum Likelihood Estimate

The likelihood of the data of a two-stage adaptive design is $\prod_{i=1}^{n_1+\tilde{n}_2} f_\theta(x_i)$ where $f_\theta(x_i)$ is the density of the i-th observation, n_1 the first stage sample size, and $\tilde{n}_2$ is the adaptively chosen second stage sample size. In the simple case of estimating the mean $\theta = \mu$ of a normal response with variance σ^2, the ML-estimate is the overall mean $\bar{x} = (n_1\bar{x}_1 + \tilde{n}_2\bar{x}_2)/(n_1 + \tilde{n}_2) = (\bar{x}_1 + \tilde{r}\bar{x}_2)/(1 + \tilde{r})$ where $\bar{x}_k$ is the mean of the data from stage k and $\tilde{r} = \tilde{n}_2/n_1$ the second stage sample size as multiple of the first stage sample size n_1.

If there is a stopping rule and/or the sample size is reassessed based on the first stage data, then $\bar{x}$ will in general exhibit some type of mean (and median) bias. Setting $v = 1/(1 + \tilde{r})$ and recognizing that the conditional mean of $\bar{x}_2$ given $\bar{x}_1$ and $\tilde{n}_2$ is $E_\mu(\bar{x}_2|\bar{x}_1, \tilde{n}_2) = \mu$, the mean bias of $\bar{x}$ is given by

$$\begin{aligned} E_\mu(\bar{x}) - \mu &= E_\mu\big(v\bar{x}_1\big) + E_\mu\big((1 - v)\bar{x}_2\big) - \mu \\ &= E_\mu\big(\tilde{v}\,(\bar{x}_1 - \mu)\big) = \mathrm{Cov}_\mu(\tilde{v}, \bar{x}_1) = \mathrm{Cov}_\mu\left(\frac{1}{1 + \tilde{r}}, \bar{x}_1\right) \end{aligned} \tag{8.12}$$

(Liu et al. 2002). This formula is helpful for the understanding of the effects of adaptive sample size reassessments on the mean of the ML-estimate.

If, for example, the sample size $\tilde{n}_2$ becomes smaller for increasing $\bar{x}_1$, i.e., the covariance $\mathrm{Cov}_\mu\big(1/(1 + \tilde{r}), \bar{x}_1\big)$ is positive, then the bias will be positive as well. A negative relationship between second stage sample size and first stage estimate occurs, for instance, in group sequential designs where the trial is stopped early if a sufficient large first stage estimate is observed. It also occurs when the second stage sample size is reassessed for achieving a target conditional power of, say, 90 %. This is because the conditional power is larger for larger first stage estimates, see Sect. 7.1.

In other examples, the relationship between first stage estimate and second stage sample size is positive such that the mean bias is negative. A positive relationship between sample size and first stage estimate is typically caused by futility rules where the trial is stopped early when the first stage estimate is small, or by treatment selection where treatment arms are terminated when the treatment effect is too small. Often the dependence between the first stage estimate and second stage sample size is neither positive nor negative (for example, in a group sequential or adaptive trial with early rejection and acceptance). In this case the correlation between $\tilde{v}$ and $\bar{x}_1$ and thereby the mean bias can have different sign for different μ. As a consequence, it is usually difficult to quantify the mean bias.

To determine or estimate the bias, we would need to know or estimate μ and the second stage sample size $\tilde{n}_2$ at all interim outcomes. However, confirmatory adaptive designs are intended to deal with cases where $\tilde{n}_2$ does not necessarily follow a pre-specified rule and in this case the mean bias will be unknown even for given μ. Since the rule for $\tilde{n}_2$ can (in general) not be estimated, the bias cannot be estimated in a reliable way. However, based on formula (8.12), an absolute upper bound for the mean bias of $\bar{x}$ that applies to every sample size adaptation rule can be derived (Brannath et al. 2006b). This bound follows from the sample size rule that maximizes the conditional bias. The maximum bias is

$$|E_\mu(\bar{x}) - \mu| \leq 0.4\,\frac{\sigma}{\sqrt{n_1}}\left(\frac{1}{1+r_{\min}} - \frac{1}{1+r_{\max}}\right)\,, \tag{8.13}$$

where $r_{\min}$ and $r_{\max}$ are pre-specified upper and lower bounds for the second stage sample size as multiple of n_1. They must be set to 0 and ∞, respectively, if no such bounds have been pre-specified. Note that according to (8.13) the mean bias is at most 40 % of the standard deviation of the first stage mean.

As mentioned before, the MSE is another important property of an estimator. The following formula for the MSE of the overall ML-estimate has been derived in Brannath et al. (2006b):

$$E_\mu(\bar{x}-\mu)^2 = E_\mu\left(\frac{(\bar{x}_1-\mu)^2}{(1+\tilde{r})^2}\right) + \frac{\sigma^2}{n_1}E_\mu\left(\frac{\tilde{r}}{(1+\tilde{r})^2}\right)\,.$$

From this formula one can see that the MSE also depends on the rule for $\tilde{n}_2$ and hence is in general unknown as well. A similar conclusion holds for the variance for which a formula can also be found in Brannath et al. (2006a)

Similar results can be expected to hold at least approximately for any maximum likelihood estimate $\hat{\theta}$ of a parameter θ if $\bar{x}_k$ are replaced by stage-wise maximum likelihood estimates $\hat{\theta}_k$ and n_1, n_2, and $\tilde{n}_2$ by Fisher informations I_1, I_2, and $\tilde{I}_2$, where I_1 is the information of the first stage sample, I_2 is the information of the second stage sample (excluding the first stage sample) under the pre-planned sample size, and $\tilde{I}_2$ the second stage information under the actual (adapted) sample size.

8.3.2 *Fixed Weighted ML-Estimate*

Let us assume for a moment that the trial is always continued to the second stage, i.e., $\tilde{n}_2 > 0$ for all interim outcomes, and that the only goal of the interim analysis is to recompute the total sample size. In this case we can define estimates that are mean unbiased independently of the sample size adaptation rule. We focus again on estimates for the mean of i.i.d. responses. However, the same conclusions hold for any mean unbiased stage-wise estimates $\hat{\theta}_k$, $k = 1, 2$.

At first we note that the first stage ML-estimate $\bar{x}_1$ and the second stage ML-estimate $\bar{x}_2$ are both mean unbiased estimates. The first stage estimate is unbiased because the first stage sample sizes are fixed, and hence $\bar{x}_1$ has the same statistical properties as in a fixed sample size design. The second stage ML-estimate is unbiased because the sample sizes are chosen based on the first stage data, and these data are not used for $\bar{x}_2$. This means that we are in the same situation as if the sample size would have been determined from an independent trial like a pilot study or from historical data. More formally speaking, if $\tilde{n}_2 > 0$, then the second stage estimate $\bar{x}_2$ is conditionally unbiased given the first stage data and decision on the second stage sample size.

First and second stage maximum likelihood estimate have the disadvantage to utilize only part of the data. For this reason, Liu et al. (2002) and Proschan et al. (2003) suggested to prefix a number $0 < \tau < 1$ and to compute the weighted mean

$$\hat{x}_\tau = \tau\,\bar{x}_1 + (1-\tau)\,\bar{x}_2 \ .$$

Since the weight τ is fixed a-priori, we call $\hat{x}_\tau$ *fixed weighted ML-estimate*. The reason for the unbiasedness of $\hat{x}_\tau$ is that $\bar{x}_1$ and $\bar{x}_2$ are unbiased and $E_\mu(\hat{x}_\tau) = \tau\,\mu + (1-\tau)\,\mu = \mu$. Certainly, the latter conclusion holds only if τ is a fixed number that is specified independently of $\bar{x}_1$ and $\bar{x}_2$.

If a number $n = n_1 + n_2$ has been pre-specified for the total sample size, and the weight $\tau = n_1/(n_1+n_2)$ is according to the pre-planned n_2, then keeping the sample size as pre-planned, i.e., $\tilde{n}_2 = n_2$, implies that $\hat{x}_\tau$ equals the overall ML-estimate $\bar{x}$. However, if $\tilde{n}_2 \neq n_2$, then $\hat{x}_\tau$ and $\bar{x}$ are different. Since $\hat{x}_\tau$ is unbiased its MSE equals its variance. As shown in Brannath et al. (2003), the variance is

$$\text{Var}_\mu(\hat{x}_\tau) = (\sigma^2/n_1)\left(\tau^2 + (1-\tau)^2\,E_\mu(1/\tilde{r})\right) ,$$

which depends on the sample size adaptation rule $\tilde{r}$.

8.3.3 Median Unbiased Point Estimation

Another suggestion for point estimation is to use an estimate which has median equal to μ independently from the adaptation rule (Brannath et al. 2003; Lawrence and Hung 2003; Proschan et al. 2003). Median unbiased point estimates can be obtained from one-sided exact confidence intervals $I_{0.5} = (l^e_{0.5}, \infty)$ that have coverage probability equal to 0.5. Since $P_\theta(\theta \in I_{0.5}) = P_\theta(\theta > l^e_{0.5}) = 0.5$, the estimate $l^e_{0.5}$ is larger than the true θ only with probability 0.5. If $l^e_{0.5}$ has a continuous distribution, then $P_\theta(\theta < l^e_{0.5}) = 0.5$ which implies that its median is equal to the unknown true θ. By a similar reasoning, the upper confidence bound of an exact one-sided 50 %-confidence interval $I_{0.5} = (-\infty, l^e_{0.5})$ provides a median unbiased estimate as well.

Median unbiasedness of $l_{0.5}$ requires that the coverage probability of the interval $(l^e_{0.5}, \infty)$ is exactly equal to 0.5. We have defined one-sided confidence intervals with exact coverage probability for combination tests in Sect. 8.2.1. The corresponding lower confidence bound $l^e_{0.5}$ is the solution of the equation

$$Q_\delta(p_{1,\delta}, p_{2,\delta}) = 0.5$$

in δ where $Q_\delta(p_{1,\delta}, p_{2,\delta})$ is the p-value (8.8) from the stage-wise ordering of combination tests. The solution gives a median unbiased estimate. This estimate can be calculated only once at the end of the trial (at the stage the trial stops).

The lower bound $l_{0.5}$ of a strictly conservative one-sided confidence interval $I_{0.5} = (l_{0.5}, \infty)$ at nominal level 0.5 may also provide an acceptable estimate even though it will be median biased because the coverage probability of $I_{0.5}$ can be larger than 0.5. Such lower bound is, for instance, the lower bound of the repeated confidence interval at level 0.5. A strictly conservative lower bound $l_{0.5}$ is larger than the true value θ only with a probability that is smaller than 0.5 because $\theta < l_{0.5}$ implies $\theta \not\in I_{0.5}$. If larger values of θ correspond to higher efficacy of the experimental treatment, this means that overestimation of efficacy is less likely than underestimation. Such a conservative estimate would prevent systematic overestimation of the treatment effect and hence could be acceptable, in particular, if the probability for underestimation is not too large. In general, the conservatism (for example, of the 50 % repeated confidence bound) will depend on the true parameter value and the adaptation rule, and its investigation will in general require numerical simulations.

8.3.4 Adaptively Weighted ML-Estimate

Several authors (Brannath et al. 2006b; Cheng and Shen 2004; Lawrence and Hung 2003) have suggested using the following weighted mean of first and second stage maximum likelihood estimates $\hat{\theta}_1$ and $\hat{\theta}_2$

$$\tilde{\theta} = \tilde{\tau}\,\hat{\theta}_1 + (1-\tilde{\tau})\,\hat{\theta}_2 \quad \text{with} \quad \tilde{\tau} = \frac{w_1/\mathrm{se}_1}{w_1/\mathrm{se}_1 + w_2/\mathrm{se}_2}\,, \tag{8.14}$$

where se_k is the standard error of $\hat{\theta}_k$, and w_k $(k = 1, 2)$ are fixed positive numbers that satisfy $w_1^2 + w_2^2 = 1$. Note that the weight $\tilde{\tau}$ of $\hat{\theta}_1$ in (8.14) depends on the data, in general, for two reasons. On the one hand, se_2 depends on the first stage data via the data-driven second stage sample size, and on the other hand, se_k usually involves some variance estimates which depend on the data of the respective stage.

We will denote $\tilde{\theta}$ as *adaptively weighted maximum likelihood estimate*. There are several ways to introduce this estimate. Cheng and Shen (2004) derived (8.14) by the method of moments. Recall from Sect. 8.2.2 that $\tilde{\theta}$ equals the estimate in expression (8.10) of the repeated lower confidence bound that results from the inverse normal combination test with weights w_k and p-values $p_{k,\delta} = 1 - \Phi\big(\hat{\theta}_k - \delta)/\mathrm{se}_k\big)$.

If $\alpha_1 = 0$, $\alpha_0 = 1$ and $\alpha = 0.5$, then the repeated lower confidence bound coincides with the estimate $\tilde{\theta}$. Recall also that the repeated confidence interval has exact coverage if $\alpha_1 = 0$ and $\alpha_0 = 1$, and hence $\tilde{\theta}$ is a median unbiased estimate, according to the previous subsection. With a sufficiently large first stage sample size, we would expect that the median bias of the estimate $\tilde{\theta}$ is small also when early stopping is possible ($\alpha_1 > 0$ or $\alpha_0 < 1$). This is because the trial could always be continued with only one (or two) more observations (if the stopping rule applies) which, as we expect, would not effect $\tilde{\theta}$ much (if se_1 is sufficiently small) but would make $\tilde{\theta}$ median unbiased. Mean and median bias as well as mean square of this and the other estimates will be investigated in the next subsection.

8.3.5 Comparison of Point Estimates

The maximum likelihood, the fixed and adaptively weighted maximum likelihood, and the median unbiased estimates are consistent if both, the first and second stage sample sizes, converge to infinity and both stage-wise estimates are consistent. If the trial is never stopped at the interim analysis ($\alpha_1 = 0$ and $\alpha_0 = 1$), then the maximum likelihood and adaptively weighted maximum likelihood estimates are consistent also if only one of the sample sizes (for example, the second stage sample size) increases and the other (for example, of the first stage) remains finite. This stronger consistency property is not satisfied for the fixed weighted maximum likelihood estimate because the weights are independent from the sample size. In trials with $\alpha_1 > 0$ and/or $\alpha_0 < 1$, consistency of any estimate requires that both sample sizes converge to infinity because the probability to stop the trial at the interim analysis is either constant (under H_0 or when the first stage sample size remains fixed) or converges to one (for positive treatment effects) or zero (for negative treatment effects).

We show the results of a simulation study comparing the performance of the different estimates. Like in Lawrence and Hung (2003) and Brannath et al. (2006b), where similar investigations were done, we assume normally distributed responses with mean μ and variance σ, and consider estimates for μ. In contrast to Brannath et al. (2006b) and Lawrence and Hung (2003) we consider trials with and without early stopping. We simulated under four different scenarios. In all scenarios the second stage sample size is reassessed in order to achieve a conditional power of 80 % with the z-test for $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$ at an overall one-sided significance level $\alpha = 0.025$. The alternative is estimated by the positive part of the interim estimate. Given this rule (which could result in an infinite sample size) the second stage sample size is truncated at a given multiple $r_{\max}$ of the first stage sample size. We furthermore assume a specific minimal second stage sample size when the trial continues to the second stage. The minimal second stage sample size is also specified as a multiple $r_{\min}$ of the first stage sample size. We used the weights $w_1 = w_2 = 1/\sqrt{2}$ for the adaptively weighted ML-estimate, and the inverse normal combination test in the computation of the median unbiased estimate.

For all simulated scenarios, $r_{\min} = 0.1$ and $r_{\max} = 5$. In the first scenario, Scenario A, we let $\alpha_1 = 0$, $\alpha_0 = 1$ such that the trial is never stopped at the interim analysis. In two further scenarios, Scenarios B and C, we assume $\alpha_1 = 0.005$ and futility boundaries $\alpha_0 = 0.5$ (Scenario B) and $\alpha_0 = 1$ (Scenario C). In Scenario D we set $\alpha_1 = 0$ and $\alpha_0 = 0.5$. One can show for all four scenarios and for each of the estimates that when dividing the estimate by $\text{se}_1 = \sigma/\sqrt{n_1}$ the distribution becomes independent from n_1 and σ^2 for $\mu = 0$. For $\mu \neq 0$ the distribution depends on n_1, σ^2 and μ only via the non-centrality parameter $\delta_1 = \mu/\text{se}_1$ of $\bar{x}_1$. Hence, we consider the distribution of the estimates divided by standard error se_1 in dependence on the parameter δ_1.

The results are displayed for 100,000 simulation runs. Figure 8.2 illustrates the mean bias of the four previously introduced estimates divided by se_1. Note that in Scenario A the adaptively weighted maximum likelihood estimate is median unbiased and coincides with the median unbiased estimate from the inverse normal combination function with weights $w_1 = w_2 = 1/\sqrt{2}$ as defined in Sect. 8.3.3. The figure illustrates that the fixed weighted ML-estimate is mean unbiased only when there is no early stopping (Scenario A) and exhibits a mean bias in all other scenarios with early stopping. In Scenarios B to D the median unbiased estimate seems to have the smallest maximum mean bias whereas all other estimates are comparable with

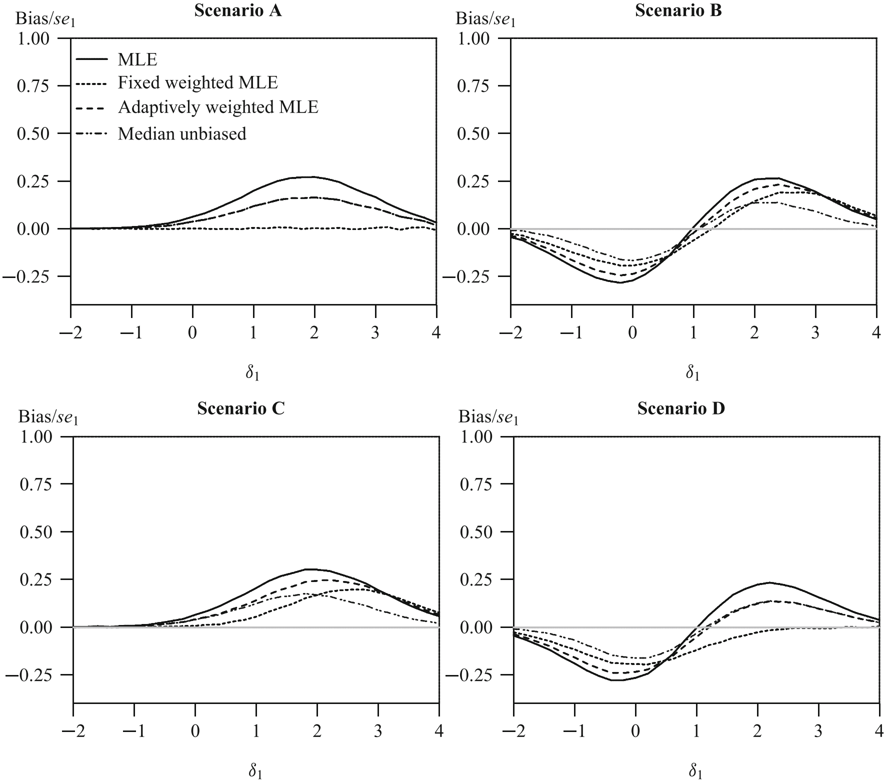

Fig. 8.2 Mean bias of the estimates explained in the text divided by standard error of the first stage estimate $\text{se}_1 = \sigma/\sqrt{n_1}$ in dependence of $\delta_1 = \mu/\text{se}_1$, for the four scenarios explained in the text

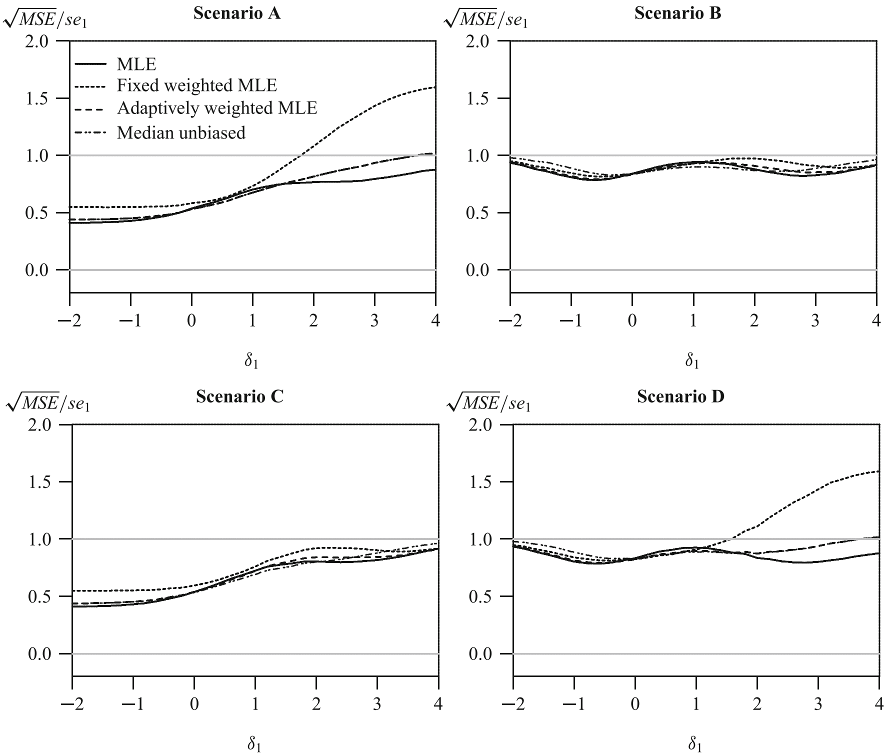

Fig. 8.3 Square root of the MSE of the estimates explained in the text divided by standard error of the first stage estimate $\mathrm{se}_1 = \sigma/\sqrt{n_1}$ in dependence of $\delta_1 = \mu/\mathrm{se}_1$, for the four scenarios explained in the text

regard to mean bias. Note that in Scenario A the adaptively weighted ML-estimate has a smaller mean bias than the usual ML-estimate. However, the mean bias of the two estimates is similar in all other scenarios. This shows that the adaptively weighted ML-estimate has the potential to reduce the bias from the sample size adaptation rule but fails to reduce the bias resulting from the stopping rule.

Figure 8.3 illustrates the square root of the MSE of the estimates divided by se_1. Recall that the MSE is the sum of variance and squared mean bias. The figure clearly shows that the fixed weighted ML-estimate is inferior to all other estimates with regard to the MSE, in particular, in Scenario A ($\alpha_1 = 0$ and $\alpha_0 = 1$) where it is mean unbiased. This implies that the fixed weighted ML-estimate has a too large variance and hence cannot be recommended for practical use even in the case where no early stopping is considered. One can also see that the ML- and adaptively weighted ML-estimates perform best with regard to the MSE. Also the median unbiased estimates appears to have an acceptable MSE, although, the median bias corrections seem to come for the price of somewhat increased MSE as well. Note that the mean bias is

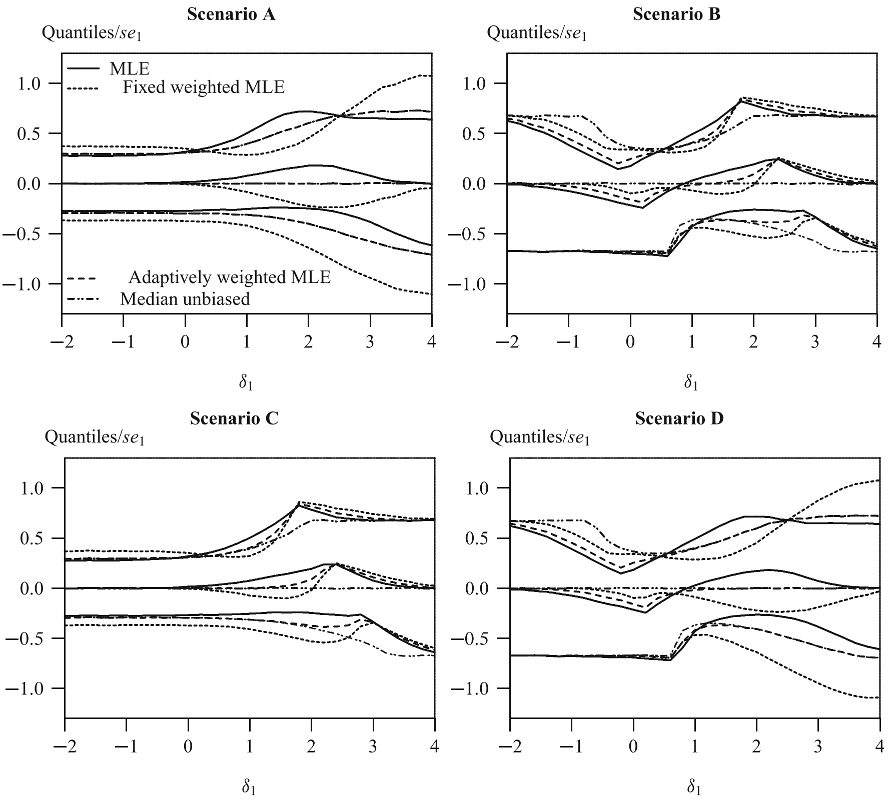

Fig. 8.4 75 %-, 50 %- and 25 %-percentiles of the estimates explained in the text minus μ divided by the standard error of the first stage estimate $\mathrm{se}_1 = \sigma/\sqrt{n_1}$ in dependence of $\delta_1 = \mu/\mathrm{se}_1$, for the four scenarios explained in the text

much smaller than the square root of the MSE. This shows that variance should be a major concern as well.

Figure 8.4 is a plot of the median as well as first and third quartiles of the estimates minus μ (i.e., the residuals) divided by se_1. It shows the tendency of the maximum likelihood estimate to overestimate efficacy. The weighted maximum likelihood estimate appears to be less biased. The median unbiased estimate does not only perform well with regard to the median but also with regard to first and third quartile. Note that the fixed weighted ML-estimate exhibits a median bias for all four scenarios, also in Scenario A where its median bias is mainly negative. The unsmooth behavior of the quartiles is caused by the stopping rule and appears only in Scenarios B to D.

In summary, the median unbiased estimate seems to perform best and be most suitable at least for trials where sample sizes are reassessed based on conditional power. The ML- and adaptively weighted ML-estimates could be considered as reasonable estimate as well, because of their good performance with the regard to the MSE and their computational simplicity. However, these estimate are expected to exhibit some kind of mean and median bias. Since the distribution of an estimator depends on the adaptation rule, we recommend investigating mean and median bias as well as MSE (or quartiles of the residuals) in the planning phase of the trial for the most reasonable adaptation rules.

8.4 Extensions

So far the derivation of overall p-values and especially confidence intervals was based on the known variance assumption. The p-value combination approach, however, only assumes the p-values to be p-clud, so testing means for normally distributed observation can also be performed with p-values obtained from t-tests that do not assume the variance to be known. This includes testing a single mean, or testing the difference and the ratio of two means in the parallel group design. Using these p-values per stage of the trial, overall p-values can directly be obtained. Note that this provides an *exact* solution to the unknown variance problem for normally distributed observations. The overall p-values are exact also in the sense that the hypothesis is rejected at level α if and only if the overall p-value is smaller than α.

Using this principle, one can also easily derive exact RCIs for one mean and the difference or the ratio of two means by considering the p-values $p_{k,\delta}$ for testing the hypothesis

$$H_0^\delta : \theta = \delta \ , \tag{8.15}$$

where $\theta = \mu$, $\theta = \mu_2 - \mu_1$, or $\theta = \mu_2/\mu_1$. The p-values for H_0^δ are calculated with the use of the t-distribution with appropriate df and variance estimate (see Sect. 5.1). The values δ for which the test does not reject H_0^δ can be found numerically and it is easy to see that due to the monotonicity of the p-values in δ there is always a unique solution. The extension of the exact confidence intervals to the case of unknown variance is less straightforward. For the normal case with unknown variance an exact solution has not been provided yet and appears difficult or even impossible. However, there a two ways to obtain confidence intervals that are at least asymptotically exact. The first approach is to simply replace the unknown variance by the actual overall estimate of the variance in both, the two stage-wise p-values and the early rejection boundaries $\alpha_{j,\delta}, j = 0, 1$, that depend on the unknown variance as well (see Sect. 8.2.1.) An alternative, somewhat better (but still only asymptotically valid) approach is to replace the stage-wise p-values by the exact p-values of the t-test, in order to not rely on the normal approximation at this place. This approach is not fully exact because we still do not account for the fact that $\alpha_{j,\delta}$ is data dependent due the variance estimate. Obviously, the second approach controls the confidence level better than the first. Unpublished simulation results indicate that the latter approach provides intervals that are more or less exact also for moderate sample sizes.

In the same way on can derive overall p-values and confidence intervals for binary observations. If π denotes the unknown rate in a one-sample situation, π_1, π_2 the unknown rates in a two-sample situation the problem is to define tests for (8.15) with $\theta = \pi$, $\theta = \pi_2 - \pi_1$, or $\theta = \pi_2/\pi_1$. Appropriate tests were described in Sect. 5.2. For example, through the approximate approach of Farrington and Manning (1990) there is always a unique solution and the exact correspondence between the test decision and the RCIs is generally fulfilled (see Wassmer 2003).

Recall that in this section we considered overall p-values and confidence intervals for two-stage adaptive designs. If we use the combination testing principle, however, it is easy to see that most of the techniques derived for the two-stage case can also be applied for more than two stages. Specifically, if the inverse normal combination test is used, the calculation of exact and repeated overall p-values is straightforward since the same formulas as for the group sequential case with the group sequential test statistic replaced by the inverse normal test statistic can be used. This is also true for more general testing situation, for example, the unknown variance or the binary case. If Fisher's combination test is used, the calculation of exact and repeated overall p-values might be a bit more complicated (see Sect. 8.1.2) but possible in principle.

RCIs can be derived for any combination tests as long as suitable tests for (8.15) can be derived. The derivation is finding those δ that do not lead to the rejection of H_0^δ at a given stage k possibly taking into account a (binding) stopping for futility boundary (see Sect. 8.2.4). The calculation of exact confidence intervals for multi-stage combination tests is not solved in general yet. A solution has been given for recursive combination tests where two-stage combinations test are applied recursively to the overall p-values of combination tests (Brannath et al. 2002). We note that there is a solution for deriving exact confidence intervals when using the CRP principle (Brannath et al. 2009a). There might be cases, however, where the confidence limit is the solution of a function in δ that is not monotone in δ. It is in fact easier to derive RCIs when using the CRP principle (Mehta et al. 2007) because here such monotonicity problems do not occur. We further refer to the "backward mapping" approach of Gao et al. (2013a,b) which also provides confidence intervals for multi-stage designs that are based on the stage-wise ordering. Finally, Liu and Anderson (2008) proposed a new ordering yielding sequential p-values and confidence intervals that is applicable in the adaptive setting.

Chapter 9
Adaptive Designs with Survival Data

The previously reviewed methods for designs with adaptive sample size modifications have been extended to survival data by Schäfer and Müller (2001), Shen and Cai (2003), Wassmer (2006), Desseaux and Porcher (2007), Jahn-Eimermacher and Ingel (2009), and Irle and Schäfer (2014). Schäfer and Müller (2001) and Irle and Schäfer (2014) extend the conditional error rate principle. Wassmer (2006) and Jahn-Eimermacher and Ingel (2009) consider the inverse normal and Desseaux and Porcher (2007) Fisher's combination test approach. Shen and Cai (2003) extend the variance spending method of Shen and Fisher (1999). Most of the mentioned approaches were developed in the multi-stage context. In this chapter we focus on the combination testing approach and briefly describe some problems that might arise in survival trials.

9.1 Combination of *p*-Values from Log-Rank Tests

Consider a trial where two treatment groups are compared with regard to a survival endpoint, for example, overall survival. As described in Sect. 5.3 a common assumption in survival studies is that the hazard ratio $\omega = \lambda_2(x)/\lambda_1(x)$ or the log hazard ratio $\theta = \log(\omega)$ is constant over the time x. Typically, one is then interested in the null hypothesis $H_0 : \theta = 0$ versus the alternative $H_1 : \theta > 0$. In a randomized, parallel group design with fixed sample sizes it is common to use the one-sided log-rank test for testing H_0. As we have outlined in Sect. 5.3 this log-rank test can be extended to group sequential designs.

If at an interim analysis the number of individuals or the target number of events is changed based on the interim data, then the usual log-rank test may not keep the nominal Type I error rate. It has been illustrated by Wassmer (2006), Desseaux and Porcher (2007), and Jahn-Eimermacher and Ingel (2009) that in this case the

G. Wassmer, W. Brannath, *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*, Springer Series in Pharmaceutical Statistics,
DOI 10.1007/978-3-319-32562-0_9

combination test approach can be applied. The main difficulty with this approach is the definition of the stage-wise data and the p-values. The problem is that some interim subjects may not have their event until the end of the first stage and then will also be followed up after the interim analysis. These subjects provide information for both sequential stages. At the end of the first stage we know that these subjects have survived the time from their entry to the interim analysis. At the end of the second stage we know the times they survive from the interim analysis to the end of the study, and so on. Hence, the event data observed before and after an interim analysis are not from independent cohorts of patients.

9.1.1 Use of Independent Increments

One approach to deal with this difficulty is to use similar arguments as for group sequential survival trials. For illustration, in a two-stage design this leads us to consider the following two log-rank statistics. The first statistic is computed from the information observed until the end of the first stage. The first stage information can formally be viewed as the survival times of the patients recruited before the interim analysis right censored at the time point of the interim analysis. This means that we view the first stage as an individual survival study that ends at the interim analysis. We use the corresponding log-rank test for the first stage p-value. The second statistic is the log-rank statistic for all the survival information observed in the two-stage trial and the second stage p-value is build from a weighted difference between the first and second log-rank statistics. Like for group sequential trials, this approach (under mild regularity assumptions) leads to p-values that are approximately independent (or p-clud) for sufficiently large number of events.

To describe the approach in the more general multi-stage setting, assume that we observe d_1 events until the end of the first interim analysis and, when the trial continues to the second stage, the cumulative number of events d_2 at the second interim analysis of the trial, and so on. We assume that there are no ties and denote the ordered event times observed until the end of stage $k = 1, \ldots, K$ by $x_{1k} < x_{2k} < \cdots < x_{d_k k}$. We further define $I_{2ik} = 1$ if the event observed until the end of stage k at time x_{ik} occurs in the second treatment group and $I_{2ik} = 0$ otherwise. We further let N_{jik} be the number of subjects recruited until the end of stage k in treatment group j that are at risk at time x_{ik}. As in Sect. 5.3, the one-sided log-rank statistics are defined as

$$\mathrm{LR}_k = \frac{\sum_{i=1}^{d_k} \left(I_{2ik} - \frac{N_{2ik}}{N_{1ik}+N_{2ik}} \right)}{\sqrt{\sum_{i=1}^{d_k} \frac{N_{1ik}N_{2ik}}{(N_{1ik}+N_{2ik})^2}}}, \quad k = 1, \ldots, K.$$

It is known that $Z_1 = \mathrm{LR}_1$ is approximately standard normally distributed under H_0. Furthermore, with constant treatment allocation ratio the statistic

$$Z_k = \frac{\sqrt{d_k}\mathrm{LR}_k - \sqrt{d_{k-1}}\mathrm{LR}_{k-1}}{\sqrt{d_k - d_{k-1}}}\,, \tag{9.1}$$

where $\mathrm{LR}_0 = d_0 = 0$ is approximately standard normal, too, and Z_{k-1} and Z_k are approximately independent (see Sect. 5.3). We can therefore use the stage-wise p-values $p_k = 1 - \Phi(Z_k)$ and the inverse normal test statistic reads as

$$Z_k^* = \left(\sum_{\tilde{k}=1}^{k} w_{\tilde{k}}^2\right)^{-1/2} \sum_{\tilde{k}=1}^{k} w_{\tilde{k}} \frac{\sqrt{d_{\tilde{k}}}\,\mathrm{LR}_{\tilde{k}} - \sqrt{d_{\tilde{k}-1}}\,\mathrm{LR}_{\tilde{k}-1}}{\sqrt{d_{\tilde{k}} - d_{\tilde{k}-1}}}\,. \tag{9.2}$$

Let the weights be fixed through

$$w_1 = \sqrt{\zeta_1}\,,\;\; w_k = \sqrt{\zeta_k - \zeta_{k-1}}\,,\; k = 2, \ldots, K, \tag{9.3}$$

where ζ_k denote the planned (or expected) number of accumulated events at stage k. Then, if $d_k = \zeta_k, k = 1, \ldots, K$,

$$Z_k^* = \frac{1}{\sqrt{\zeta_k}} \sum_{\tilde{k}=1}^{k} \sqrt{\zeta_{\tilde{k}} - \zeta_{\tilde{k}-1}} \frac{\sqrt{\zeta_{\tilde{k}}}\,\mathrm{LR}_{\tilde{k}} - \sqrt{\zeta_{\tilde{k}-1}}\,\mathrm{LR}_{\tilde{k}-1}}{\sqrt{\zeta_{\tilde{k}} - \zeta_{\tilde{k}-1}}} = \mathrm{LR}_k\,.$$

That is, if the observed number of events equals the planned number, the inverse normal test statistic is the same as the usual log-rank test statistic. Therefore, it is reasonable to use the test statistic (9.2) and to specify the weights according to (9.3) in the planning phase.

9.1.2 Applying Left Truncation at the Second Stage

Jahn-Eimermacher and Ingel (2009) consider an alternative approach to build the second stage p-value p_2. In this approach the second stage p-value is directly from the events observed at the second stage. This is achieved by left truncation and right censoring. The approach goes back to Keiding et al. (1987) which consider the possibility of reusing subjects from an exploratory survival study in a subsequent confirmatory study (see also Parner and Keiding 2001; Keiding 2006). The possibility of using this approach for combination tests has already been mentioned in Bauer and Köhne (1994).

To describe the approach in more detail let R_i be the calendar time of entry for individual i. Let further X_i be its time from entry to the event and C_i its time from entry to censoring. We assume that X_i, R_i, and C_i are independent and we denote by r_i, x_i, and c_i the observed outcomes of R_i, X_i, and C_i, respectively.

The multi-stage data are distinguished by different survival data and risk intervals as follows. As in the previous approach, the first stage data of individual i consists of the time and status variable $y_{i,1} = \min\{x_i, c_i, t_1 - r_i\}$ and $\delta_{i,1} = \mathbf{1}_{\{x_i \leq \min\{c_i, t_1 - r_i\}\}}$, whereby we consider only individuals recruited before the interim analysis ($r_i < t_1$). The risk interval for individual i is $(0; y_{i,1})$, which means that individual i belongs to the risk set at the event time $x_j = y_{j,1}$ if $\min\{c_i, t_1 - r_i\} \geq x_j$.

The second stage data consists of the survival data $y_{i,2} = \min\{x_i, c_i, t_2 - r_i\}$ and $\delta_{i,2} = \mathbf{1}_{\{x_i \leq \min\{c_i, t_1 - r_i\}\}}$ from individuals with $r_i + \min\{x_i, c_i\} > t_1$, i.e., restricted to events that are observed after the interim analysis. The risk interval is $(t_1 - r_i; y_{i,2})$ which means individual i belongs to the risk set at $x_j = y_{j,2} \geq t_1 - r_j$ if $t_1 - r_i < x_j \leq y_{i,2}$. The subsequent stages are defined analogously. Keiding et al. (1987) demonstrate by a factorization of the empirical likelihood that the above defined stage-wisely defined data are independent (see also Jahn-Eimermacher and Ingel 2009). Hence, the resulting log-rank tests and stage-wise p-values are independent as well.

For illustrating this approach consider a two-stage design. Subjects of this trial can be divided into three groups. The first group consists of the subjects which are recruited before the interim analysis and have their event or are censored during the first stage. This group contributes data only to the first stage data. The second group are the subjects which are recruited after the interim analysis. They provide survival data only at the second stage. The third group consists of the subjects that are recruited before the interim analysis and have their event after the interim analysis. These subjects provide information to both sequential stages.

The first stage information of subjects in the third group is the survival time right censored at the interim analysis. Hence, the first stage log-rank statistic LR_1 and p-value p_1 can be computed as in the previous subsection. Events from patients in the third group may either occur during the second stage or after the end of the study. In the latter case the survival time is censored at the study end. The second stage information is the time from the interim analysis to the event or end of the second stage (whatever comes first) plus the censoring status at the end of the study. For the second stage they are considered at risk only for time points between the interim and final analysis. This risk interval can easily be implemented in statistical software that supports the counting process style (like SAS and R) by defining appropriate start and stop variables. For the first stage subjects from the second stage are considered at risk only for time points before the interim analysis, whereas for the second stage this restriction does not apply.

9.2 Restriction of the Information Used for the Adaptations

Since event or censoring times of interim subjects observed after the interim analysis are part of the second stage data, we must be careful in using information other than the interim survival status of these patients. It has been described by Bauer and Posch (2004) that there is a major limitation in adaptive survival trials with regard to the information that may be used for the adaptations at the interim analysis. Although the arguments have been given for the conditional error rate method these restrictions also apply to combination tests.

Bauer and Posch (2004) noted that for survival designs only the censored survival times or the log-rank statistic from the confirmatory phase may be used for subsequent planning, no other information from patients under risk can be used. The reason is that such interim information may be correlated with the survival time observed at the second stage data, and with the presence of such a correlation the occurrence of an event can be predicted for the patients under risk. Adaptations based on such information could lead to inflations of the Type I error rate and control of Type I error rate cannot be guaranteed any more. Bauer and Posch (2004) show that in extreme cases the inflated Type I error rate could be as large as twice the nominal level. Since their example is based on a worst case scenario where a surrogate endpoint is used which is perfectly correlated to survival, and the sample size adaptations directly intend to maximally inflate the Type I error rate, a less pronounced inflation is expected to appear in practice. However, to preserve the trial from potential inflation we should avoid using information, for example, baseline characteristics, secondary, surrogate, and safety endpoints from patients that are recruited before the interim analysis but are event free until then.

The design adaptations can always be based on the censored survival times from all first stage subjects and, in particular, on the first stage log-rank statistic. If some limited information other than the survival data shall be used for the adaptation decisions, then the second stage p-value must be from a test that is conditional on this information. Such a p-value can be obtained by stratification if the utilized information is categorial (see Brannath et al. 2009b) or from a regression model. In the latter case Type I error rate control relies on the correctness of the regression model.

Suggestions have also been made to construct valid adaptive test statistics including as much information as possible and allowing interim decision making on all collected data for multi-arm and population enrichment designs (Friede et al. 2011; Jenkins et al. 2011; Mehta et al. 2014; Irle and Schäfer 2014; Carreras et al. 2015). We describe these very important applications of adaptive designs in more detail in Part III of this book. Recently, Magirr et al. (2014a) show that the proposals have the common disadvantage that the final test statistic may ignore a substantial subset of the observed survival times. They show if the goal is to use all the data, a worst case adjustment of the critical boundaries guarantees Type I error control for the price of reduced power.

Other methods require assumptions regarding the joint distribution of survival times and short-term secondary endpoints (Di Scala and Glimm 2011; Stallard 2010). Note that related problems arise in overrunning, for example, patients having been recruited before or during the interim analysis such that their data could not be used for interim decision making (Faldum and Hommel 2007). Such kind of delayed responses were also considered by Hampson and Jennison (2013) in the context of classical group sequential designs.

9.3 Sample Size Reassessment Rules

Sample size calculation for the log-rank test is usually performed in two steps. In the first step the required number of events is determined for a specific power and hazard ratio. In the second step the number of patients for achieving the event number is estimated from assumptions on the survival distributions, see Sect. 5.3. Schäfer and Müller (2001), Wassmer (2006), and Desseaux and Porcher (2007) apply similar methods for sample size reassessments at the interim analysis. As discussed in Chap. 7 sample size reassessments are most naturally based on conditional power (rather than overall power) under specific assumptions on the treatment effect, here the hazard ratio, which can but need not be influenced by the interim data and external information. According to the restrictions mentioned in the previous section such estimate can depend on the survival data of all interim patients, however, not on other information from the yet event free and uncensored subjects.

The conditional power at a given interim stage is the probability of a significant outcome in one of the subsequent stages given the results observed so far under a specified parameter value of interest. If only one stage is left or one is interested in the conditional power for the subsequent stage, the necessary number of events to reach a specified power can easily be found by considering the univariate standard normal cdf with an appropriate shift of the decision region. For example, when using the inverse normal statistic (9.2), for given allocation ratio r, from (7.4) in Sect. 7.3 the number of additional events for the next (kth) stage of the trial to reach conditional power cp for this stage given the results up to stage $k-1$ is found to be

$$d_k - d_{k-1} = \frac{(1+r)^2}{r} \, \frac{\left(\Phi^{-1}(\mathrm{cp}) + \left((\sum_{\tilde{\mathrm{k}}=1}^{\mathrm{k}} \mathrm{w}_{\tilde{\mathrm{k}}}^2)^{1/2}\, \mathrm{u}_{\mathrm{k}} - \sum_{\tilde{\mathrm{k}}=1}^{\mathrm{k}-1} \mathrm{w}_{\tilde{\mathrm{k}}}\, \mathrm{z}_{\tilde{\mathrm{k}}}\right)/\mathrm{w}_{\mathrm{k}}\right)^2}{\left(\ln(\omega_a)\right)^2}\,, \tag{9.4}$$

where ω_a is the assumed hazard ratio, in accordance with Schoenfeld (1981) sample size estimation formula (see Wassmer 2006).

In the general case, the multivariate normal integral can be used to calculate the overall power at given number of events. For specified overall conditional power the necessary number of events during the stages is found by a linear search. For this calculation, as a shortcut it is also possible to use the last stage critical value, u_K, in place of u_k in (9.4). In most cases this supplies an adequate approximation (more precisely, a lower bound) to the required number of events.

Given the observed data, the conditional power can be calculated in two different ways: Either ω_a is set equal to an overall estimate $\hat{\omega}$ of the hazard ratio at given stage $k-1$, or ω_a is set equal to a pre-specified value, for example, the minimum relevant hazard ratio that is worthwhile to detect. It is also possible to compute the conditional power for a range of parameter values ω in order to reassess the necessary amount of information. How to estimate ω is considered next.

9.4 Estimation of the Hazard Ratio

As an estimate $\hat{\omega}$ for ω at stage k, $k = 1, \dots, K$, given the log-rank test statistic, the number of events d_k, and the allocation ratio r_k, one might use the unadjusted estimate

$$\hat{\omega} = \exp\left(\frac{\mathrm{LR}_k}{\sqrt{d_k}} \, \frac{1 + r_k}{\sqrt{r_k}} \right) , \tag{9.5}$$

that is based on

$$E(\mathrm{LR}_k) = \sqrt{d_k} \, \frac{\sqrt{r_k}}{1 + r_k} \, \log(\omega) \tag{9.6}$$

(see Sect. 5.3). This estimate is reasonable if the hazard ratio is not too far from 1 and if the number of patients at risk is roughly equal in the treatment groups.

Based on (9.6), an approximate solution to the computation of repeated confidence intervals for the hazard ratio is easily derived. Consider an adaptive group sequential design with no provision for early stopping in favor of H_0. Given the critical values $u_1, \dots, u_K$ at two-sided level α (or one-sided level $\alpha/2$), the approximate two-sided $(1 - \alpha)100\,\%$ RCI for ω is computed as

$$\exp\left(\frac{\sum_{\tilde{k}=1}^{k} w_{\tilde{k}} z_{\tilde{k}}}{h} \pm \frac{u_k \left(\sum_{\tilde{k}=1}^{k} w_{\tilde{k}}^2 \right)^{1/2}}{h} \right) \tag{9.7}$$

with $z_{\tilde{k}}$ from (9.1), and h is given by

$$h = \sum_{\tilde{k}=1}^{k} \frac{w_{\tilde{k}}}{\sqrt{d_{\tilde{k}} - d_{\tilde{k}-1}}} \left(d_{\tilde{k}} \, \frac{\sqrt{r_{\tilde{k}}}}{1 + r_{\tilde{k}}} - d_{\tilde{k}-1} \, \frac{\sqrt{r_{\tilde{k}-1}}}{1 + r_{\tilde{k}-1}} \right) ,$$

where $d_0 = r_0 = 0$. Formula (9.7) is derived by finding the values ω for which

$$z_k^* - E(Z_k^*) \le u_k \, ,$$

where, given $d_{\tilde{k}}$,

$$E(Z_k^*) = \left(\sum_{\tilde{k}=1}^{k} w_{\tilde{k}}^2 \right)^{-1/2} \sum_{\tilde{k}=1}^{k} \frac{w_{\tilde{k}}}{\sqrt{d_{\tilde{k}} - d_{\tilde{k}-1}}} \left(d_{\tilde{k}} \, \frac{\sqrt{r_{\tilde{k}}}}{1 + r_{\tilde{k}}} - d_{\tilde{k}-1} \, \frac{\sqrt{r_{\tilde{k}-1}}}{1 + r_{\tilde{k}-1}} \right) \ln(\omega) \, .$$

At stage k, these intervals are merely based on the observed values of the log-rank test statistics $\mathrm{LR}_{\tilde{k}}$, the observed number of events $d_{\tilde{k}}$, and the allocation ratios $r_{\tilde{k}}$

specifying the proportion of patients randomized to treatment group 1 and 2 at analysis stage $\tilde{k}$, $\tilde{k} = 1, \ldots, k$. Similar intervals were proposed by Jennison and Turnbull (1989) for group sequential tests. Wassmer (2006) showed by simulation that the confidence level of the RCIs as given by (9.7) is satisfactory fulfilled for exponentially distributed survival times. Note that usually the number of patients at risk is not a constant during the course of the trial. The proposed intervals account for this change in the allocation ratios.

The midpoint of the confidence interval (9.7) might serve as a reasonable point estimate for the hazard ratio, too (see Sect. 8.3). The performance of the different estimates, including (9.5) and the ones described in Sect. 8.3, were evaluated in Ligges (2012). She also extended the simulation result for many practically relevant situations includes dropouts, ties, and different survival time distributions. As a final note, there might be difficulties to supply the final analysis confidence intervals and median unbiased estimates based on the stage-wise ordering, at least for the general multi-stage case (see Sect. 8.4).

Part III
Adaptive Designs with Multiple Hypotheses

Chapter 10
Multiple Testing in Adaptive Designs

In the first two parts of this book we have focused on group sequential and adaptive designs for a single hypothesis. In many clinical trials multiple study objectives are investigated, leading to multiple hypotheses of interest. Sometimes the wish is then to use the information available up to an interim analysis to adapt the hypotheses under investigation. We consider two examples to motivate the need for adaptive designs with multiple hypotheses. Other applications and case studies will be described and discussed in detail later in Chap. 11.

As a first example, we consider a prospectively planned adaptive clinical study described by Posch et al. (2005), which consisted of two stages. The first stage aimed at comparing efficacy and safety of three treatments with placebo. An interim analysis after the end of the first stage was planned to select the treatment with the best benefit/risk ratio for the second stage of the study. The final analysis then aimed at the comparison of efficacy and safety for the single treatment selected at interim with placebo while using the combined information from both stages. In this example, the hypotheses adaptation consists of selecting a treatment at interim such that for the final analysis we are left with testing a single hypothesis out of initially three hypotheses (one corresponding to each treat-control comparisons).

Brannath et al. (2009b) described a clinical trial for the comparison of a new treatment with an active comparator. The new treatment was hypothesized to be a targeted treatment, i.e., a treatment which possibly had a higher benefit in a specific patient subgroup. Nevertheless, the new treatment could also have a substantial effect in the full population. In this example, an adaptive design was envisaged to test a targeted therapy for a selected population. An interim analysis was therefore planned to (1) validate the added benefit in the pre-specified patient subgroup, (2) decide whether the study is warranted to continue to the next stage, or should be

G. Wassmer, W. Brannath, *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*, Springer Series in Pharmaceutical Statistics,
DOI 10.1007/978-3-319-32562-0_10

stopped for futility, and, in case that the study continues, and (3) decide whether to include only patients from the specific subgroup or patients from the full population. If it is decided at interim to continue with the full population, it is then determined whether to test the hypothesis of a treatment effect in the specific subpopulation in addition to that of the full population.

In this chapter we introduce the basic statistical methods to design and analyze clinical trials with hypotheses adaption at an interim analysis. In Sect. 10.1 we describe the various sources of multiplicity which can arise in such trial designs. To avoid an inflation of the overall Type I error rate beyond a pre-specified significance level α, such considerations have to be taken into account. So far, we have concentrated on those sources of multiplicity that arise in group sequential and adaptive designs. In this part of the book we additionally are faced with more "traditional" multiplicity problems such as comparing multiple treatment with a control or identifying a suitable (sub-)population. The rest of this chapter is organized as follows. In Sect. 10.2 we review the closure principle, which is a powerful and flexible method to construct multiple test procedures controlling the overall Type I error rate for a given family of null hypotheses in a fixed sample design. In Sect. 10.3 we then describe the core methodology to test adaptively multiple hypotheses by using adaptive combination tests, which essentially apply the methods described in Part II to the closure principle. The resulting methods are very general and allow for flexible adaptations of hypotheses at interim. Much of the material in this chapter is based on the review papers from Bretz et al. (2006, 2009a). In Chap. 11 we will describe how the general principle can be applied to the most important application of confirmatory adaptive designs which are adaptive treatment selection and adaptive population enrichment designs. We also provide some examples on other types of adaptation.

10.1 Sources of Multiplicity

The principal differentiation of adaptive designs compared to traditional designs without adaptations is the ability to perform interim analyses in order to take decisions affecting the further conduct of the trial. This leads to repeatedly testing of one or multiple hypotheses and the possibility to change design features based on interim data. Since the same interim data is subsequently used for hypothesis testing and estimation such approaches may cause bias in estimation and inflation of the Type I error rate if not adequately controlled.

Table 10.1 summarizes the key sources of multiplicity and how to control the overall Type I error rate in each case. As described in Part I of this book, repeatedly looking at the data with the option to stop the trial early for success by rejecting the null hypothesis may inflate the overall Type I error rate beyond the pre-specified significance level α. Consequently, proper statistical methods have to be employed, for example, group sequential methods or α-spending function approaches. Similarly, the possibility of adapting design and analysis features

Table 10.1 Sources of multiplicity and control of Type I error rate inflation (adapted from Maurer et al. 2010)

Sources of multiplicity	Related techniques to control the overall Type I error rate
Repeated hypothesis testing with early rejection of null hypotheses at an interim analysis	Group sequential methods; α spending function approaches
Adaptation of design and analysis features with combination of information across trial stages	Combination of p-values; conditional error function approaches; methods conditional on known adaptation rules
Testing multiple hypotheses testing (for example, because of comparing multiple treatment with a control or identifying a suitable population)	Traditional multiple testing methods, such as the closure principle

while combining the accrued information across the various trial stages in the final analysis may also inflate the overall Type I error rate. In Part II we have described related approaches to control the overall Type I error rate, such as the methods that are based on a p-values combination function or the conditional error function. Finally, when testing more than one null hypothesis (for example, because of comparing multiple treatment with a control or identifying a suitable population), traditional multiple testing methods, such as the closure principle, have to be employed.

In the following sections we demonstrate how to design and analyze adaptive designs involving multiple null hypotheses with possible adaptations at interim, such that the overall Type I error rate is strictly controlled by combining different elements from Table 10.1.

10.2 The Closure Principle

In this section we briefly review the closure principle due to Marcus et al. (1976), which is a common technique to construct powerful multiple test procedures. The closure principle provides a large degree of flexibility to map the difference in importance as well as the relationship between the various study objectives onto an adequate multiple test procedure. This flexibility will be the key for the considerations in the subsequent sections. Note that in this section we do not consider adaptive testing.

Assume that G directional (i.e., one-sided) null hypotheses H_0^g, $g = 1, \ldots, G$, are to be tested (for example, the comparison of G treatments with a control). Performing an α-level test for each of the G hypotheses H_0^g may lead to a substantial inflation of the overall Type I error rate. That is, the probability to reject at least one true null hypothesis may be larger than the pre-specified significance level α. However, it is mandatory for confirmatory clinical trials that the probability to declare at least one ineffective treatment as effective is bounded by α. Hence the need of multiple test procedures which strongly control the familywise error rate (FWER) at level α are needed, where the strong control of the FWER is defined as the maximum probability of rejecting at least one true null hypothesis being lower than α, irrespective of the configuration of true and false null hypotheses (Hochberg and Tamhane 1987).

The closure principle is a general methodology to construct multiple test procedures controlling the FWER in the strong sense. We define the global hypothesis as

$$H_0 = \bigcap_{g=1}^{G} H_0^g.$$

Intersection hypotheses are, for example,

$$H_0^{\{2,3\}} = H_0^2 \cap H_0^3.$$

The closure principle considers all intersection hypotheses which are constructed from the initial set of elementary null hypotheses $H_0^1, \ldots, H_0^G$. To control the FWER, an intersection null hypothesis $H_0^{\mathcal{I}}$, $\mathcal{I} \subseteq \{1, \ldots, G\}$, can only be rejected if all intersection hypotheses implying $H_0^{\mathcal{I}}$ are rejected, too, at level α. The resulting closed test procedure is operationally defined as follows:

1. Define a set of elementary hypotheses $H_0^1, \ldots, H_0^G$ of interest.
2. Derive the corresponding closed system of hypotheses consisting of all possible intersection hypotheses $H_0^{\mathcal{I}} = \bigcap_{g \in \mathcal{I}} H_0^g, \mathcal{I} \subseteq \{1, \ldots, G\}$.
3. For each of the hypotheses in the closed system of hypotheses find a suitable local level-α test.
4. Reject $H_0^{\mathcal{I}}$ at FWER α, if all hypotheses $H_0^{\mathcal{J}}$ with $H_0^{\mathcal{J}} \subseteq H_0^{\mathcal{I}}$ are rejected, each at (local) level α.

It can be shown that this procedure strongly controls the FWER at level α (Marcus et al. 1976). It can be seen (intuitively) as follows: whatever collection $\mathcal{I}$ of null hypotheses is true, in order to reject any of them, we need to reject their intersection $H_0^{\mathcal{I}}$ with the given (local) level α test. Hence, the probability to reject any true null hypothesis is necessarily bounded by α. Note that we can specify any α-level test for the intersection hypotheses. In particular, different tests can be used for different hypotheses. This property will be exploited when constructing adaptive tests based on the closure principle.

An elementary hypothesis H_0^g is rejected if H_0^g itself and all hypotheses containing (i.e., implying) H_0^g are rejected, too, each at level α. For $G = 3$, the closed system of hypotheses is illustrated in Fig. 10.1 indicating the way of how rejection of an elementary hypothesis is reached: Only if the global hypothesis can be rejected, hypotheses consisting two elementary hypothesis are tested, only if an intersection hypothesis is rejected, subsequent elementary hypotheses are tested.

$$\begin{tikzcd}
 & H_0^1 \cap H_0^2 \cap H_0^3 \arrow[dl] \arrow[d] \arrow[dr] & \\
H_0^1 \cap H_0^2 \arrow[d] \arrow[dr] & H_0^1 \cap H_0^3 \arrow[dl] \arrow[dr] & H_0^2 \cap H_0^3 \arrow[dl] \arrow[d] \\
H_0^1 & H_0^2 & H_0^3
\end{tikzcd}$$

Fig. 10.1 Closed test procedure for three null hypotheses H_0^1, H_0^2, and H_0^3

10.3 Closed Testing in Adaptive Designs

In this section we discuss how to test adaptively multiple hypotheses by combining the techniques from the previous sections. The following results are rather general and allow flexible adaptations of hypotheses at interim analyses, as illustrated later in this section with two generic examples.

Assume that we are interested in testing G hypotheses $H_0^1, \ldots, H_0^G$ using a two-stage design with a single interim analysis. The generalization of the subsequent methods and results to more than two stages is mostly straightforward. The general rule is to apply the closure principle by constructing all necessary intersection hypotheses and testing each of them with a suitable combination test (Bauer and Kieser 1999; Kieser et al. 1999; Lehmacher et al. 2000; Hommel 2001). Following the closure principle, a null hypothesis H_0^g is rejected if all hypotheses implying H_0^g are also rejected. In the sequence we call the resulting closed test procedure as "adaptive closed test."

Consider Fig. 10.2 for an example of adaptively testing $G = 2$ hypotheses. As before, let H_0^1, H_0^2, and $H_0^{\{1,2\}}$ denote the hypotheses to be tested according to the closure principle. Let further $p_k^{\mathcal{J}}$ denote the one-sided p-value for hypothesis $H_0^{\mathcal{J}}, \mathcal{J} \in \{1, 2, \{1, 2\}\}$ at stage $k = 1, 2$. Finally, let $C(p_1^{\mathcal{J}}, p_2^{\mathcal{J}})$ denote the combination function C applied to the p-values $p_k^{\mathcal{J}}$ from stage $k = 1, 2$, as introduced in Sect. 6.2. Note that different combination functions as well as different stopping boundaries could be used for the different intersection hypotheses (for simplicity we omit this generalization here). According to the closure principle, H_0^1 (say) is rejected while strongly controlling the FWER at level α if H_0^1 and $H_0^{\{1,2\}}$ are both rejected each at local level α. In terms of the combination function C, H_0^1 is rejected if $C(p_1^1, p_2^1) \leq c$ and $C(p_1^{\{1,2\}}, p_2^{\{1,2\}}) \leq c$ for the critical value c for which (6.1) is satisfied. In other words, H_0^1 is rejected if the combination function test is significant when applied to the first stage and second stage p-values for H_0^1 and $H_0^{\{1,2\}}$, respectively.

Recall from Sect. 6.3 that the combination function approach is closely related to the conditional error function approach, i.e., the CRP principle. Accordingly, the

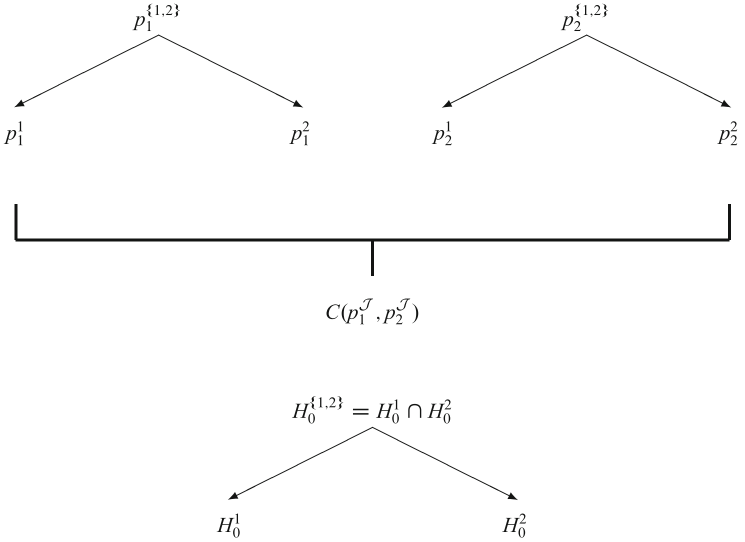
$p_1^{\{1,2\}}$
$p_2^{\{1,2\}}$
p_1^1
p_1^2
p_2^1
p_2^2
$C(p_1^{\mathcal{J}}, p_2^{\mathcal{J}})$
$H_0^{\{1,2\}} = H_0^1 \cap H_0^2$
H_0^1
H_0^2

Fig. 10.2 Closure principle for testing adaptively $G = 2$ null hypotheses H_0^1 and H_0^2. *Top*: stage-wise p-values for the hypotheses H_0^1, H_0^2, and $H_0^{\{1,2\}}$. *Bottom*: combination function $C(p_1^{\mathcal{J}}, p_2^{\mathcal{J}})$ applied to the stage-wise p-values and the underlying closed test procedure (adapted from Bretz et al. 2006)

rejection rules derived from Fig. 10.2 can also be expressed in terms of conditional error functions $A(p_1)$. Following this alternative formulation, H_0^1 is rejected if $p_2^1 \leq A(p_1^1)$ and $p_2^{\{1,2\}} \leq A(p_1^{\{1,2\}})$. An analogous rejection rule applies also to H_0^2.

Note that the rejection rules simplify if early rejection or non-rejection of any hypothesis is achieved at the interim analysis. Following the notation from Sect. 6.2, let α_0 and α_1 denote early stopping boundaries such that a null hypothesis is rejected (retained) early at the interim analysis if its associated p-value is less (greater) than or equal to α_1 (α_0). Assume now that for the intersection hypothesis $H_0^{\{1,2\}}$ the first stage p-value $p_1^{\{1,2\}} \leq \alpha_1$. By definition, the associated conditional error function becomes $A(p_1^{\{1,2\}}) = 1$ and the condition $p_2^{\{1,2\}} \leq A(p_1^{\{1,2\}}) = 1$ is always satisfied. Thus, $H_0^{\{1,2\}}$ is already rejected at the interim analysis, irrespective of the second stage data. As a consequence, the rejection rules for the elementary hypotheses H_0^1 and H_0^2 simplify since now they depend only on the associated conditional error functions $A(p_1^1)$ and $A(p_1^2)$, respectively. For example, H_0^1 is rejected if $p_2^1 \leq A(p_1^1)$ as long as $p_1^{\{1,2\}} \leq \alpha_1$. Conversely, if $p_1^{\{1,2\}} \geq \alpha_0$, then $A(p_1^{\{1,2\}}) = 0$ and H_0^1 cannot be rejected irrespective of the second stage data. In practice, this would lead to an early futility stop of the entire trial after the interim analysis, since by the closure principle neither H_0^1 nor H_0^2 could be rejected. Finally, note that if both p_1^g, $g \in \{1, 2\}$, and $p_1^{\{1,2\}}$ are less than or equal to α_1, H_0^g is already rejected at the interim analysis and there is no need to continue testing H_0^g. Kieser et al. (1999) gave a flow chart to depict graphically the complete decision process for $G = 2$ hypotheses.

In the following we consider two generic examples of adaptively modifying multiple hypotheses after an interim analysis. The first example illustrates the selection of a treatment at interim. The second example considers a treatment switch at interim.

Generic Example 1: Treatment Selection

Assume a two-stage design to compare two treatments with a control. At the interim analysis it is decided which of the two treatments to carry forward to the second stage. The final analysis for the selected treatment includes the patients from both stages by applying an adaptive combination test. Assume without loss of generality that one decides at the interim analysis to continue with treatment 1 and let H_0^1 be the related null hypothesis. No data is therefore available for treatment 2 from the second stage. Consequently, the intersection hypothesis $H_0^{\{1,2\}}$ for the second stage needs to be tested with the test for H_0^1. This is possible because the test for H_0^1 can also be used for $H_0^{\{1,2\}}$: if H_0^1 is rejected, $H_0^{\{1,2\}}$ is rejected in particular. Figure 10.3 depicts the closed test procedure associated with the two null hypotheses H_0^1 and H_0^2 together with the related stage-wise p-values as well as the resulting combination of both stages in terms of conditional error functions.

From the closure principle it follows that we have to reject H_0^1 and $H_0^{\{1,2\}}$ in order to declare treatment 1 to be significantly different from the control. As seen from

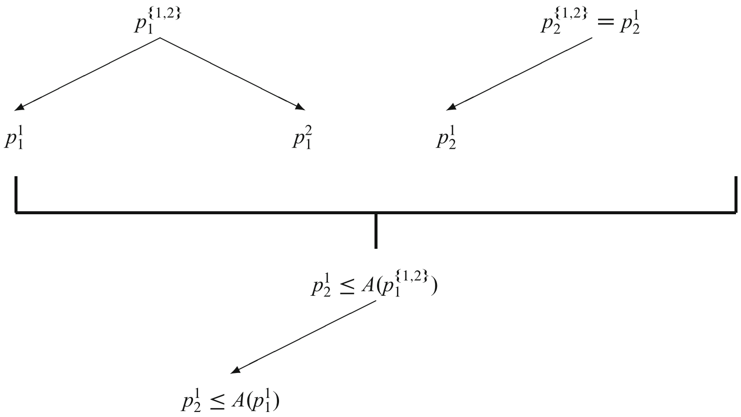

Fig. 10.3 Closed test procedure with treatment selection at interim. *Top*: stage-wise p-values for the hypotheses H_0^1, H_0^2, and $H_0^{\{1,2\}}$, assuming that treatment 1 is selected at the interim analysis. *Bottom*: rejection rule for H_0^1 in the final analysis in terms of conditional error functions (adapted from Bretz et al. 2006)

Fig. 10.3, we thus require that

$$p_2^1 \leq \min\{A(p_1^1), A(p_1^{\{1,2\}})\} \ .$$

Equivalently, we require

$$C(p_1^{\{1,2\}}, p_2^1) \leq c \quad \text{and} \quad C(p_1^1, p_2^1) \leq c$$

for both combination tests for rejecting H_0^1.

Note that the approach above also applies to other adaptive selection problems involving two hypotheses, such as, selecting a suitable population from two prespecified populations. We will generalize and discuss these approaches in Sects. 11.1 and 11.2.

Generic Example 2: Treatment Switch

Assume that a study is planned to investigate a single treatment (1, say) in comparison with a control. Assume further that at the interim analysis safety problems are detected and it is decided to discontinue the present treatment arm. Instead, it is decided to continue the study with a new treatment (2, say, which could, for example, be a lower dose of treatment 1) being investigated at the second stage. Figure 10.4 depicts the resulting closed test procedure associated with the two null hypotheses H_0^1 and H_0^2 being tested in the course of the study.

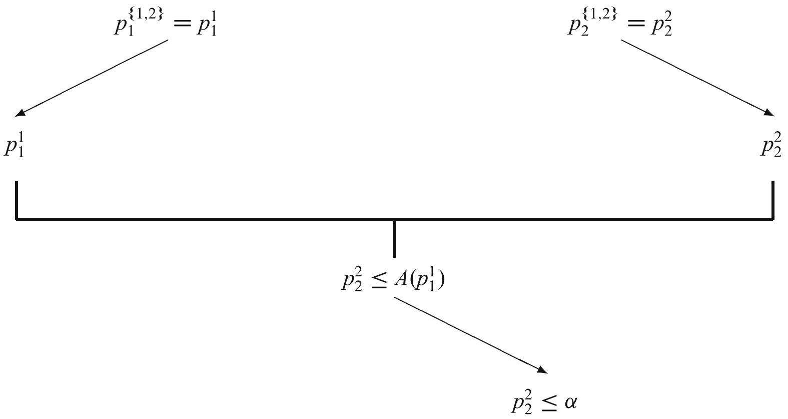

Fig. 10.4 Closed test procedure with treatment switch at interim. *Top*: stage-wise p-values for the hypotheses H_0^1, H_0^2, and $H_0^{\{1,2\}}$, assuming that treatment 1 is discontinued at the interim analysis and treatment 2 is investigated instead during the second stage. *Bottom*: rejection rule for H_0^2 in the final analysis in terms of conditional error functions (adapted from Bretz et al. 2006)

Since at stage 1 no data is available for treatment 2, and vice versa at stage 2 no data is available for treatment 1, the related stage-wise p-values for the intersection hypothesis $H_0^{\{1,2\}}$ are just the corresponding p-values from the elementary hypotheses H_0^1 and H_0^2, respectively. As concluded from Fig. 10.4, H_0^2 is rejected if $p_2^2 \leq \min\{A(p_1^1), \alpha\}$, i.e., if the second stage p-value p_2^2 associated with treatment 2 is less than α and less than the conditional error resulting from the first stage p-value p_1^1 associated with treatment 1.

Note that in this example we use the constant conditional error function α for H_0^2 while we use $A(p_1^1)$ for $H_0^{\{1,2\}}$. Hence, we have here an example where we use different conditional error functions for the different null hypothesis. Note also that, in practice, the treatment switch example will probably never be applied as described here. One would rather stop the entire study after the interim analysis and start a second (seemingly independent) study investigating treatment 2 at full level α. The above considerations are not only instructive for illustrating adaptive combination tests, but they also put the current statistical practice into a different perspective.

Chapter 11
Applications and Case Studies

As noted at the beginning of Chap. 10, we will now apply the general principle of adaptive combination tests to two important situations in clinical trial designs. The first is the design with multiple treatment arms where, based on interim results, one or more arms are selected. The second is the design where one or more prespecified subsets of a population are selected for further investigation, the latter designs are called *adaptive enrichment designs*. The combination testing principle together with the closed testing principle can be used in both settings. We will describe the procedures in detail, particularly, which intersection tests can be used for specific situations and provide examples for the assessment of these designs. We also provide real trial examples to illustrates how these designs were used in practice. We then discuss other types of adaptations that were discussed in the literature. In the last section of this chapter, we briefly discuss the added logistical and regulatory complexity when performing adaptive designs.

11.1 Adaptive Treatment Selection in Multi-Arm Clinical Trials

Group sequential pairwise comparisons in multi-arm clinical trial designs were already considered by Follmann et al. (1994) and Proschan et al. (1994). They derived critical values for the overall test statistic through simulation and showed that, in the all-pairs comparison setting, the procedure preserves the FWER in a strong sense if a treatment arm is dropped at some stage k of the trial if it was shown to be inferior. Hellmich (2001) showed that otherwise (for example, if a superior

G. Wassmer, W. Brannath, *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*, Springer Series in Pharmaceutical Statistics,
DOI 10.1007/978-3-319-32562-0_11

treatment arm is dropped due to severe side effects) the FWER indeed is no more preserved. An alternative approach, where the number of selected treatment arm needs to be specified, was proposed by Stallard and Todd (2003) (see also, Kelly et al. 2005). Friede and Stallard (2008) and Stallard and Friede (2008) compared adaptive designs to group sequential designs with treatment selection. A fully flexible approach is derived from the flexible combination testing (or the CRP) principle together with the closed testing procedure which we will now describe in detail (see also, Chap. 10).

Historically, multiple hypotheses testing within adaptive designs was first developed for the many-to-one comparison setting (Bauer and Kieser 1999; Hommel 2001; Lehmacher et al. 2000; Posch et al. 2005). Multi-arm clinical trials designs with an adaptive interim analysis in order to select treatment arms have also been referred to as adaptive seamless designs (Bretz et al. 2006; Jennison and Turnbull 2007; Maca et al. 2006; Posch et al. 2005; Schmidli et al. 2006; Wassmer 2011; Hampson and Jennison 2015). Recently, the classical group sequential approach has been made flexible using the closed testing and conditional error principle (Magirr et al. 2014b), see also Gao et al. (2014). Furthermore, König et al. (2008) proposed a procedure that is based on the conditional error of the single stage Dunnett test. In the following, these procedures will be described.

11.1.1 Test Procedure

We consider the many-to-one comparison setting, where G experimental treatment groups are tested against a control group. We consider testing means μ_g of normally distributed variables, i.e., we are considering G elementary (null) hypotheses

$$H_0^g : \mu_0 = \mu_g,\ g = 1, \ldots, G,$$

where μ_0 is the mean of the control group and μ_g, $g = 1, \ldots, G$, refer to the means in the active treatment groups. The global hypothesis is

$$H_0 = \bigcap_{g=1}^{G} H_0^g : \mu_0 = \cdots = \mu_G\ .$$

When performing the closed test we will also consider intersection hypotheses

$$H_0^{\mathcal{J}} = \bigcap_{g \in \mathcal{J}} H_0^g,\ \mathcal{J} \subset \{1, \ldots, G\}\ ,$$

which are, for example,

$$H_0^{23} = H_0^2 \cap H_0^3 : \mu_0 = \mu_2 = \mu_3 \ ,$$

and all elementary hypotheses H_0^g (see Sect. 10.2).

The problem of adaptive treatment arm selection is essentially stated as follows: If one or more treatment arms were dropped in an interim analysis and hence the set of initial hypotheses was reduced, usually no data is available from the excluded treatment arm(s). In this case, we define tests for intersection hypotheses involving excluded treatment arms as tests for hypotheses for the non-excluded treatment arms. For example, if one treatment arm, s, is selected out of G treatment arms, all intersection hypotheses contained in H_0^s can be tested with a level alpha test for H_0^s. This is a valid test procedure since $H_0^{\mathcal{J}} \subseteq H_0^s$ for all $s \in \mathcal{J}$ or equivalently: if H_0^s is rejected, $H_0^{\mathcal{J}}$ with $s \in \mathcal{J}$ is rejected in particular, under control of the Type I error rate (compare Generic example 1 and 2 in Sect. 10.3).

Generally, denoting $\mathcal{E} \subset \{1, \ldots, G\}$ the index set of all excluded H_0^g, a test for $H_0^{\mathcal{J}}$ with $\mathcal{J} \cap \mathcal{E} \neq \emptyset$ is performed as a test for $H_0^{\mathcal{J}\setminus\mathcal{E}}$. For example, when using Bonferroni tests, the adjusted p-value for the hypothesis $H_0^{\mathcal{J}}$ at stage k of the group sequential test procedure is given by

$$p_{\mathcal{J},k}^{\text{adj}} = \min\{|\mathcal{J}\setminus\mathcal{E}| \min_{g \in \mathcal{J}\setminus\mathcal{E}} \{p_{g,k}\}, 1\} \ ,$$

where $p_{g,k}$ denotes the p-value for H_0^g at stage k. This, in general, reduces the degree of adjustment that is necessary for testing $H_0^{\mathcal{J}}$. As a consequence, it increases the power of the test procedure when deselecting one or more treatment arms. Note that the use of the Bonferroni test together with the inverse normal combination function is problematic because the adjusted p-value can be equal to 1 which means that the inverse normal combination test cannot reject a hypothesis at a subsequent stage and hence produces an implicit futility rule.

Applying this testing procedure, at the first interim analysis, it is possible to stop the trial while showing significance of one (or more) treatment arms. This might be possible if the closed testing procedure already shows significance for these treatment arms using early efficacy bounds α_1 (see Sect. 10.2). In an interim analysis, it is also possible to stop the trial due to futility arguments. This is usually based on conditional power calculations (see Sect. 7.2). It is expected, however, that the first stage is specifically used to select a treatment arm to be considered in the subsequent stages of the trial and/or to reassess the sample size for the subsequent stages.

Given a combination function C, at the second stage the decision rule can be formulated as follows. Let $p_{\mathcal{J},1}^{adj}$ and $p_{\mathcal{J},2}^{adj}$ denote the first and second stage p-values for elementary or intersection hypotheses tests. At the second stage, the hypotheses belonging to a selected treatment arm s is rejected if

$$\max_{\mathcal{J}\ni s} C(p^{\text{adj}}_{\mathcal{J},1}, p^{\text{adj}}_{\mathcal{J}\setminus\mathcal{E},2}) \le c\ , \tag{11.1}$$

where c denotes the critical value for the combination test (see Sect. 6.2). This can be easily extended to designs with three or more stages. With this procedure, it is possible to deselect treatment arms at any stage of the trial. Additionally, other types of adaptation, e.g., a recalculation of the necessary sample size, can be performed. Note that inconsonance can occur which means that after the rejection of the global null hypothesis no subset or single hypothesis can be rejected (Friede and Stallard 2008). We illustrate this later on by an example.

If one treatment arm, s, is selected, condition (11.1) reduces to

$$\max_{\mathcal{J}\ni s} C(p^{\text{adj}}_{\mathcal{J},1}, p_{s,2}) \le c\ . \tag{11.2}$$

For $G = 3$ and $s = 3$, the latter case is illustrated in Fig. 11.1, the curved arrows indicating which combination tests have to be carried out and showing (11.2) in order to show significance of the selected treatment arm.

This procedure can be easily generalized for the multi-stage case using appropriate more general combination functions (see Sect. 6.5). We also note that this procedure might come with a loss in power as compared to a procedure which is based on the intersection tests for the pooled data (not a combination test of intersection tests) if no adaptation were performed. It has the advantage, however, to enable design adaptations (including treatment arm selection), thereby reducing the necessary adjustment for the subsequent stages and improving power.

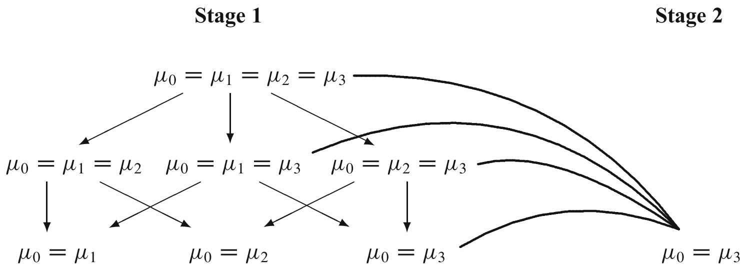

Fig. 11.1 Closed system of hypotheses for $G = 3$ if treatment arm 3 referring to hypothesis H_0^3 is selected for the second stage. The *arrows* indicate logical implications for hypotheses, the *solid curves* indicate combination tests to be performed to show significance for H_0^3 (adapted from Wassmer 2011)

11.1.2 Intersection Tests

The choice of testing the intersection hypotheses is free. The resulting p-values of the global and intersection tests will be used in the adaptive setting where they will be combined over the stages. In the following, five tests for the intersection hypotheses will be defined. We consider the first stage of the test procedure where all G hypotheses are tested. For the following stage(s), the set of indices for the selected treatment arms is a subset of $\mathcal{G} = \{1, \ldots, G\}$.

Let p_g denote the p-value for testing H_0^g, $g = 1, \ldots, G$ and let $p_{(1)} \leq \ldots \leq p_{(G)}$ denote the ordered p-values of the G comparisons. Finally, let $\mathcal{J} \subset \mathcal{G}$, and $|\mathcal{J}|$ be defined as the number of all indices $g \in \mathcal{J}$.

Dunnett Test for Many-to-One Comparisons

Let $\bar{x}_0$ and $\bar{x}_g$ denote the sample means in the control group and the experimental treatment arm g, respectively, and let $\hat{\sigma}$ denote the residual standard deviation estimate.

The adjusted p-value for testing the global hypothesis H_0 is calculated through

$$p^{\text{adj}} = 1 - F_{\boldsymbol{\Sigma}, \sum_{g=0}^{G}(n_g - 1)}\left(\max_{g \in \mathcal{G}} \frac{\bar{x}_g - \bar{x}_0}{\hat{\sigma}\sqrt{1/n_0 + 1/n_g}}\right), \tag{11.3}$$

where $F_{\boldsymbol{\Sigma}, \sum_{g=0}^{G}(n_g-1)}(\cdot)$ denotes the value of the cdf of the multivariate t distribution with correlation matrix $\boldsymbol{\Sigma}$ and $\sum_{g=0}^{G}(n_g - 1)$ degrees of freedom when all G arguments are equal. The elements of the correlation matrix $\boldsymbol{\Sigma}$ are $\sigma_{gg'} = \varsigma_g \varsigma_{g'}$ for $g \neq g'$ where

$$\varsigma_g = \sqrt{\frac{n_g}{n_0 + n_g}}, \; g = 1 \ldots, G. \tag{11.4}$$

This specific multivariate t distribution (with correlation matrix having product correlation structure) is known as the Dunnett distribution and the corresponding test is the Dunnett test (Genz and Bretz 2009). The test can be used for normally distributed observations with comparing means in the many-to-one comparisons setting. The sample sizes in the treatment arms need not be the same. The distribution was also derived for the known variance case. In an approximate sense, the latter can also be used for comparing rates or time to event data in survival designs (Follmann et al. 1994; Proschan et al. 1994).

Corresponding adjusted p-values for a subset $\mathcal{J} \subset \mathcal{G}$ are given by

$$p_{\mathcal{J}}^{\text{adj}} = 1 - F_{\boldsymbol{\Sigma}_{\mathcal{J}}, \sum_{g=0}^{G}(n_g - 1)} \left(\max_{g \in \mathcal{J}} \frac{\bar{x}_g - \bar{x}_0}{\hat{\sigma}\sqrt{1/n_0 + 1/n_g}} \right) ,$$

where $\boldsymbol{\Sigma}_{\mathcal{J}}$ is the corresponding submatrix of the matrix $\boldsymbol{\Sigma}$.

Bonferroni Test for Many-to-One Comparisons

Using the Bonferroni test, the adjusted p-value for testing a hypothesis $H_0^{\mathcal{J}}$, $\mathcal{J} \subseteq \mathcal{G}$, is given by

$$p_{\mathcal{J}}^{\text{adj}} = \min\{|\mathcal{J}| \min_{g \in \mathcal{J}}\{p_g\}, 1\} .$$

We note that this adjusted p-value may become equal to 1. Hence, when using it with the inverse normal method we may obtain combination test statistics (6.7) equal to 0 by just one of the two adjusted p-values. This implies, for example, an implicit futility stopping criterion if the first stage adjusted p-value is equal to 1.

Šidák Method for Many-to-One Comparisons

With the Šidák test, the adjusted p-value for testing a hypothesis $H_0^{\mathcal{J}}$, $\mathcal{J} \subseteq \mathcal{G}$, is given by

$$p_{\mathcal{J}}^{\text{adj}} = 1 - (1 - \min_{g \in \mathcal{J}}\{p_g\})^{|\mathcal{J}|} .$$

With the Šidák test the adjusted p-values are always smaller than 1 and hence, even not with the inverse normal combination function, no implicit early acceptance can occur.

Simes Method for Many-to-One Comparisons

With the Simes intersection test, the adjusted p-value for testing a hypothesis $H_0^{\mathcal{J}}$, $\mathcal{J} \subseteq \mathcal{G}$, is given by

$$p_{\mathcal{J}}^{\text{adj}} = \min_{g \in \mathcal{J}}\{\frac{|\mathcal{J}|}{g} p_{(g_{\mathcal{J}})}\} ,$$

where $p_{(1_{\mathcal{J}})} \le \ldots \le p_{(|\mathcal{J}|_{\mathcal{J}})}$ denote the ordered p-values from the subset $\mathcal{J} \subset \mathcal{G}$.

We note that both Simes and Šidák derived p-values for intersection hypotheses yield valid level α test procedures because the elements of $\boldsymbol{\Sigma}$ are always positive. In this case, under normality and if the residual standard deviation $\hat{\sigma}$ estimate is used for calculating the p-values, the underlying inequalities hold (Sarkar and Chang 1997; Finner et al. 2015; Huque 2016).

A Priori Hierarchical Test

This test assumes a fixed ordering of the hypotheses under consideration with regard to their importance. It is a stepwise procedure starting with the hypothesis of highest order. If this hypothesis is rejected, the hypothesis of second order is tested, and only if this is rejected the next hypothesis is tested, and so on. Generally, a hypothesis is rejected if all hypotheses with higher order are rejected as well. Assume that hypotheses with higher indices have higher importance. Consequently, the adjusted p-value for testing a hypothesis $H_0^{\mathcal{J}}$, $\mathcal{J} \subseteq \mathcal{G}$, is given by

$$p_{\mathcal{J}}^{\mathrm{adj}} = p_{\max\{g \in \mathcal{J}\}} ,$$

where $\max\{g \in J\}$ corresponds to the hypothesis in $\mathcal{J}$ with the highest importance.

11.1.3 Overall p-Values and Confidence Intervals

In the multi-arm testing situation, overall (repeated) p-values for a hypothesis H_0^g are defined as smallest significance level for which the test results yield rejection of the considered (single) hypothesis H_0^g (see Sect. 8.1):

$$p_g^k = \min\{\alpha : H_0^g \text{ can be rejected at stage } k\} .$$

These are generally found by numerical root finding and can be calculated at any stage of the trial with the full adaptive closed test at multiple level α. By definition, at stage k, an overall p-value p_g^k falls below the overall significance level α if and only if the corresponding hypothesis H_0^g can be rejected at stage k. Hence, these p-values account for the sequential adaptive and step-down nature of the closed testing principle. By this they are completely consistent with the test decision.

The calculation of overall confidence intervals is more problematic. In general, confidence intervals based on stepwise testing procedures are difficult to construct. This is a specific feature of multiple testing procedures and not of adaptive or sequential testing. On the other hand, Posch et al. (2005) proposed to construct repeated confidence intervals based on the single-step adjusted overall p-values as follows (see also, Bretz et al. 2009a):

Consider the hypotheses

$$H_0^g(\delta_g) : \mu_g - \mu_0 = \delta_g,\ g = 1, \ldots, G,$$

with corresponding p-values $p_{g,k}(\delta_g)$ at stage k. Essentially, adjusted p-values for testing $H_0^g(\delta_g)$ are defined as functions of $p_{g,k}(\delta_g)$ alone, without taking into account the step-down nature of the testing procedure. For example, for the Bonferroni and

Simes test,

$$p_{g,k}^{\mathrm{adj}}(\delta_g) = \min\{1, |\mathcal{S}_k|\, p_{g,k}(\delta_g)\} ,$$

where $\mathcal{S}_k$ denotes the set of active treatment arms that are selected at stage k (i.e., $\mathcal{S}_1 = \{1, \ldots, G\}$). For the Šidák test,

$$p_{g,k}^{\mathrm{adj}}(\delta_g) = 1 - (1 - p_{g,k}(\delta_g))^{|\mathcal{S}_k|} ,$$

and for Dunnett test

$$p_{g,k}^{\mathrm{adj}}(\delta_g) = 1 - F_{\boldsymbol{\Sigma}_{\mathcal{S}_k}, \sum_{g\in\{0\}\cup\mathcal{S}_k}(n_g-1)} \left(\frac{\bar{x}_{g,k} - \bar{x}_{0,k} - \delta_g}{\hat{\sigma}\sqrt{1/n_{0,k} + 1/n_{g,k}}} \right) .$$

For the a priori hierarchical intersection test, this p-value and hence the repeated confidence interval are not defined.

The confidence intervals are computed separately for each treatment arm g that was selected for stage k. By use of the given combination function C with critical value c the lower bound of the repeated confidence interval for treatment arm $g \in \mathcal{S}_k$ at stage k is found by the values δ_k for which the test yields non-rejection at stage k, i.e.,

$$C(p_{g,1}^{\mathrm{adj}}(\delta_g), \ldots, p_{g,k}^{\mathrm{adj}}(\delta_g)) \geq c . \tag{11.5}$$

The upper bound can be found analogously.

For example, when using the unweighted inverse normal method, the Bonferroni adjustment and one active treatment arm $s \in \mathcal{G}$ selecting for the second stage, the lower bound of the confidence interval for treatment arm s by using (11.5) is obtained by

$$\max\{\delta_s : 1 - \Phi((\Phi^{-1}(1 - \min\{1, G p_{s,1}(\delta_s)\}) + \Phi^{-1}(1 - p_{s,2}(\delta_s)))/\sqrt{2}\big) \geq c\} .$$

If a binding stopping for futility criterion is chosen, the determination is equivalent to the way as described in Sect. 8.2.2.

These RCIs are not, in general, consistent with the test decision. It might happen that, for example, a hypothesis is rejected but the lower bound of the CI is below 0. If only the treatment arm with the smallest p-value was selected in the first stage of the trial, however, the resulting confidence intervals are completely consistent with the test decision. Simultaneous confidence intervals that are in general consistent with adaptive closed tests are provided in Magirr et al. (2013). We further note that these RCIs may fail to become narrower with increasing sample size for the second or subsequent stages. This is the case if the adjustment leads to an implicit futility stopping rule like with the Bonferroni adjustment and the inverse normal method.

11.1.4 Numerical Example

We illustrate the performance of the test procedure and the way to calculate overall p-values and confidence intervals for a two-stage adaptive design with the use of the unweighted inverse normal method. As for the example in Sect. 8.2.4 we use critical values according to the Wang and Tsiatis class with $\Delta = 0.25$ at a one-sided significance level $\alpha = 0.025$. These are given by $\alpha_1 = 0.00768$ and $c = 0.0208$ (or $u_1 = 2.4239$ and $u_2 = 2.0382$). We consider testing the mean μ of normally distributed observations in $G = 3$ active treatment arms and one control arm. The variance, σ^2, is assumed to be unknown. Suppose the study was planned with sample size $n_1 = n_2 = 20$ per stage and treatment arm and suppose that at the first stage of the trial the observation as summarized in Table 11.1 were made.

The p-value of the Dunnett test for testing the global hypothesis $H_0 : \mu_0 = \mu_1 = \mu_2 = \mu_3$ is 0.0224 which is calculated from the Dunnett t distribution with $\max_{g \in \mathcal{G}} t_g = 2.442$, $df = 80$, and $\varsigma_g = \sqrt{n_g/(n_0 + n_g)} = 0.6708, 0.7149, 0.6988$ from (11.4). This can be done, for example, with the R package mvtnorm (Genz et al. 2014) using the syntax (note that a slightly different value can be obtained which is due to the Monte Carlo integration error)

```
corr <- diag(3)
corr[2,1] <- 0.6708*0.7149
corr[3,1] <- 0.6708*0.6988
corr[3,2] <- 0.7149*0.6988
p <- 1 - pmvt(lower=-Inf, upper=2.442, delta=rep(0, 3),
  df=80, corr = corr)
```

Since $\Phi^{-1}(1 - 0.0224) = 2.006 < u_1$, the global hypothesis cannot be rejected and therefore none of the hypothesis $H_0^g : \mu_0 = \mu_g, g = 1, 2, 3$.

Note that the same holds true if we would have pre-planned the Bonferroni, Šidák, or Simes test, respectively, for testing the global intersection hypothesis. For the Bonferroni test, the p-value for testing H_0 is

$$p_{\mathcal{G}}^{\text{adj}} = \min\{3 \times 0.0084, 1\} = 0.0252\ ,$$

Table 11.1 First stage results (means $\bar{x}_g$, standard deviation $\hat{\sigma}_g$, sample sizes n_g, test statistics t_g, and p-values p_g, p-values p_g, $g = 0, 1, 2, 3$, the latter two based on an overall estimate of the standard deviation which is $\hat{\sigma} = 2.197$) of a hypothetical trial with three treatment arms and a control

Treatment arm	0	1	2	3
$\bar{x}_g$	1.1	1.4	2.7	2.6
$\hat{\sigma}_g$	1.8	2.1	2.5	2.3
n_g	22	18	23	21
t_g	–	0.430	2.442	2.237
p_g	–	0.3343	0.0084	0.0140

for the Šidák test, the p-value is

$$p_{\mathcal{G}}^{\text{adj}} = 1 - (1 - 0.0084)^3 = 0.0250\,,$$

and for the Simes test, the p-value is

$$p_{\mathcal{G}}^{\text{adj}} = \min_{g \in \mathcal{G}}\{3 \times 0.0084, 3/2 \times 0.0140, 0.3343\} = 0.0210\,.$$

For all these tests the p-value is larger than α_1 or, equivalently, the inverse normal transformation is smaller than u_1. Note that even if the second treatment arm was fixed as the arm that is related to the hypothesis with highest order, a hierarchical test does not achieve significance at the first stage.

Obviously, the effect in the first treatment arm is so small that the arm should be dropped for further analysis. The effects in the other treatment arms are of similar magnitude such that there is no clear preference which treatment arm to select. For example, if treatment arms refer to increasing doses of a drug, choosing the second arm might be preferable because it is undesirable to administer unnecessarily high doses. On the other hand, if there are no serious safety concerns, it might be reasonable to select both treatment arms for stage 2. It is important to recognize that the adaptive procedure allows a free selection of treatment arms and hence all options are valid choices.

Assume that treatment arms 2 and 3 are selected for the second stage. Since the effect sizes in both treatment arms yield small p-values there is no reason to increase the sample size per treatment arm for the second stage. We therefore assume that the sample sizes are again planned to be 20 per treatment arm at stage 2. Let us assume the stage 2 results are as summarized in Table 11.2.

The p-value of the Dunnett test for the second stage data alone (for testing the hypothesis $H_0^{23} : \mu_0 = \mu_2 = \mu_3$) is 0.1219 which is calculated with the Dunnett t cdf with $\max_{g \in \mathcal{G} \setminus \mathcal{E}} t_g = 1.479$, where $\mathcal{E} = \{1\}$, $df = 55$, and $\varsigma_g = \sqrt{n_g/(n_0 + n_g)} = 0.7264, 0.7511$ from (11.4).

Table 11.2 Second stage results (means $\bar{x}_g$, standard deviation $\hat{\sigma}_g$, sample sizes n_g, test statistics t_g, and p-values p_g, $g = 0, 2, 3$, the latter two based on an overall estimate of the standard deviation which is $\hat{\sigma} = 2.026$) of a hypothetical trial with three treatment arms and a control if treatment arm 1 is deselected

Treatment arm	0	1	2	3
$\bar{x}_g$	1.3	–	2.3	2.1
$\hat{\sigma}_g$	2.1	–	2.2	1.8
n_g	17	–	19	22
t_g	–	–	1.479	1.223
p_g	–	–	0.0725	0.1133

Now, $(\Phi^{-1}(1-0.0224)+\Phi^{-1}(1-0.1219))/\sqrt{2} = 2.242 > 2.038$, and therefore the global null hypothesis can be rejected. For performing the closed testing procedure, we also calculate the inverse normal test statistics for testing the hypotheses in the closed system of hypotheses. We obtain

- for the hypothesis $H_0^{12}: \mu_0 = \mu_1 = \mu_2$:
$$(\Phi^{-1}(1-0.0158)+\Phi^{-1}(1-0.0725))/\sqrt{2} = 2.550,$$
- for the hypothesis $H_0^{13}: \mu_0 = \mu_1 = \mu_3$:
$$(\Phi^{-1}(1-0.0261)+\Phi^{-1}(1-0.1133))/\sqrt{2} = 2.228,$$
- for the hypothesis $H_0^{23}: \mu_0 = \mu_2 = \mu_3$:
$$(\Phi^{-1}(1-0.0157)+\Phi^{-1}(1-0.1219))/\sqrt{2} = 2.345,$$
- for the hypothesis $H_0^{2}: \mu_0 = \mu_2$:
$$(\Phi^{-1}(1-0.0084)+\Phi^{-1}(1-0.0725))/\sqrt{2} = 2.721,$$
- for the hypothesis $H_0^{3}: \mu_0 = \mu_3$:
$$(\Phi^{-1}(1-0.0140)+\Phi^{-1}(1-0.1133))/\sqrt{2} = 2.408.$$

Note that for the hypotheses $H_0^{12}: \mu_0 = \mu_1 = \mu_2$ and $H_0^{13}: \mu_0 = \mu_1 = \mu_3$ the second stage p-value for testing $H_0^2: \mu_0 = \mu_2$ and $H_0^3: \mu_0 = \mu_3$, respectively, are used. The first stage p-values 0.0158, 0.0261, and 0.0157 are obtained with the Dunnett t distribution with $df = 80$ and appropriate correlation structure, for example, 0.0158 is from the Dunnett t cdf with $\max_{g\in\{1,2\}} t_g = 2.442$, $df = 80$, and $\varsigma_g = \sqrt{n_g/(n_0+n_g)} = 0.6708, 0.7149$.

All test statistics exceed the critical boundary, 2.038, and therefore the hypotheses $H_0^2: \mu_0 = \mu_2$ and $H_0^3: \mu_0 = \mu_3$ can be rejected. We leave as an exercise to check that the Simes test as well as the Bonferroni or the Šidák method yield the same results.

The overall repeated p-values can be found by a bisection search yielding the values $p_2^2 = 0.0147$ for the hypothesis belonging to the second treatment arm, and $p_3^2 = 0.0157$ for the hypothesis belonging to the third treatment arm. Since both are smaller than 0.025, this corresponds to the test decision. For an overall significance level of 0.0147, the critical value for the second stage within the Wang and Tsiatis class with $\Delta = 0.25$ is $u_2 = 2.242$. Thus, the global hypothesis' test statistic is on the boundary. Furthermore, all test statistics for hypotheses containing the hypothesis belonging to the second treatment arm can be rejected, too. For an overall significance level of 0.0157, $u_2 = 2.217$ and thus the p-value for $H_0^{13}: \mu_0 = \mu_1 = \mu_3$ is also on the boundary. This illustrates that the overall p-value accounts for the step-down nature of the closed testing procedure yielding an exact correspondence to the test decision.

Unfortunately, this is not true for the repeated confidence intervals. Applying the method described in Sect. 11.1.3, a bisection search yields the 95 % confidence intervals (0.0925; 2.53) and (–0.0602; 2.37) for the second and third treatment arm effect size, respectively. Although, using the closed test procedure, both related hypotheses could be rejected, the confidence interval related to the third treatment arm contains the null value. This is an undesirable property of the described confidence intervals that follows from the step-down nature of the closed testing approach.

If one treatment arm with the largest test statistic is selected for the second stage, the rejection of the global hypothesis implies the rejection of the hypothesis related to the selected hypothesis. This is true for the Dunnett test as well as for the other intersection tests (and clearly for the hierarchical test procedure, too). In general, however, one has to calculate the p-values of all intersection tests within the closed system of hypotheses in order to decide which hypotheses can be rejected. Brannath and Bretz (2010) derived an algorithm to construct shortcuts and illustrated it with several applications.

We finally note a property of the test procedure that might be considered as undesirable. Suppose, for example, that the third treatment arm was dropped for further analysis, for example, due to safety concerns. Suppose further that for the third treatment arm and also for testing the intersection hypothesis H_0^{13} , p-values less than $0.00768 = \alpha_1$ were obtained. Furthermore, assume that significance for testing the global hypothesis was not achieved yet. So we proceed to the second stage as described before and obtain significance for testing the hypothesis belonging to the second treatment arm. Because the global hypothesis can be rejected now, and all intersection hypotheses belonging to the third treatment arm can be rejected, too, at the second stage the hypothesis related to treatment arm 3 can be rejected although no additional data was observed. This looks curious but is a described and well-known property of stepwise testing in multiple testing theory.

11.1.5 Adaptive Dunnett Test

König et al. (2008) proposed a procedure that is based on applying the CRP principle as described in Sect. 6.3.3. They showed that this procedure uniformly improves the Dunnett test if treatment arms where selected in an interim stage. The test coincides with the classical Dunnett test if no treatment arm selections (or other adaptations) were performed. Application within the closed testing procedure is straightforward. The procedure is different to the inverse normal method, when a Dunnett test is used for testing intersection hypotheses, and is generally more powerful. We will see, however, that the gain in power is only small.

The proposed test is a two-stage procedure where in the interim analysis the conditional Type I error rate for the Dunnett test is calculated. Assume that the variance σ^2 is known and treatment means $\bar{x}_0^{(1)}, \ldots, \bar{x}_G^{(1)}$ from the first stage data with sample sizes $n_0, \ldots, n_G$ are calculated. Assume that the information rates that relate the first stage sample size to the pre-planned overall sample size N_g in treatment group g are $t_{1g} = n_g/N_g$, $g = 1, \ldots, G$. The conditional Type I error rate is then given by

$$\begin{aligned}\alpha_{\mathcal{G}} &= \alpha_{\mathcal{G}}(z_1^{(1)}, \ldots, z_G^{(1)}) \\ &= 1 - \int_{-\infty}^{\infty} \prod_{g \in \mathcal{G}} \Phi\left(\frac{c_G(\alpha) - \sqrt{t_{1g}} z_g^{(1)} + \sqrt{1 - t_{1g}}\, \varsigma_g\, x}{\sqrt{(1 - t_{1g})(1 - \varsigma_g^2)}}\right) \varphi(x)\, dx\,,\end{aligned} \tag{11.6}$$

where

$$\begin{aligned}\varsigma_g &= \sqrt{\frac{n_g}{n_0 + n_g}}\,, \\ z_g^{(1)} &= \frac{\bar{x}_g^{(1)} - \bar{x}_0^{(1)}}{\sigma(\sqrt{1/n_g + 1/n_0})} = \frac{\bar{x}_g^{(1)} - \bar{x}_0^{(1)}}{\sigma} \varsigma_g \sqrt{n_0}\,,\ g = 1, \ldots, G,\end{aligned}$$

and $c_G(\alpha)$ is the critical value for performing the Dunnett test at level α with G active treatment arms.

Formula (11.6) needs to be computed with a numerical integration method. Before we show (11.6) we first derive the formula for the overall Type I error rate of the Dunnett test:

Unconditionally, under H_0, for the maximum of standard normally distributed test statistics $Z_1^{(2)}, \ldots, Z_G^{(2)}$, that are the summarized test statistics for the treatment arms,

$$\begin{aligned}&P(\max\{Z_1^{(2)}, \ldots, Z_G^{(2)}\} \le c) \\ &= P\left(\bigcap_{g \in \mathcal{G}} \{Z_g^{(2)} \le c\}\right) \\ &= P\left(\frac{\bar{x}_g^{(2)} - \bar{x}_0^{(2)}}{\sigma} \varsigma_g \sqrt{N_0} \le c\,,\ g = 1 \ldots, G\right) \\ &= P\left(\frac{\bar{x}_g^{(2)}}{\sigma} \sqrt{N_0} - \underbrace{\frac{\bar{x}_0^{(2)}}{\sigma} \sqrt{N_0}}_{\sim N(0;1)} \le \frac{c}{\varsigma_g}\,,\ g = 1 \ldots, G\right) \\ &= \int_{-\infty}^{\infty} \prod_{g \in \mathcal{G}} P\left(\frac{\bar{x}_g^{(2)}}{\sigma} \sqrt{N_0} \le \frac{c}{\varsigma_g} + x\right) \varphi(x)\, dx\end{aligned}$$

$$= \int_{-\infty}^{\infty} \prod_{g \in \mathcal{G}} P\left(\frac{\bar{x}_g^{(2)}}{\sigma} \underbrace{\frac{\varsigma_g}{\sqrt{1-\varsigma_g^2}} \sqrt{N_0}}_{=\sqrt{N_g}} \leq \frac{c}{\sqrt{1-\varsigma_g^2}} + x \frac{\varsigma_g}{\sqrt{1-\varsigma_g^2}} \right) \varphi(x)\, dx$$

$$= \int_{-\infty}^{\infty} \prod_{g \in \mathcal{G}} \Phi\left(\frac{c + \varsigma_g x}{\sqrt{1-\varsigma_g^2}} \right) \varphi(x)\, dx \, .$$

We now turn to the proof of (11.6). Since

$$Z_g^{(2)} = \sqrt{t_{1g}} Z_g^{(1)} + \sqrt{1-t_{1g}} Z_g^2 \, , \tag{11.7}$$

where Z_g^2 is the observed test statistic from the second stage data and $Z_g^{(2)}$ is the observed overall test statistic for treatment arm g at the final stage, conditional on the first stage test statistic,

$$\begin{aligned} \alpha_{\mathcal{G}} &= P(\max\{Z_1^{(2)}, \ldots, Z_G^{(2)}\} \geq c_G(\alpha) \mid z_1^{(1)}, \ldots, z_G^{(1)}) \\ &= 1 - P\left(\bigcap_{g \in \mathcal{G}} \{ Z_g^2 \leq \frac{c_G(\alpha) - \sqrt{t_{1g}} z_g^{(1)}}{\sqrt{1-t_{1g}}} \} \right) \, , \end{aligned}$$

and therefore, by the same arguments as above, we obtain

$$\begin{aligned} \alpha_{\mathcal{G}} &= 1 - \int_{-\infty}^{\infty} \prod_{g \in \mathcal{G}} \Phi\left(\frac{\frac{c_G(\alpha) - \sqrt{t_{1g}} z_g^{(1)}}{\sqrt{1-t_{1g}}} + \varsigma_g x}{\sqrt{1-\varsigma_g^2}} \right) \varphi(x)\, dx \\ &= 1 - \int_{-\infty}^{\infty} \prod_{g \in \mathcal{G}} \Phi\left(\frac{c_G(\alpha) - \sqrt{t_{1g}} z_g^{(1)} + \sqrt{1-t_{1g}}\, \varsigma_g x}{\sqrt{(1-t_{1g})(1-\varsigma_g^2)}} \right) \varphi(x)\, dx \, . \end{aligned}$$

At the interim stage, treatment selection and/or sample size recalculation (for example, based on conditional power) can be performed. If the second stage test is a test at level $\alpha_{\mathcal{G}}$, the procedure preserves the overall level. Assume a set $\mathcal{S} \subseteq \{1, \ldots, G\}$ of treatment arms were selected for the second stage. König et al. (2008) proposed two tests for the global test of $H_0 \cap H_0^{\mathcal{S}}$ at the final stage. The first is to perform a conditional second stage Dunnett test at level $\alpha_{\mathcal{G}}$, the second is based on a corresponding unconditional test.

For the conditional second stage Dunnett test, the p-value at the final analysis is given by

$$1 - \int_{-\infty}^{\infty} \prod_{g \in \mathcal{S}} \Phi \left(\frac{\max_{g \in \mathcal{S}} z_g^{(2)} - \sqrt{t_{1g}} z_g^{(1)} + \sqrt{1 - t_{1g}}\, \varsigma_g\, x}{\sqrt{(1 - t_{1g})(1 - \varsigma_g^2)}} \right) \varphi(x)\, dx\,. \tag{11.8}$$

(11.8) is similar to (11.6), with $c_G(\alpha)$ replaced by the observed maximum statistic $\max_{g \in \mathcal{S}} z_g^{(2)}$ where $z_g^{(2)}$ is calculated as in (11.7) with the pre-planned weights t_{1g}, and the integration performed over the selected treatment arms.

If the unconditional test is used, the p-value is calculated from the second stage data alone and given by

$$1 - \int_{-\infty}^{\infty} \prod_{g \in \mathcal{S}} \Phi \left(\frac{\max_{g \in \mathcal{S}} z_g^2 + \varsigma_g\, x}{\sqrt{1 - \varsigma_g^2}} \right) \varphi(x)\, dx\,, \tag{11.9}$$

where ς_g can be different from stage 1.

Application of the tests based on (11.6)–(11.9) within the closed testing procedure as described in Sect. 11.1.1 yields a multiple testing procedure for the considered closed system of hypotheses. That is, for each intersection hypothesis $H_0^{\mathcal{J}}$, $\mathcal{J} \subset \mathcal{G}$ the conditional error $\alpha_{\mathcal{J}}$ is calculated. $H_0^{\mathcal{J}}$ is rejected if the p-value is smaller than $\alpha_{\mathcal{J}}$ where the test for $H_0^{\mathcal{J}}$ with $\mathcal{S} \subset \mathcal{J}$ is formally performed as test for $H_0^{\mathcal{J} \cap \mathcal{S}}$.

The test based on (11.8) coincides with the classical (single stage) Dunnett test if no treatment arms were selected and no sample size adaptations were conducted. If only one treatment arm is selected for the second stage, it is easy to see that (11.8) coincides with (11.9) (König et al. 2008) and so the procedures are exactly the same. If sample size recalculations were performed the test based on (11.8) assumes the ς_g, $g = 1, \ldots, G$, to be constant over the stages, otherwise the test based on (11.9) should be used (where ς_g in (11.9) is obtained from the second stage data). Both tests require the variance to be known and therefore the sample sizes to be quite large in order to obtain valid approximate results. If this is not the case the conditional error rate is difficult to calculate and relies on additional assumptions (Posch et al. 2004; Timmesfeld et al. 2007; Gutjahr et al. 2011). Note also that no formal stopping rules are foreseen, i.e., the interim analysis is carried out solely to perform a data-dependent treatment arm selection and/or a sample size recalculation. If stopping rules are taken into account it is straightforward though computationally cumbersome to derive adjusted limits for the test decision. An alternative is to use the separate stage p-values from the Dunnett test and combine them with the use of a combination function. The latter procedure has the additional advantage to account for the unknown variance case if the residual variance estimate per stage is used. The use of both (11.3) and (11.9) provides exact Type I error control that can be used within the closed test procedure.

Power Comparison

An interesting question is whether the use of the CRP principle considerably improves the power of the adaptive design. We will compare three procedures for a multi-arm design with $G = 3$ treatment arms at the first stage at given overall significance level $\alpha = 0.025$. For a more extensive comparison of the conditional Dunnett test with other approaches, see Friede and Stallard (2008).

The first two procedures are the ones that use the conditional error rate (11.6) and use either the conditional p-value (11.8) or the unconditional p-value (11.9). We compare these two with the inverse normal combination test where at each stage the Dunnett test is used (assuming the variance to be known) and no interim stops are foreseen. We assume that the treatment arms are corresponding to a linear dose–response relationship with standardized slope δ. That is, for example, if $\delta = 0.3$, the standardized treatment arm effects are 0.10, 0.20, and 0.30 for the first, second, and third treatment arm, respectively; if $\delta = 0.6$, the effects are 0.20, 0.40, and 0.60, and so on. We consider standardized slopes δ ranging from 0 to 1. For fixed sample sizes $n = 20$ for each treatment arm and stage, this provides power values (defined as the probability of declaring at least one treatment arm as effective) in a reasonable range. We consider four rules for selecting the treatment arm(s) for the second stage:

1. select the best,
2. select the best and the second best,
3. select all,
4. select the best and the arm(s) that are not worse than $\epsilon = 0.25$ than the best,

where "best" is defined in terms of the observed effect which is the difference between the means. The last selection rule was proposed in Kelly et al. (2005) and used in Friede and Stallard (2008). Here, the number of actually selected treatment arms is flexible, and hence the overall sample size is random. Since no other adaptation than treatment arm selection is performed, we use constant weights for the inverse normal combination function and $t_{1g} \equiv 0.50$ in (11.6). Figure 11.2 illustrates the power of the three procedures for the four selection rules.

As shown in König et al. (2008), the conditional Dunnett test performs best for all situations. However, for the considered situations, the gain in power as compared to the other procedures is only small. If one treatment arm is selected, there is even no remarkable difference at all as compared to the combination test (the two conditional Dunnett tests coincide). Only if no treatment is dropped, the "pure" conditional Dunnett test considerably outperforms the other two tests and it seems that the combination test procedure is very similar to the separate stage conditional Dunnett test. The last selection rule is some kind of mixture between the other three selection rules and might represent a practically relevant situation. Accordingly, the conditional Dunnett performs best though only to a small amount. We note that the use of this procedure specifically assumes the allocation ratios to be the same over the stages and the variance to known. It is therefore questionable if the smaller power gain compensates the more relevant disadvantages in practicability.

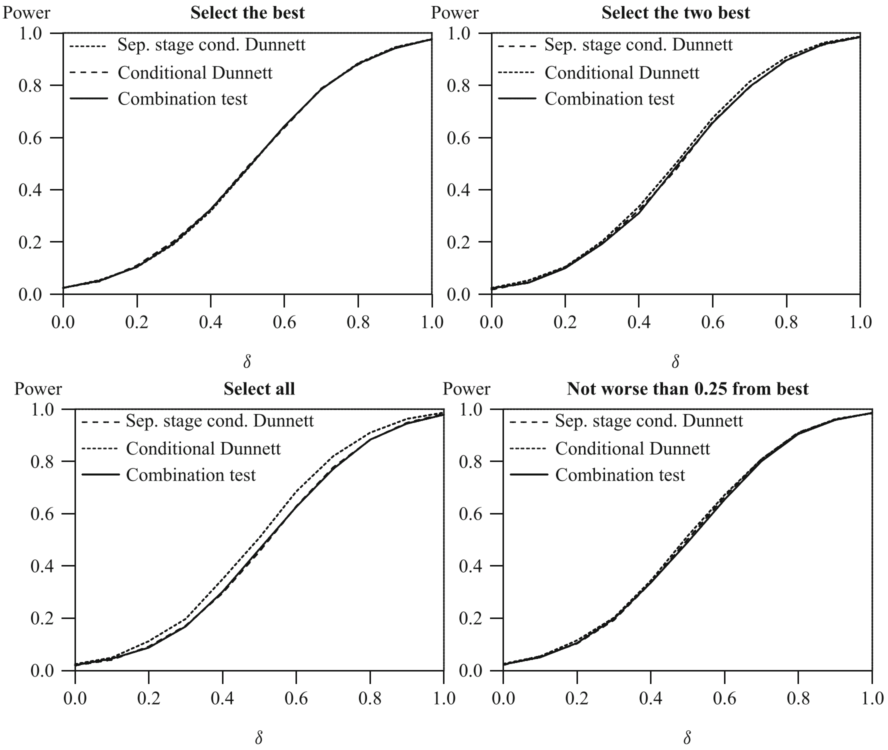

Fig. 11.2 Power (reject at least one hypothesis) in dependence of $\delta = \mu/\sigma$ of three adaptive multi-arm design procedures, for four selection rules as explained in the text

Here, we only consider one effect size pattern which is defined through a linear relationship between treatment arms and effect sizes. We can also consider other relationships, for example, defined through an exponentially shaped parameter curve, an umbrella like shape, a logistic shape, a sigmoid Emax shape, or a parameter shape with constant effect sizes for all treatment arms, for details see, for example, Bretz et al. (2005, 2009a) or Dragalin et al. (2007). Generally, the performance of the procedure depends also on other issues, such as the selection rule for the selected treatment arm(s), or on the sample size reassessment rule at the interim stages. We also note that we used a simple definition of power in this simulation. If there are multiple hypotheses, there are different ways to define the power (Senn and Bretz 2007).

An interesting selection rule is the r-s-selection rule which was originally suggested by Bretz and Maurer (unpublished) (Bretz et al. 2009a). The number and the way to select the treatment arms are not fixed, instead this is done according to specified probabilities. The selection rule is defined by two probability vectors $\mathbf{r} = (r_1, \dots, r_G)$ and $\mathbf{s} = (s_1, \dots, s_G)$ with $\sum_{g=1}^{G} r_g = \sum_{g=1}^{G} s_g = 1$ and the element r_g and s_g, $g = 1, \ldots, G$ defined by

$$\begin{aligned} r_g &= \text{P(Select } g \text{ treatment arms)}\ , \\ s_g &= \text{P(Start selection at the } g\text{th best treatment arm)}\ . \end{aligned}$$

For example, $r_g = 1$ and $r_{g'} = 0$ for $g' \neq g$, and $s_1 = 1$ and $s_{g'} = 0$ for $g' \neq 1$, results in the g best treatment arms selection rule. This rule might represent the variety of selection rules applied in different clinical studies and hence might realistically represent a company's strategy to select treatment arms (Bretz and Wang 2010).

11.1.6 Other Endpoints

So far, the described procedures assume normally distributed endpoints. In this case, the individual intersection tests are either exactly exhausting the Type I error rate (for example, when the Dunnett t test within the inverse normal combination test is used) or conservative. For example, the Bonferroni, Simes, and Šidák based procedures are generally conservative. Note that the estimation of the variance for the conditional Dunnett test in Sect. 11.1.5 involves a slight anti-conservatism and therefore assumes the sample sizes to be large enough in real applications.

As described in Sect. 5.2, the group sequential theory can be easily adapted for the use of binary data where the test statistic (5.13) for testing

$$H_0^g : \pi_0 = \pi_g,\ g = 1, \ldots, G,$$

can be repeatedly used for group sequential testing. This provides approximate Type I error rate control. It is now straightforward to calculate the test statistic (5.13) from the stage-wise data and to combine the stages with a suitable combination test. Furthermore, for the many-to-one treatment arm comparison, this can be done pairwisely such that this procedure can be used within the closed test procedure as described above. It approximately controls the Type I error rate even for the Dunnett intersection test if the correlation is calculated as in the normal case with the use of (11.4), and df is set to infinity. Also, the conditional Dunnett test can be used analogously. The calculation of confidence intervals is possible with the use of a combination test and an appropriately defined test statistic or p-value $p_{g,k}^{\text{adj}}(\delta_g)$. As already mentioned in Sect. 5.2, a number of approaches are available for this which can be used for the adaptive combination test. Note also that it is possible to use Fisher's exact test for the pairwise comparisons per stage, however, this test turns out to be very conservative within the combination test approach.

For survival trials, in Sect. 9.1 we described how to use the combination testing principle. For many-to-one comparisons this can be done pairwisely by testing

$$H_0^g : \omega^g = 1, \; g = 1, \ldots, G.$$

Denoting $d_{g,k}$, $g = 1, \ldots, G$, $k = 1, \ldots, K$, the number of events in the comparison of treatment group g with the control group up to stage k,

$$Z_{g,k} = \frac{\sqrt{d_{g,k}}\, LR^*_{g,k} - \sqrt{d_{g,k-1}}\, LR^*_{g,k-1}}{\sqrt{d_{g,k} - d_{g,k-1}}}\ , \; g = 1, \ldots, G, \; k = 1, \ldots, K,$$

is the approximately independent increment of the log-rank statistic for analysis set g in stage k. This test statistic can be used within the combination test yielding approximate control of Type I error rate in the adaptive case (Wassmer 2006).

Therefore, in principle there is no problem to apply the methods as described before for survival trials. Also the Dunnett test can be used with the stage-wise test statistics, and a suitable estimate of the correlation is given by (11.4) where the sample sizes are replaced by the corresponding stage-wise events. A problem arises if not only the test statistic (or the p-value) itself but also a correlated endpoint is used for the design adaptation at interim. We already described the problem in Sect. 9.2 and gave some relevant literature. The use of correlated information, however, is typical in survival trials with treatment arm selection. Here it often happens that the selection of a treatment arm at an interim analysis is based on the efficacy and safety for a surrogate parameter or some other correlated endpoint, rather than the clinical (primary) endpoint alone. This is because the primary outcome is usually assumed to be a long-term endpoint and hence the selection procedure would only be performed at a late stage of the trial if based on this endpoint. Hence, it is reasonable to use an endpoint with an earlier availability for the selection of the treatment arm. Recently, some proposals were made to overcome the problem that Type I error rate control cannot be guaranteed anymore with the naïve use of the closed adaptive test. Essentially, proposals are either based on a modification of the combination testing principle or the CRP approach (Jenkins et al. 2011; Magirr et al. 2014a; Mehta et al. 2014; Irle and Schäfer 2014; Stallard et al. 2014; Carreras et al. 2015) or requiring additional assumptions regarding the joint distribution of the primary and the short-term endpoints (Di Scala and Glimm 2011; Stallard 2010). There is intensive ongoing research in this area.

11.1.7 Case Studies

We briefly summarize some trials with adaptive treatment selection which we take from a recent summary in Bauer et al. (2016). First of all, Zeymer et al. (2001) conducted an international, prospective, randomized, double-blind, placebo-controlled

Phase II trial in patients undergoing thrombolytic therapy or primary angioplasty for acute ST-elevation myocardial infarction applying a two-stage adaptive design. This is the first major clinical trial using adaptive design methodology for many-to-one comparisons with dose-selection at an adaptive interim analysis. A two-stage design with Fisher's combination test and trend tests for the stage-wise p-values was used. Two doses out of four were selected. Finally, no hypotheses could be rejected, unfortunately, and so the trial did not succeed in showing that the drug was superior to placebo at any of the investigated dose levels (for some more details, see Bauer et al. 2016).

Other examples of adaptive treatment selection designs have been implemented since then (Morgan et al. 2014): INHANCE (Donohue et al. 2010) was a multinational, multicenter, double blind, double dummy, two-stage adaptive, parallel group study design with blinded formoterol, and open label tiotropium as active controls in patients with chronic obstructive pulmonary disease (COPD). This trial was one of two pivotal trials to support registration and label claims of indacaterol as novel therapy for the treatment of COPD. The aim of the trial was to provide pivotal confirmation of efficacy, safety, and tolerability of the selected doses of indacaterol, where the dose selection is done at a pre-specified interim analysis. In this case study, a two-stage Phase III adaptive design was an appropriate option, because "dose" was the only major remaining question and a large body of evidence was available at the end of Phase II. Overall, this design led to a reduction of approximately 15 % in terms of development program length, number of patients, and costs as compared to a more traditional design of two sequential trials. INHANCE was included as a pivotal study in submissions to regulatory agencies globally and indacaterol is now approved in all major markets globally for once-daily maintenance bronchodilator treatment of airflow obstruction in adult patients with COPD. The results of the interim analysis of INHANCE have been published in full (Barnes et al. 2010), as have those of the final analysis (Donohue et al. 2010). More details on the methodology employed in this trial can be found in Lawrence and Bretz (2014) and Lawrence et al. (2014).

Hemangeol was developed as the treatment for proliferating infantile hemangioma requiring systemic therapy. It was developed for the use in pediatric population following the guidelines of health regulatory agencies. The Phase III development of Hemangeol was based on a two-stage confirmatory adaptive trial with regimen selection at the end of the first stage, in order to identify the appropriate dose and duration for further study in the second stage. Early stopping for futility and sample size reestimation were also considered at the interim analysis. The aim of this trial was to demonstrate the superiority of the selected dose(s) over placebo and to document its safety profile. Marketing authorization of Hemangeol was granted by both FDA and EMA in 2014. Heritier et al. (2011) provided statistical details of this adaptive design.

Secretory diarrhea in HIV positive patients remains a serious unmet clinical need, even and especially in the age of highly active anti-retroviral therapy. In late 2012 crofelemer was approved by the FDA as a first-in-class anti-diarrheal agent indicated for the symptomatic relief of non-infectious diarrhea in adult HIV patients on antiretroviral therapy. The safety and efficacy of crofelemer were established through ADVENT, a two-stage adaptive clinical trial with dose selection at the end of stage 1 (Chaturvedi et al. 2014).

11.2 Adaptive Enrichment Designs

Adaptive enrichment designs are applicable where studies of unselected patients might be unable to detect a drug effect and it seems necessary to “enrich” the study with potential responders, defined as a subpopulation of the unselected patient population. If this is done in an adaptive and data-driven way (i.e., it is not clear upfront whether to use the selected population and this is decided based on data observed at an interim stage) we might use “adaptive population enrichment designs” (Wang et al. 2007, 2009). We note that Temple (1994) was the first who used the term “enrichment” for this and similar kinds of patient selection.

Adaptive population enrichment designs enable the data-driven selection of one or more pre-specified subpopulations in an interim analysis, and the confirmatory proof of efficacy in the selected subset(s) at the end of the trial. Sample size reassessment and other adaptive design changes can be performed as well. As for the adaptive treatment arm selection designs described in the last section, strong control of the FWER is guaranteed by use of the combination testing principle together with the closed testing procedure.

Enrichment factors may be predictive biomarkers, or they may be biomarkers or clinicopathologic or demographic characteristics associated with a predictive biomarker or with the target of a therapeutic agent. The lower the proportion of truly benefiting patients, the more advantageous it is to consider studying an enriched population. However, instead of limiting the enrollment only to the narrow subpopulation of interest, prospectively specified adaptive designs may also be used to consider the effect of the experimental treatment both in the wider entire patient population under investigation and in various subpopulations.

In this section we briefly describe the general methodology that makes use of the combination testing principle and the closed testing principle. This is equivalent to the methodology described for adaptive treatment arm selection design (see Sect. 11.1.1). It was first proposed in Brannath et al. (2009b) who used Bayesian decision tools for the selection rule (see also, Götte et al. 2015; Graf et al. 2015; Krisam and Kieser 2014). We describe the methodology and designing issues when planning such a design, and give some clinical trial examples where such design can be used. This will be along the lines of the presentation in Wassmer and Dragalin (2015). We note that alternative approaches were proposed in the literature, for example, the approach proposed by Rosenblum and van der Laan (2011). This procedure is restricted to a two-stage design, a predefined selection rule, and the normal case with known variance. As an essential feature, this restriction is not required for the procedure described here (see also, Rosenblum 2015). A systematic review of (also exploratory) procedures for enrichment designs is provided in Ondra et al. (2016).

11.2.1 Test Procedure

Assume there is a full population F with $G - 1$ pre-specified subpopulations of interest denoted as $S_1, S_2, \ldots, S_{G-1}$ such that $S_g \subset F$. Let S_G denote the full population F. We consider a set of G elementary hypotheses

$$H_0^{S_g} : \mu_0^g = \mu_1^g,\ g = 1, \ldots, G,$$

where $H_0^{S_g}$ tests the effect of the experimental treatment μ_1^g versus control μ_0^g in subpopulation S_g.

As before, the closed system of hypotheses consists of all possible intersection hypotheses

$$H_0^{\mathcal{J}} = \bigcap_{g \in \mathcal{J}} H_0^{S_g},\ \mathcal{J} \subseteq \{1, \ldots, G\}.$$

If each hypothesis is tested by a suitable local level α test, a hypothesis $H_0^{\mathcal{J}}$ can be rejected with strong control of the FWER α if all intersection hypotheses $H_0^{\mathcal{I}}$ with $\mathcal{J} \subseteq \mathcal{I}$ are rejected, each at local level α. Specifically, any elementary hypothesis $H_0^{S_g}$ can be rejected if all intersection hypotheses $H_0^{\mathcal{J}}$ with $\mathcal{J} \ni g$ are rejected at local level α.

For a multi-stage design with no interim selection of subpopulations, the complete set of intersection hypotheses is tested for each stage k, yielding p-values $p_{\mathcal{J},k}$ for each intersection hypothesis $H_0^{\mathcal{J}}$. These p-values are combined according to the specified combination test, for example, the inverse normal method or Fisher's combination test. As for the treatment arm selection case, we note that this procedure might have a power disadvantage as compared to the procedure where the p-values are obtained from the pooled data. However, we have the advantage that data-driven adaptations including subgroups selection are possible, thereby improving power.

In this case, the global test decision at stage k is determined based on testing the global hypothesis

$$H_0 = \bigcap_{g=1}^{G} H_0^{S_g}$$

with the selected combination test.

If a subpopulation selection has been performed at an interim stage, the same problem as described in Sect. 11.1.1 arises: not all tests of the intersection hypotheses $H_0^{\mathcal{J}}$ are available for the subsequent stages. As a solution, we define tests for intersection tests involving excluded subpopulations as tests for the non-excluded subpopulations. That is, if a subset $\mathcal{S}_k$ of analysis sets is selected for stage k, the p-values for $H_0^{\mathcal{J}}$ are replaced by the p-values for $H_0^{\mathcal{J}\cap\mathcal{S}_k}$. The closed testing procedure is then performed as described above, combining the p-values from the earlier stage(s) with their counterparts, if they exist, or with the replaced p-values.

Formally, given a combination function C, at the second stage the hypotheses belonging to a selected subpopulation s are rejected if

$$\max_{\mathcal{J}\ni s} C(p_{\mathcal{J},1}^{\text{adj}}, p_{\mathcal{J}\setminus\mathcal{E},2}^{\text{adj}}) \leq c \ , \tag{11.10}$$

where $\mathcal{E} \subset \{1,\ldots,G\}$ denotes the index set of all excluded H_0^g, and c denotes the critical value for the combination test. If one subpopulation, s, is selected, condition (11.10) reduces to

$$\max_{\mathcal{J}\ni s} C(p_{\mathcal{J},1}^{\text{adj}}, p_{s,2}) \leq c \ .$$

For $G = 3$ and one selected subgroup, the procedure is illustrated in Fig. 11.3, the curved arrows indicating which combination tests have to be carried out in order to show significance of the selected subgroup S_2.

Designing a population enrichment trial is a complex task since the trial consists of many elements that influence the operating characteristics of the design. In the following we briefly describe possible options for a population enrichment design. These can be used to assess the benefits of such a design and help to decide whether it is appropriate for the objectives of a specific trial.

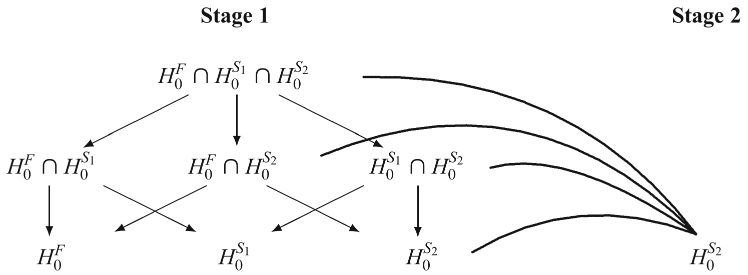

Fig. 11.3 Closed system of hypotheses for $G = 3$ if subpopulation S_2 referring to hypothesis $H_0^{S_2}$ is selected for the second stage. The *arrows* indicate logical implications for hypotheses, the *solid curves* indicate combination tests to be performed to show significance for $H_0^{S_2}$ (from Wassmer and Dragalin 2015)

11.2.2 Intersection Tests

As described in the last section, for the closed testing procedure, several choices of the intersection tests for testing the global and intersection hypotheses are available. To describe these, let as before p_g denote the p-value for testing H_0^g, $g = 1, \ldots, G$ and let $p_{(1)} \leq \ldots \leq p_{(G)}$ denote the ordered p-values of the G comparisons. Finally, let $\mathcal{J} \subset \mathcal{G}$, and $|\mathcal{J}|$ be defined as the number of all indices $g \in \mathcal{J}$.

Bivariate *t* Test for One Subgroup of Interest

For enrichment designs, the choice of an intersection test that is based on the maximum statistic similar to Dunnett's test is possible for $G = 2$, i.e., for one subpopulation $S \subset F$. This was proposed by Spiessens and Debois (2010) and Friede et al. (2012) for the known variance case. More general, for normal responses with a common unknown variance the adjusted p-value, p^{adj}, for testing the global null hypothesis H_0 is given by

$$p^{\text{adj}} = 1 - F_{\boldsymbol{\Sigma},df}\big(\max\{Z^F, Z^S\}\big) ,$$

where $F_{\boldsymbol{\Sigma},df}(\cdot)$ is the value of the cdf of the bivariate t distribution with correlation

$$\sigma_{12} = \sqrt{\frac{n_0^S + n_1^S}{n_0^F + n_1^F}}$$

and $df = n_0^F + n_1^F - 4$ degrees of freedom when the two arguments are equal,

$$Z_g = \frac{\bar{x}_1^g - \bar{x}_0^g}{\hat{\sigma}\sqrt{1/n_0^g + 1/n_1^g}}$$

is the directional test statistic for analysis set g, and $\hat{\sigma}^2$ is the residual variance estimate corresponding to a two-factorial ANOVA model.

This test provides exact Type I error rate control under the specified assumption of normal responses with an unknown common variance. For $G = 2$, we note that it is also possible to derive a test procedure that is based on the CRP principle. This conditional procedure was proposed in Friede et al. (2012) who reported in the correction that in realistic situation the difference in power as compared to the combination test approach is only small and there is actually no power advantage of the CRP methodology.

For a general number of subpopulation, the description of the Bonferroni, Šidák, Simes, and the a priori hierarchical is the same as for the treatment arm selection case:

Bonferroni Test for Many Subgroups

Using the Bonferroni test, the adjusted p-value for testing a hypothesis $H_0^{\mathcal{J}}$, $\mathcal{J} \subseteq \mathcal{G}$, is given by

$$p_{\mathcal{J}}^{\text{adj}} = \min\{|\mathcal{J}| \min_{g \in \mathcal{J}}\{p_g\}, 1\} .$$

Šidák Method for Many Subgroups

With the Šidák test, the adjusted p-value for testing a hypothesis $H_0^{\mathcal{J}}$, $\mathcal{J} \subseteq \mathcal{G}$, is given by

$$p_{\mathcal{J}}^{\text{adj}} = 1 - (1 - \min_{g \in \mathcal{J}}\{p_g\})^{|\mathcal{J}|} .$$

Simes Method for Many Subgroups

With the Simes intersection test, the adjusted p-value for testing a hypothesis $H_0^{\mathcal{J}}$, $\mathcal{J} \subseteq \mathcal{G}$, is given by

$$p_{\mathcal{J}}^{\text{adj}} = \min_{g \in \mathcal{J}}\{\frac{|\mathcal{J}|}{g} p_{(g_{\mathcal{J}})}\} ,$$

where $p_{(1_{\mathcal{J}})} \leq \ldots \leq p_{(|\mathcal{J}|_{\mathcal{J}})}$ denote the ordered p-values from the subset $\mathcal{J} \subset \mathcal{G}$.

We note that both Simes and Šidák derived p-values for intersection hypotheses yield valid level α test procedures since the elements of $\boldsymbol{\Sigma}$ are always positive. Therefore, Type I error rate control can be guaranteed in general (see Sect. 11.1.2).

A Priori Hierarchical Test

If an ordering of the hypotheses can be assumed (for example, the hypothesis relating to the full population is considered first, then S_{G-1}, S_{G-2}, etc.), the adjusted p-value for testing a hypothesis $H_0^{\mathcal{J}}$, $\mathcal{J} \subseteq \mathcal{G}$, is given by

$$p_{\mathcal{J}}^{adj} = p_{\max\{g \in \mathcal{J}\}},$$

where $\max\{g \in J\}$ corresponds to the hypothesis in $\mathcal{J}$ with the highest importance. Note that, of course, hierarchical testing in enrichment designs is often questionable and may be applied only in very specific situations (for example, showing a subgroup effect is only of interest after showing an effect in the full population).

11.2.3 Effect Specification

In order to assess the statistical performance of an enrichment design (usually by simulation), the operating characteristics depend on the effect sizes and prevalences of the considered subpopulations S_g. If one subpopulation S is considered, this is to specify the prevalence of S and the assumed effect sizes in S and $F \backslash S$. The effect in F is a weighted average of the effect sizes in the disjunct subgroups S and $F \backslash S$, respectively. There is typically a large number of possible configurations because each effect size in S is combined with the effect sizes in $F \backslash S$.

For $G = 3$, assume there are two dichotomous indicators I_1 and I_2 that classify the patients in the full population as I_1^- or I_1^+ and I_2^- or I_2^+. I_1 may be a baseline characteristic such as gender, race, performance status, or disease stage. I_2 may be a genomic biomarker (positive versus negative) or a genomic signature (good versus poor). We may be interested in showing the treatment effect of a test drug versus placebo in the full population and in I_1^+ and I_2^+. In this case, we have three analysis sets: $S_1 = I_1^+$, $S_2 = I_2^+$, and $S_3 = F$.

Often, $S_1 \cap S_2 \neq \emptyset$ and therefore the operating characteristics of the test procedure depend on the effect sizes in $S_1 \backslash S_2$, $S_2 \backslash S_1$, $S_1 \cap S_2$ and $F \backslash (S_1 \cup S_2)$ and their assumed prevalences. $S_1 \cap S_2 = \emptyset$ can be assumed if an investigation of effect sizes is considered in different patients populations (for example, countries) besides the effect in a full population. This reduces the number of necessary specifications for the prevalences and the effect sizes. If we consider the nested case where $S_1 \subset S_2 \subset F$, such as in Wang et al. (2009), $S_1 \cap S_2 = S_1$ and $S_1 \backslash S_2 = \emptyset$. Note that if we are also interested in showing an effect in $S_1 \cap S_2$ (i.e., interaction effects), then we specify $G = 4$ subpopulations $S_1 = I_1^+$, $S_2 = I_2^+$, $S_3 = I_1^+ \cap I_2^+$, and $S_4 = F$. These four different situations are illustrated in Fig. 11.4.

11.2.4 Overall p-Values and Confidence Intervals

As in Sect. 11.1.3, overall (repeated) p-values are defined as smallest significance level for which the test results yield rejection of the considered (single) hypothesis $H_0^{S_g}$ at stage k. They can generally be found by a numerical search and can be calculated at any stage of the trial. They account for the step-down nature of the closed testing principle and are completely consistent with the test decision.

In analogy to Posch et al. (2005), overall confidence intervals can also be found in an equivalent way as described in Sect. 11.1.3. The idea is to consider the shifted

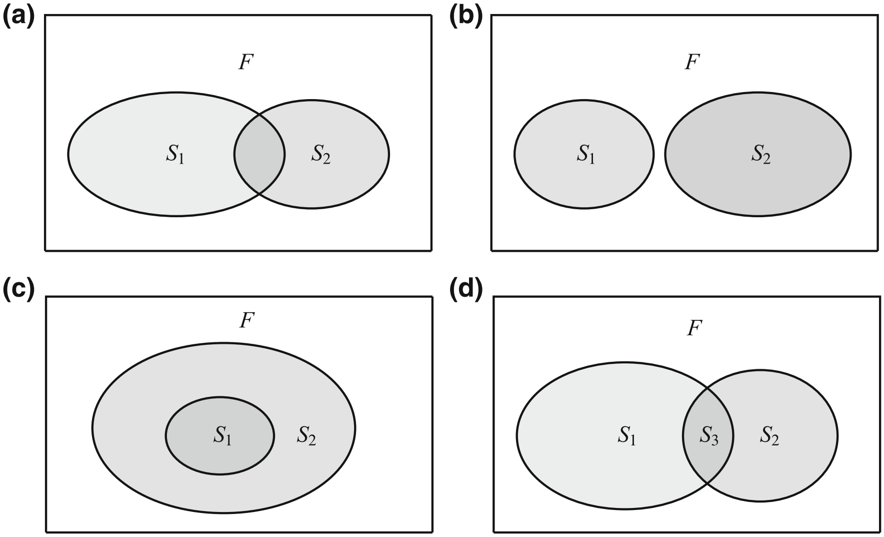

Fig. 11.4 Four different configurations of subpopulation, see text. (**a**) Two subpopulations of interest. (**b**) Two non-overlapping subpopulations. (**c**) Two nested subpopulations. (**d**) Three subpopulations of interest

hypothesis

$$H_0^{S_g}(\delta_g) : \mu_1^g - \mu_0^g = \delta_g,\ g = 1, \ldots, G,$$

with corresponding p-values $p_{g,k}(\delta_g)$ at stage k and deriving the adjusted shifted p-values with the use of the selected intersection test. Note that we have the same difficulties with hierarchical testing as in adaptive designs with treatment selection (see Sect. 11.1.3).

11.2.5 Other Endpoints

In the binary case, the derivation of closed adaptive tests for the set of G elementary null hypotheses

$$H_0^{S_g} : \pi_0^g = \pi_1^g,\ g = 1, \ldots, G,$$

is formally straightforward when using stage-wise adjusted p-values like the Bonferroni, Šidák, or Simes test which do not account for the correlation structure of the test statistics. However, one may need to modify the standard error estimates in (5.13), for instance, in the full population, because of overdispersion effects from the mixture of (potentially) different binomial distributions in the different subpopulations. If we wish to account for the correlation between the G test statistics by using the asymptotic multivariate normal distribution of the effect estimates for the adjusted stage-wise p-values additional problems may arise in the estimation of the correlation matrix which depends on the different null proportions in the different subpopulations. As far as we know, no research has been done on these issues yet.

In the survival case, when testing

$$H_0^{S_g} : \omega^g = 1,\ g = 1, \ldots, G,$$

one needs to use the stratified log-rank test with the disjoint subgroups as strata in order to achieve asymptotic Type I error rate control (see Brannath et al. 2009b). Utilization of the approximate multivariate normal distribution for improving the stage-wise p-values is more straightforward here but, as far as we know, has not been published yet.

For the survival case it was described what problems can arise if the population selection is based on a short-term rather than on the primary endpoint. A solution specifically referring to the population enrichment case was given in Jenkins et al. (2011).

11.2.6 A Clinical Trial Example

Different therapeutic areas for the application of adaptive enrichment designs have been published recently (Mehta et al. 2009; Mehta and Gao 2011; Mehta et al. 2014; Tournoux-Facon et al. 2011a,b). Wassmer and Dragalin (2015) presented typical case studies for designing enrichment designs for different kind of endpoints and goals of the trial. We present one example from this article. It considers an enrichment design with one subpopulation of interest for comparing rates. Note that it is hypothetical because it has not been conducted in the described way. Nevertheless, it illustrates how the described confirmatory adaptive enrichment designs might be suitable alternatives to the traditional trial designs.

Investigation of **S**erial Studies to **P**redict **Y**our **T**herapeutic **R**esponse with **I**maging **A**nd mo**L**ecular Analysis (I-SPY 2 TRIAL) is an ongoing clinical trial in patients with high-risk primary breast cancer. It involves a randomized phase II screening process in which a series of experimental drugs are evaluated in combination with standard neoadjuvant chemotherapy which is given prior to surgery. The primary endpoint is pathologic complete response (pCR) at the time of surgery (for details, see Barker et al. 2009).

The screening process includes a magnetic resonance imaging to establish tumor size at baseline and a biopsy to identify the tumor's hormone receptor status (HR) and the HER2/neu status (HER2). Triple negative breast cancer (TNBC) refers to breast cancer that does not express the genes for the estrogen receptor, the progesterone receptor or HER2.

Assume that one of the experimental drugs has been identified from I-SPY 2 TRIAL not only with the biomarker signature of TNBC but also with some promising effect in the HER2 negative (HER2$-$) biomarker signature. The sponsor may consider a confirmatory Phase III trial in TNBC patients only. The prevalence of TNBC, however, is only 34 %, while the prevalence of HER2$-$ is 63 %. Therefore, an alternative option is to run a confirmatory trial with a two-stage enrichment design starting with the HER2$-$ patients as the full population, but with the pre-planned option of selecting the TNBC patients after the first stage if the observed effect is not promising in the HER2$-$ patients with positive hormone receptor status HR+.

If a pCR rate in the control arm of 0.3 and a treatment effect of 0.2 is assumed (the treatment effect is measured as the difference in pCR rates between the new drug and control), the required total sample size for a conventional two-arm test with power 90 % and one-sided significance level 0.0125 (i.e., applying Bonferroni correction) is 294. This can be found with formula (5.12) from Sect. 5.2. It will serve as a first guess for the actually needed sample size and we illustrate the enrichment design for this study assuming that a total sample size of 300 subjects will be enrolled in the trial.

The interim analysis is planned after 150 subjects and no early stopping is intended. A subpopulation selection using the ϵ-selection rule with $\epsilon = 0.1$ will be considered. That is, the decision at the interim analysis will be to either select the TNBC subpopulation or going on with the full population of HER2$-$ patients. If the observed treatment effect difference exceeds 0.1 in favor of the TNBC population, the TNBC subpopulation will be selected, otherwise no selection will be considered and the test for the full population only will be conducted if the observed treatment effect difference exceeds 0.1 in favor of the F population; otherwise the test for both populations will be conducted. The inverse normal combination testing strategy together with the Bonferroni intersection test will be used.

In I-SPY 1 TRIAL, a prevalence of TNBC patients in the HER2$-$ population of about 54 % and a control pCR rate in TNBC patients of 0.34 was observed. The pCR rate in the HER2$-$ patients with HR+ hormone receptor is 0.23. The operating characteristics of the enrichment design are investigated for treatment effect differences ranging from 0 to 0.3 by an increment of 0.05 in the TNBC subpopulation and ranging from 0 to 0.2 by an increment of 0.10 in the HER2$-$ patients with HR+ hormone receptor. This yields 21 different scenarios for the effect sizes resulting in different treatment effects for the full population F. The results of 10,000 simulations per scenario are reported in Table 11.3, the software ADDPLAN PE (see Appendix) was used for the simulation.

The power of the design (the probability to reject at least one null hypothesis) is greater than 90 % for scenarios 17–21, mainly corresponding to treatment effects 0.25 and 0.3 for S. Hence, in these cases a total sample size of 300 patients reaches the desired power, and the rough estimate provided through the use of the Bonferroni correction provides a good estimate for the necessary sample size. Note that the term

Table 11.3 Simulation results for population enrichment design with one subgroup. S and $\bar{S}$ refer to the TNBC population and HER2− patients with HR+ hormone receptor population, respectively. Results are based on a TNBC prevalence of 54 % within HER2− patients (from Wassmer and Dragalin 2015)

Scenario	Effect S	Effect $\bar{S}$	Effect F	Power	P(Reject F)	P(Reject S)	P(Select F)	P(Select S)	P(Select 1 set)
1	0	0	0	0.015	0.009	0.011	0.933	0.929	0.139
2	0	0.1	0.046	0.065	0.062	0.019	0.982	0.785	0.233
3	0	0.2	0.092	0.249	0.249	0.022	0.997	0.549	0.454
4	0.05	0	0.027	0.070	0.034	0.063	0.869	0.962	0.169
5	0.05	0.1	0.073	0.158	0.141	0.076	0.960	0.868	0.172
6	0.05	0.2	0.119	0.412	0.407	0.089	0.990	0.673	0.337
7	0.1	0	0.054	0.206	0.083	0.195	0.788	0.981	0.231
8	0.1	0.1	0.1	0.320	0.265	0.220	0.919	0.928	0.154
9	0.1	0.2	0.146	0.596	0.580	0.236	0.978	0.778	0.244
10	0.15	0	0.081	0.434	0.161	0.425	0.675	0.992	0.333
11	0.15	0.1	0.127	0.537	0.418	0.437	0.865	0.958	0.176
12	0.15	0.2	0.173	0.768	0.725	0.470	0.955	0.863	0.182
13	0.2	0	0.108	0.669	0.228	0.660	0.551	0.998	0.451
14	0.2	0.1	0.154	0.755	0.538	0.689	0.780	0.982	0.238
15	0.2	0.2	0.2	0.896	0.817	0.701	0.920	0.919	0.162
16	0.25	0	0.135	0.860	0.264	0.857	0.421	0.999	0.580
17	0.25	0.1	0.181	0.901	0.572	0.873	0.675	0.991	0.335
18	0.25	0.2	0.227	0.960	0.821	0.864	0.861	0.957	0.182
19	0.3	0	0.162	0.958	0.246	0.957	0.301	1.000	0.699
20	0.3	0.1	0.208	0.972	0.518	0.962	0.548	0.996	0.456
21	0.3	0.2	0.254	0.991	0.772	0.952	0.781	0.978	0.241

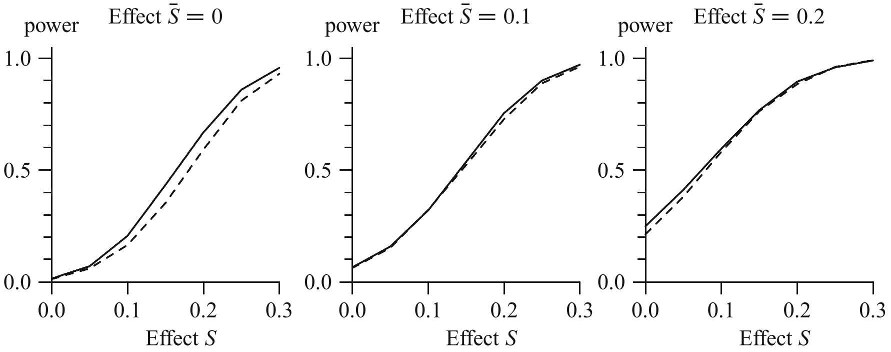

Fig. 11.5 Power of enrichment design as compared to the power if no population selection takes place. The *solid line* refers to the enrichment design, the *dashed line* to the classical design with no selection of study population (from Wassmer and Dragalin 2015)

"power" is used here also for the cases where the null hypothesis is true (Scenario 1–3). This, however, illustrates *strong* control of FWER (see column "P(Reject S)").

The table also shows that the power to reject in the full population is (except for effect size ≥ 0.2 and < 0.3 in both subsets, i.e., scenarios 15 and 18) smaller than 80 %, for largest effect sizes the power even decreases a bit. The latter is due to the fact that in this case the probability to deselect F and to select S increases. For most scenarios the probability to reduce the confirmatory proof to one hypothesis, H_0^S or H_0^F, is quite small, see column "P(Select 1 set)".

The case for enrichment, i.e., the selection of S at the interim stage, varies between 1 and 70 % over the scenarios and can be derived from P(Select F): P(Enrichment) = 1 − P(Select F). The question arises if this might reduce power (defined as above) due to wrongly selecting a population. The answer is no, as illustrated in Fig. 11.5. Here it is shown that for all effect sizes in $\bar{S}$ there is no decrease, for effect size 0 there is even a clear increase in power showing the advantage of an adaptive enrichment design as compared to the non-adaptive case.

11.3 Other Types of Adaptations

When using the combination testing approach or the CRP principle, the class of possible data-driven adaptations is rich. We have shown that not only a sample size recalculation but also a selection of treatment arms or population subsets at interim stages is possible. Generally, although the latter two applications might be considered as a specific type of sample size recalculation (the sample size of the deselected treatment arms or subsets is set equal to 0), it involves a change of the hypotheses that are under consideration at the different stages. Hommel (2001) showed that generally, using an adaptive test, the hypotheses within a predefined class of hypotheses can be changed during the course of the trial. It is even possible to include a completely new hypothesis but this is mainly of theoretical interest because this hypothesis should be tested with the data of the subsequent stages only such that a completely new trial is always at least as good as the adaptive procedure. The application of an adaptive change of hypotheses was illustrated by several examples, including an adaptive modification of the hierarchy of hypotheses (Hommel and Kropf 2001; Kropf et al. 2000; Kieser 2005).

Another attractive option is to select the way of testing a hypothesis, i.e., the choice of the test statistic or the structure of the underlying test procedure in a multiple testing situation. This application was illustrated for trend tests (Lang et al. 2000) and non-parametric procedures (Neuhäuser 2001). Kieser et al. (2002) proposed a bootstrap procedure for the adaptive selection of the test statistic. Lawrence (2002) considered the change of the test statistic in survival trials.

The adaptive choice of test statistics might also be of interest in clinical trials with multiple endpoints. These trials address several outcome variables within a single confirmatory experiment and multiple tests are part of the confirmatory statistical analysis. A range of possible ways to construct statistical test procedures were proposed (for a review, see Wassmer et al. 1999) and it might be appropriate to reselect a specific test based on interim outcomes, including the selection of endpoints. It is often straightforward to define an adaptive closed test to get strict error control for all considered endpoints. We illustrate this application in the example below.

Finally, due to increase sophistication of multiple testing strategies the communication of the related multiple testing procedures and its interpretation to clinicians and sometimes even to statistical colleagues gets complex. Hence graphical methods to visualize the logical structure have been suggested to facilitate planning, execution, and interpretation of such complex multiple tests for conventional fixed size sample designs (Bretz et al. 2009b, 2011; Burman et al. 2009). Recently, adaptive graph based methods for multiple comparisons based on combination tests (Sugitani et al. 2013, 2014) or the CRP principle (Klinglmüller et al. 2014) have been proposed.

11.3.1 A Case Study with Adaptive Multiple Endpoint Selection

We describe a trial with adaptive endpoint selection that involved an adaptive change of hypotheses where we take the report from Bauer et al. (2016). This placebo-controlled multicenter trial was performed to test three co-primary endpoints—two superiority and one non-inferiority hypotheses have been involved. It was investigated whether clonidine as a co-medication with fentanyl and midazolam is superior to fentanyl and midazolam alone in ventilated newborns, and infants up to 2 years of age, as measured by the endpoints: total consumption of fentanyl (superiority hypothesis 1) and midazolam (superiority hypothesis 2). Additionally, non-inferiority to placebo with respect to the need for the rescue thiopentone use has to be shown (hypothesis 3). It was a study of the PAED-Net which is a corporation of pediatric modules in six German university locations in a small and vulnerable population. For details of the study conduct see Hünseler et al. (2014) (Trial Registration Number ISRCTN77772144).

Regulatory authorities usually expect statistical significance in all three co-primary endpoints simultaneously, which would mean that no further multiple testing correction is needed. But in the setting of pediatric populations the investigator was convinced that under double blind conditions it would be worth to achieve significance in at least one of the two superiority hypotheses. The non-inferiority margin for thiopentone use was set to 20 %. The corresponding global null hypothesis was tested using the ordinary least squares (OLS) test due to O'Brien (1984), and a closed testing was planned in order to show significant differences in specific endpoints (Lehmacher et al. 1991).

The study was planned in three stages with critical values according to O'Brien and Fleming adjusted significance levels 0.0003, 0.0071, and 0.0225, respectively. The results of the three stages were combined using the inverse normal method together with the closed testing principle as described in Sect. 10.3. Under reasonable assumption of the effect sizes and taking into account the correlation of the endpoints, by simulation a sample size of 210 patients was estimated to achieve 80 % overall power at one-sided significance level $\alpha = 2.5\,\%$. Overall power was defined for detecting at least one significant difference. Considering the power to reject all hypotheses required a lot more patients and was considered inappropriate.

Three types of adaptations were pre-specified in the study protocol:

1. a sample size reassessment based on the results observed at interim stages,
2. the possibility to redefine the test statistic if it turned out that the OLS test statistic was clearly inferior to a better overall test,
3. dropping a superiority endpoint if the effect size in this endpoint was too low.

The last option seemed to be useful since it was not clear at the beginning if both fentanyl and midazolam consumption could be reduced with clonidine.

The first interim analysis yielded very promising results: The overall p-value for the OLS test was 0.0009, thus already near showing significance. The midazolam result, however, was already weak ($p = 0.0472$) as compared to the others ($p = 0.0051$ for fentanyl and $p = 0.0012$ for thiopentone). This trend was dramatically confirmed at the second interim analysis: a negative effect in midazolam was observed for the second stage data, yielding a p-value of 0.678. Nevertheless, the OLS test for the global multivariate hypothesis yielded a p-value of 0.0017. This was statistically significant, though a study continuation was recommended because superiority with regard to fentanyl was not very clear anymore. Particularly, it was not significant within the closed test procedure, only non-inferiority with regard to thiopentone was significant within the closed test procedure. The study recommendation was also to drop midazolam consumption as a primary endpoint from the further analysis since it might jeopardize an overall positive result of the study.

Although a positive result was likely for the reduced clinical question, at the final analysis even fentanyl could not be shown to be significant. The main study results were published in Hünseler et al. (2014).

The example shows the importance of keeping adaptive interim decisions secretly. If the doctors had been aware of the fact that midazolam was dropped from further confirmatory analyses this might have clearly influenced treatment and medication of the patients. In order to exclude this possibility, only the iDMC and the Independent Statistical Center who gave the recommendation were aware of the study results. The head of the study was informed that there was a recommendation, the decision on it was left to one representative of the sponsor (Boehringer Ingelheim, Germany) which needed to be involved.

The study did not show the desired effect though especially the second interim analysis illustrates the potential advantage of an adaptive way of analyzing data. There were no strict stopping criteria and the continuation of the trial produced a disappointing result but obviously reflects reality. We nevertheless think that this study serves as an interesting example for an early attempt for an adaptation which goes beyond sample size reassessment and treatment arm selection in a vulnerable, small population (EMA 2006). The study was planned 2002 and finalized in 2008. Publication of the non-convincing study results was a problematic issue. Another issue with publishing complicated adaptive designs in medical journals is to have space to communicate the statistical methodology (Bauer and Einfalt 2006). Indeed, several details of the statistical study design were not provided in Hünseler et al. (2014).

11.4 Regulatory and Logistical Issues

The repeated significance testing approach was proposed in the late 1960s. This approach introduced an increased complexity as compared to a fixed design with a single primary efficacy endpoint, a fixed sample size, and no interim looks. Particularly, looking repeatedly for efficacy involves the potential for operational bias, mainly due to unblinding the study results at interim. The increased flexibility of adaptive designs deteriorates the problem and results in several regulatory and logistical issues. It must be stated, however, that we already encounter many of the issues in adaptive designs when we are applying the classical group sequential methodology and so many problems are “inherited.”

We already noted in the Preface of this monograph that the adaptive design methodology found its way into a Reflection Paper entitled “Methodological Issues in Confirmatory Clinical Trials Planned With an Adaptive Design” from the European Medicines Agency (EMA 2007), a draft guidance on “Adaptive Clinical Trials for Drug and Biologics” from the US Food and Drug Administration (FDA 2010), and a draft guidance on “Adaptive Designs for Medical Device Clinical Studies,” also from the US Food and Drug Administration (FDA 2015). We do not review these documents but want to refer to a summary provided by Wang

(2014) who reviews the EMA guideline as well as the FDA "Adaptive Clinical Trials for Drug and Biologics" draft guidance (see also, Brannath et al. 2010; Hung et al. 2011). As an overall summary, when applied carefully, the adaptive design methodology serves as a useful tool for performing confirmatory trials with registration potential. However, it is clearly not a remedy for the inability of making clear design specification nor is it an excuse for poor study planning.

The FDA guidance defines an adaptive study as one that "includes a prospectively planned opportunity for modification of one or more specified aspects of the study design and hypotheses based on analysis of data (usually interim data) from subjects in the study." Hence, the type of adaptations needs to be prospectively planned, where "the term prospective here means that the adaptation was planned (and details specified) before data were examined in an unblinded manner by any personnel involved in planning the revision." (FDA 2010). The possible ways of doing it should be thoroughly evaluated through extensive simulations, and all the details extensively discussed with the regulator. From this, a guidance to perform the adaptations is formulated and provided in a document that is available only for the persons who actually are assessing the study results that eventually lead to an adaptation. Usually, this is an iDMC charter of which the circulation is strictly limited. Furthermore, it is absolutely required to have at hand prospectively written standard operating procedures and working processes for implementing adaptive designs. Furthermore, pharmaceutical companies are encouraged to engage contract research organizations (CROs) that are experienced in running adaptively designed trials.

As an important point, from a regulator's point of view, it is therefore required to *specify all adaptive design elements at the planning stage of the study.* This is in some sense controversial to the possible options provided by the adaptive design methodology. Specifically, the methodology theoretically allows for such decisions at an interim stage, i.e., it allows that the type of an adaptation and the way of how to conduct it need not to be pre-specified. For example, introducing—at interim—a selection of endpoints in addition to a pre-specified sample size recalculation is theoretically allowed; a sample size recalculation in addition to a treatment arm selection design is possible too; or even *introducing* a formerly unspecified sample size reassessment in a fixed sample size design is a possible option (for example, using the CRP principle). From a regulatory point of view, all this should be avoided. Otherwise, the integrity of the clinical trial cannot be maintained and the adaptive design can merely be understood in a more exploratory way. It is important to say this, because at the very beginning of the methodological investigation in the late 1990s from a part of users the adaptive design methodology was misunderstood in the sense that "all is allowed now." This is clearly not the case.

iDMCs were instituted to maintain the integrity of a trial, to ensure patient safety, and to perform the interim analyses. DeMets et al. (2006), Ellenberg et al. (2003), and Herson (2009) are excellent books that describe the role and definition of this board of people thereby providing many study examples. Recently, Gallo et al. (2014) gave a summary of the issues of iDMCs in adaptive trials. Most importantly, in adaptive designs the question arises of who is doing the adaptation and how.

Besides the fact that this problem already exists in conventional group sequential trials (for example, an early stop for efficacy or futility can be regarded as a kind of sample size reassessment) it clarifies the fact that at least the statistician in the iDMC should possess specific knowledge on the adaptive design methodology.

The iDMC's role is not to design or redesign the trial, its role is to apply specific adaptation rules that are prospectively planned and evaluated, i.e., it is implementing an adaptation. It is important to recognize, however, that the adaptation rules will usually not serve as strict rules but more as a guideline and it is even the case that the charter allows the iDMC to deviate from the guidelines. This is because it is generally impossible to take into account all possible information that is achieved at an interim analysis and formulate the rules accordingly. For example, it turns out at an interim stage that only a combined evaluation of safety *and* efficacy can be adequately used for performing an adaptation. An increase in sample size can also be recommended to obtain more information on important secondary efficacy or safety endpoints. As another example, we often do not know exactly the effect size and variability at interim because usually some patients are already observed but not yet in the data base. In this case, if the efficacy boundaries are crossed but only to a small amount, the iDMC might recommend not stopping the study and awaiting a confirmation of a significant result at the next interim (instead of running the risk that the final data base does not achieve significance anymore).

It is also essential that an iDMC will never make a decision to redesign a trial, it is only giving a recommendation. The decision is made by the sponsor. To maintain the integrity of the trial, a (small) number of persons on sponsor's site is required who is allowed and responsible for making the decision and is not involved in any other trial activities. If an adaptation is recommended by the iDMC these sponsor representatives will be informed about the necessary details that resulted in the recommendation, and make the decision. In a sense, this problem already exists in conventional group sequential designs because also here knowledge of unblinded interim data is provided to sponsor personal in case of major recommendations, such as terminating the trial due to efficacy of futility.

As a last point, we consider the problem of "reverse engineering" of the effect size that, for example, led to a change in the sample size. Based on an increased sample size at interim, say, one is able to back calculate the effect size. To reply on this, we note that this problem, at least partly, also exists if no adaptations are planned because, for example, the continuation of the trial also provides some information about the effect size. Second, if adaptive designs are properly performed, a sample size recalculation will usually be based on multiple information, for example, a combined evaluation of the effect size and a nuisance parameter. Therefore, it is simply not possible to back calculate the effect size. At least, this demagnifies the problem. Lastly and most importantly, it illustrates the importance of adequate firewalls that are put in place to guarantee that personal involved in the study conduct do not have access to any kind of unblinded study information.

Appendix: Software for Adaptive Designs

The availability of software is a necessary condition for the applicability and acceptance of a statistical methodology. Many of the procedures proposed for adaptive designs additionally require high levels of computational performance such that software should be able to perform complex computations in a relatively short time. This kind of software is available today, and we briefly review the available software in this chapter which is a bit more general review as the one provided in Bauer et al. (2016). Up to now, the reviews of software packages concentrated on packages specifically designed for group sequential methods (Emerson 1996; Wassmer 2006; Zhu et al. 2011), the reason simply being that software for adaptive designs was not available at that time. A review of software for adaptive designs is provided in Tymofyeyev (2014).

One essential core of many if not all packages available for group sequential design is the numerical computation of the multivariate normal integral as described in Chap. 1. For group sequential designs it turns out that due to the independent increment structure of the underlying stochastic process the multivariate integral can be computed through the successive computation of univariate integrals. This is a consequence of the well-known *recursive integration formula* described in Armitage et al. (1969), and makes the computation feasible. It is interesting to recognize that these authors were already able to provide accurate results for the problem for up to 100 dimensions, i.e., stages of the trial. Due to the enormous growth in the computational capacity many alternative algorithms are available today that make the computation feasible. For an overview, see Genz and Bretz (2009).

A wide range of computations necessary in the context of the assessment of group sequential designs is possible with the use of software programs freely available on the homepage of Christopher Jennison: www.bath.ac.uk/~mascj. He provides the Fortran code for all the tabulated results of the Jennison and Turnbull monograph on group sequential designs in clinical trials (Jennison and Turnbull 2000). This might

G. Wassmer, W. Brannath, *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*, Springer Series in Pharmaceutical Statistics,
DOI 10.1007/978-3-319-32562-0

serve as a very valuable tool to find the source code for algorithms to be used in group sequential designs.

Fortran programs for the computation of the use function approach are available from the University of Wisconsin School of Medicine and Public Health site www.biostat.wisc.edu/content/lan-demets-method-statistical-programs-clinical-trials Programs for Computing Group Sequential Boundaries Using the Lan-DeMets Method, Version 2.1. It comes with a Microsoft Windows graphical user interface and hence additionally provides a convenient way to perform the calculation. The last update is from 11/2003. So this tool was not further developed, and it is restricted to the use function approach. However, an R tool is available now (see below). R is free, and it also compiles and runs on a wide variety of UNIX platforms, Windows, and MacOS. This might be advantageous, and reason for its widespread use. We checked CRAN (Comprehensive R Archive Network) cran.rstudio.com on January 20, 2015, and list the available packages which are available, together with a short description and its potential use in adaptive designs. We hope to provide a more or less complete list though it is emphasized that this is a dynamic development and we expect a lot of more packages in the near future. We also note that we concentrate on tools for confirmatory adaptive designs and *not* on tools especially developed for early phase dose-finding trials.

- adaptTest: Adaptive two-stage tests (Vandemeulebroecke 2009). The functions defined in this program serve for implementing adaptive two-stage adaptive tests that are based on the combination testing principle.
- AGSDest: Estimation in adaptive group sequential trials (Hack et al. 2013). This module enables the calculation of confidence intervals in adaptive group sequential trials.
- asd: Simulations for adaptive seamless designs (Parsons 2013). This package runs simulations for adaptive seamless designs with and without early outcomes for treatment selection and population enrichment type designs.
- gMCP: Graph Based Multiple Comparison Procedures (Rohmeyer and Klinglmüller 2014). This package provides functions and a graphical user interface for adaptive (Klinglmüller et al. 2014) and non-adaptive (Bretz et al. 2009b) graph-based multiple comparison procedures.
- GroupSeq: A GUI-based program to compute probabilities regarding group sequential designs (Pahl 2014). This program can be used for assessing the test characteristics of group sequential design and providing the boundaries for a group sequential approach or an inverse normal combination test approach.
- gsDesign: Group Sequential Design (Anderson 2014). gsDesign is a comprehensive package that derives group sequential designs and describes their properties. A graphical user interface gsDesignExplorer is available as well. The resulting boundaries can be used for adaptive settings.
- interAdapt (Fisher et al. 2014). This is an interactive tool for designing and evaluating certain types of adaptive enrichment designs.
- ldbounds Lan-DeMets method for group sequential boundaries (Casper and Perez 2014) is based on the Fortran from the University of Wisconsin and can also be used to provide the test characteristics of the use function approach.

- PwrGSD: Power in a Group Sequential Design (Izmirlian 2014). This program evaluates analysis plans for sequentially monitored trials on a survival endpoint. It can also be used to perform power calculations in a group sequential setting.
- seqmon: Sequential Monitoring of Clinical Trials (Schoenfeld 2012). This program elementarily computes the probability of crossing sequential boundaries in a clinical trial and uses a method described by the author (Schoenfeld 2001).

There is also an R package called RCTDesign: Methods and Software for Clinical Trials. This package builds on the formerly available S-Plus module S+SeqTrial. RCTdesign is currently not available at CRAN but is freely available to users through a joint agreement between Tibco, Inc. (the owners of the S-Plus software system and the S-Plus code in the module S+SeqTrial) and Scott S. Emerson (the developer of the C code that serves as the engine for S+SeqTrial). RCTdesign makes the computation and evaluation of a wide range of commonly used designs possible. It also comes with an add-on for adaptive methods. Furthermore, the book (Chang 2014) contains R programs for adaptive designs. These are elementary programs for performing sample size reassessment procedures and some basic adaptive randomization designs. The book also comes with SAS macros, most of them performing simulations for the adaptive design described in the book.

Since version 6, SAS comes with some function calls in SAS/IML for doing groups sequential tests (SAS Institute Inc. 1995). Currently available are the SEQ, SEQSCALE, and SEQSHIFT calls. These procedures provide accurate results for computing decision regions, maximum and expected sample sizes, group sequential densities, etc. Examples can be found in Wassmer (1999c), SAS Institute Inc. (1995), Wassmer and Biller (1998), Dmitrienko et al. (2005). Within SAS/IML it is straightforward to produce results for group sequential designs although the calculation of, for example, bias adjusted estimates might become cumbersome. New in SAS 9.2 are procedures for doing group sequential designs in a more comfortable way (SAS Institute Inc. 2009). Specifically, the SEQDESIGN procedure designs interim analyses for clinical trials. It directly computes the boundary values and required sample sizes for the trial within a wide range of possible designs. The SEQTEST procedure performs the interim analyses (tests and confidence intervals) based on design information produced by the SEQDESIGN procedure. SAS currently does not provide any direct capabilities for doing confirmatory adaptive designs as considered in this monograph.

Since the very beginning of adaptive designs the software ADDPLAN was designed for doing confirmatory adaptive designs (www.addplan.com). It is commercially available since 2002 as a tool for designing, simulating, and performing analysis for group sequential designs with an emphasis on the confirmatory adaptive technique. The MC module provides additional multiple comparison features for more than two treatment arms in simulation and analysis, and the PE module additional features for patient enrichment designs in simulation and analysis. There is also the new DF module with capabilities for adaptive dose-finding designs (MCPMod, CRM, etc.).

East from Cytel (www.cytel.com) is a comprehensive tool for design, simulation, and analysis of trials with interim analyses. In the current release, adaptive extensions are provided with the EastAdapt and the EastSurv module. Recently, the modules EastMultiarm and EastEndpoint provide extensions to multi-arm designs and designs with multiple endpoints. An extension to dose finding trials comes with EastEscalate and Cytel's Compass.

We also mention the nQuery module for calculating designs for the group sequential case in the nQuery package (www.statsols.com/products/nquery-advisor-nterim) as well as corresponding capabilities in PASS from NCSS (www.ncss.com). Both do not provide any adaptive extensions but can be used for performing interim decisions and assessing group sequential designs, for example, with respect to maximum and expected sample size.

To summarize, some software is free and hence attractive for statistical research. This is particularly true for the increasing number of available R packages. Simulation-based evaluation of operating characteristics of adaptive designs is becoming increasingly important, some of the available adaptive R packages typically address this issue. The R and SAS packages are available only within the programming environment, whereas the ADDPLAN, EaSt, nQuery, and PASS programs come with a user-friendly graphical user interface (GUI). We note that a free GUI is also available for gsDesign and some other R packages. Within commercially available packages, only ADDPLAN, EastAdapt ,and EastSurv address the specific requirements for confirmatory adaptive designs.

References

Anderson, K. M. (2007). Optimal spending functions for asymmetric group sequential designs. *Biometrical Journal, 49*, 337–345.

Anderson, K. M. (2014). GsDesign: Group sequential design. http://cran.r-project.org/web/packages/gsDesign. R package version 2.9–3.

Anderson, K. M., & Clark, J. B. (2010). Fitting spending functions. *Statistics in Medicine, 29*, 321–327.

Armitage, P. (1957). Restricted sequential procedures. *Biometrika, 44*, 9–56.

Armitage, P. (1975). *Sequential medical trials* (2nd ed.). Oxford: Blackwell.

Armitage, P. (1991). Interim analysis in clinical trials. *Statistics in Medicine, 10*, 925–937.

Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society A, 132*, 235–244.

Barber, S., & Jennison, C. (2002). Optimal asymmetric one-sided group sequential tests. *Biometrika, 89*, 49–60.

Barker, A., Sigman, C., Kelloff, G., Hylton, N., Berry, D., & Esserman, L. (2009). I–SPY 2: An adaptive breast cancer trial design in the setting ofneoadjuvant chemotherapy. *Clinical Pharmacology and Therapeutics, 86*, 97–100.

Barnes, P. J., Pocock, S. J., & Magnussen, H. (2010). Integrating Indacaterol dose selection in a clinical study in COPD using an adaptive seamless design. *Pulmonary Pharmacology & Therapeutics, 23*, 165–171.

Bartroff, J., & Lai, T. L. (2008). Efficient adaptive designs with mid-course sample size adjustment in clinical trials. *Statistics in Medicine, 27*, 1593–1611.

Bartroff, J., Lai, T. L., & Shih, M. -C. (2013). *Sequential experimentation in clinical trials: Design and analysis*. New York: Springer Science & Business Media.

Bauer, P. (1989). Multistage testing with adaptive designs. *Biometrie und Informatik in Medizin und Biologie, 20*, 130–148.

Bauer, P. (1992). The choice of sequential boundaries based on the concept of power spending. *Biometrie und Informatik in Medizin und Biologie, 23*, 3–15.

Bauer, P., Bretz, F., Dragalin, V., König, F., & Wassmer, G. (2016). 25 years of confirmatory adaptive designs: Opportunities and pitfalls. *Statistics in Medicine, 35*, 325–347.

Bauer, P., & Einfalt, J. (2006). Application of adaptive designs - a review. *Biometrical Journal, 8*, 1–16.

Bauer, P., & Kieser, M. (1999). Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine, 34*, 1833–1848.

G. Wassmer, W. Brannath, *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*, Springer Series in Pharmaceutical Statistics,
DOI 10.1007/978-3-319-32562-0

Bauer, P., & Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics, 50*, 1029–1041. (Correction in 1996 *Biometrics, 52*, 380).

Bauer, P., & König, F. (2006). The reassessment of trial perspectives from interim data - a critical view. *Statistics in Medicine, 25*, 23–36.

Bauer, P., & Posch, M. (2004). Letter to the editor: Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections. *Statistics in Medicine, 23*, 1333–1335.

Bauer, P., & Röhmel, J. (1995). An adaptive method for establishing a dose-response relationship. *Statistics in Medicine, 14*, 1595–1607.

Bauer, P., Scheiber, V., & Wohlzogen, F. X. (1986). *Sequentielle statistische Verfahren*. Stuttgart: Gustav Fischer Verlag.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. New York: Springer.

Birkett, M. A., & Day, S. J. (1994). Internal pilot studies for estimating sample size. *Statistics in Medicine, 13*, 2455–2463.

Brannath, W., & Bauer, P. (2004). Optimal conditional error functions for the control of conditional power. *Biometrics, 60*, 715–723.

Brannath, W., Bauer, P., & Posch, M. (2006a). On the efficiency of adaptive designs for flexible interim decisions in clinical trials. *Journal of Statistical Planning and Inference, 136*, 1956–1961.

Brannath, W., & Bretz, F. (2010). Shortcuts for locally consonant closed test procedures. *Journal of the American Statistical Association, 105*, 660–669.

Brannath, W., Burger, H. U., Glimm, E., Stallard, N., Vandemeulebroecke, M., & Wassmer, G. (2010). Comments on the “Draft guidance on adaptive design clinical trials for drugs and biologics" of the U.S. Food and Drug Administration. *Journal of Biopharmaceutical Statistics, 20*, 1125–1131.

Brannath, W., Gutjahr, G., & Bauer, P. (2012). Probabilistic foundation of confirmatory adaptive designs. *Journal of the American Statistical Association, 107*, 824–832.

Brannath, W., König, F., & Bauer, P. (2003). Improved repeated confidence bounds in trials with a maximal goal. *Biometrical Journal, 45*, 311–324.

Brannath, W., König, F., & Bauer, P. (2006b). Estimation in flexible two stage designs. *Statistics in Medicine, 25*, 3366–3381.

Brannath, W., König, F., & Bauer, P. (2007). Multiplicity and flexibility in clinical trials. *Pharmaceutical Statistics, 6*, 205–216.

Brannath, W., Mehta, C. R., & Posch, M. (2009a). Exact confidence bounds following adaptive group sequential tests. *Biometrics, 65*, 539–546.

Brannath, W., Posch, M., & Bauer, P. (2002). Recursive combination tests. *Journal of the American Statistical Association, 97*, 236–244.

Brannath, W., Zuber, E., Branson, M., Bretz, F., Gallo, P., Posch, M., & Racine-Poon, A. (2009b). Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy on oncology. *Statistics in Medicine, 28*, 1445–1463.

Bretz, F., König, F., Brannath, W., Glimm, E., & Posch, M. (2009a). Tutorial in biostatistics: Adaptive designs for confirmatory clinical trials. *Statistics in Medicine, 28*, 1181–1217.

Bretz, F., Maurer, W., Brannath, W., & Posch, M. (2009b). A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine, 28*, 586–604.

Bretz, F., Pinheitro, J. C., & Branson, M. (2005). Combining multiple comparison and modeling techniques in dose-response studies. *Biometrics, 61*, 738–748.

Bretz, F., Posch, M., Glimm, E., Klinglmüller, F., Maurer, W., & Rohmeyer, K. (2011). Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes, or parametric tests. *Biometrical Journal, 53*, 894–913.

Bretz, F., Schmidli, H., König, F., Racine, A., & Maurer, W. (2006). Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: General concepts. *Biometrical Journal, 48*, 623–634.

Bretz, F., & Wang, S. -J. (2010). From adaptive design to modern protocol design for drug development: Part II. success probabilities and effect estimates for phase 3 development programs. *Drug Information Journal, 44*, 333–342.

Brittain, E. H., & Bailey, K. R. (1993). Optimization of multistage testing times and critical values in clinical trials. *Biometrics, 49*, 763–772.

Burman, C. -F., & Sonesson, C. (2006). Are flexible designs sound? *Biometrics, 62*, 664–669.

Burman, C. -F., Sonesson, C., & Guilbaud, O. (2009). A recycling framework for the construction of Bonferroni-based multiple tests. *Statistics in Medicine, 28*, 739–761.

Carreras, M., Gutjahr, G., & Brannath, W. (2015). Adaptive seamless designs with interim treatment selection: A case study in oncology. *Statistics in Medicine, 34*, 1261–1440.

Case, L. D., Morgan, T. M., & Davis, C. E. (1987). Optimal restricted two-stage designs. *Controlled Clinical Trials, 8*, 146–156.

Casper, C., & Perez, O. A. (2014). ldbounds: Lan-DeMets method for group sequential boundaries. http://cran.r-project.org/web/packages/ldbounds. R package version 1.1-1.

Chang, M. N. (1989). Confidence intervals for a normal mean following a group sequential test. *Biometrics, 45*, 247–254.

Chang, M. N. (2014). *Adaptive design theory and implementation using SAS and R* (2nd ed.). Boca Raton: Chapman and Hall/CRC.

Chang, M. N., Gould, A. L., & Snapinn, S. M. (1995). P-values for group sequential testing. *Biometrika, 82*, 650–654.

Chang, M. N., Hwang, I. K., & Shih, W. J. (1998). Group sequential designs using both type I and type II error probability spending functions. *Communications in Statistics - Theory and Methods, 27*, 1323–1339.

Chang, M. N., & O'Brien, P. C. (1986). Confidence intervals following group sequential tests. *Controlled Clinical Trials, 7*, 18–26.

Chaturvedi, P. R., Antonijevic, Z., & Mehta, C. R. (2014). Practical considerations for a two-stage confirmatory adaptive clinical trial design and its implementation: ADVENT trial. In W. He, J. Pinheiro, & O. M. Kuznetsova (Eds.), *Practical considerations for adaptive trial design and implementation* (pp. 77–93). New York: Springer, Science and Business Media.

Chen, Y. H., DeMets, D. L., & Lan, K. K. G. (2004). Increasing the sample size when the unblinded interim result is promising. *Statistics in Medicine, 23*, 1023–1038.

Cheng, Y., & Shen, Y. (2004). Estimation of a parameter and its exact confidence interval following sequential sample size reestimation trials. *Biometrics, 60*, 910–918.

Chi, G. Y. H., & Liu, Q. (1999). The attractiveness of the concept of a prospectively designed two-stage clinical trial. *Journal of Biopharmaceutical Statistics, 9*, 537–547.

Chin, R. (2012). *Adaptive and flexible clinical trials*. Boca Raton: Chapman and Hall/CRC.

Chow, S. C., & Chang, M. N. (Eds.). (2006). *Adaptive design methods in clinical trials*. Boca Raton: Chapman and Hall/CRC.

Coad, D. S., & Woodroofe, M. B. (1996). Corrected confidence intervals after sequential testing with applications to survival analysis. *Biometrika, 83*, 763–777.

Coburger, S., & Wassmer, G. (2001). Conditional point estimation in adaptive group sequential test designs. *Biometrical Journal, 43*, 821–833.

Coburger, S., & Wassmer, G. (2003). Sample size reassessment in adaptive clinical trials using a bias corrected estimate. *Biometrical Journal, 45*, 812–825.

Coe, P. R., & Tamhane, A. C. (1993). Exact repeated confidence intervals for Bernoulli parameters in a group sequential trial. *Controlled Clinical Trials, 14*, 19–29.

Cook, R. J. (1995). Interim analysis in 2 x 2 crossover trials. *Biometrics, 51*, 932–945.

Cook, R. J. (1996). Coupled error spending functions for parallel bivariate sequential tests. *Biometrics, 52*, 442–450.

Cook, T. D. (2002). P-value adjustment in sequential clinical trials. *Biometrics, 58*, 1005–1011.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B, 34*, 187–220.

Cui, L., Hung, H. M. J., & Wang, S. -J. (1999). Modification of sample size in group sequential clinical trials. *Biometrics, 55*, 853–857.

D'Agostino, R. B., Massaro, J. M., & Sullivan, L. M. (2003). Non-inferiority trials: Design concepts and issues - the encounters of academics consultants in statistics. *Statistics in Medicine, 22*, 169–186.

Dallow, N., & Fina, P. (2011). The perils with the misuse of predictive power. *Pharmaceutical Statistics, 10*, 311–317.

DeMets, D. L., Friedman, L. M., & Furberg, C. D. (2006). *Data monitoring in clinical trials*. New York: Springer.

DeMets, D. L., & Gail, M. H. (1985). Use of log-rank tests and group sequential methods at fixed calender times. *Biometrics, 41*, 1039–1044.

DeMets, D. L., & Lan, K. K. G. (1994). Interim analysis: The alpha spending function approach. *Statistics in Medicine, 13*, 1341–1352.

DeMets, D. L., & Ware, J. H. (1980). Group sequential methods for clinical trials with a one-sided hypothesis. *Biometrika, 67*, 651–660.

DeMets, D. L., & Ware, J. H. (1982). Asymmetric group sequential boundaries for monitoring clinical trials. *Biometrika, 69*, 661–663.

Denne, J. S. (2001). Sample size recalculation using conditional power. *Statistics in Medicine, 20*, 2645–2660.

Desseaux, K., & Porcher, R. (2007). Flexible two-stage design with sample size reassessment for survival trials. *Statistics in Medicine, 26*, 5002–5013.

Di Scala, L., & Glimm, E. (2011). Time-to-event analysis with treatment arm selection at interim. *Statistics in Medicine, 30*, 3067–3081. (Correction in 2013 *Statistics in Medicine, 32*, 1974).

Dmitrienko, A., Molenberghs, G., Chuang-Stein, C., & Offen, W. (2005). *Analysis of clinical trials using SAS: A practical guide*. Cary, NC: SAS Press.

Dmitrienko, A., & Wang, M. -D. (2006). Bayesian predictive approach to interim monitoring in clinical trials. *Statistics in Medicine, 25*, 2178–2195.

Dodge, H. F., & Romig, H. G. (1929). A method of sampling inspection. *Bell System Technical Journal, 8*, 613–631.

Donohue, J. F., Fogarty, C., & Lötvall, J. (2010). Once-daily bronchodilators for chronic obstructive pulmonary disease: Indacaterol versus Tiotropium. *American Journal of Respiratory and Critical Care Medicine, 182*, 155–162.

Dragalin, V., Hsuan, F., & Padmanabhan, S. (2007). Adaptive designs for dose-finding studies based on sigmoid E-max model. *Journal of Biopharmaceutical Statistics, 17*, 1051–1070.

Duffy, D. E., & Santner, T. J. (1987). Confidence intervals for a binomial parameter based on multistage tests. *Biometrics, 43*, 81–93.

Eales, J. D., & Jennison, C. (1992). An improved method for deriving optimal one-sided group sequential tests. *Biometrika, 79*, 13–24.

Elfering, G. L., & Schultz, J. R. (1973). Group sequential designs for clinical trials. *Biometrics, 29*, 471–477.

Ellenberg, S. S., Fleming, T. R., & DeMets, D. L. (2003). *Data monitoring committees in clinical trials: A practical perspective*. Chichester: Wiley.

EMA. (2006). *Guideline on clinical trials in small populations (CHMP/EWP/83561/2005)*. London, UK: European Medicines Agency.

EMA. (2007). *Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design*. London, UK: European Medicines Agency.

Emerson, S. S. (1993). Computation of the uniform minimum variance unbiased estimator of the normal mean following a group sequential trial. *Computers and Biomedical Research, 26*, 68–73.

Emerson, S. S. (1996). Statistical packages for group sequential methods. *The American Statistician, 50*, 183–192.

Emerson, S. S. (2006). Issues in the use of adaptive clinical trial designs. *Statistics in Medicine, 25*, 3270–3296; Discussion 3302–3304, 3320–3325, 3326–3347.

Emerson, S. S., & Fleming, T. R. (1989). Symmetric group sequential test designs. *Biometrics, 45*, 905–923.

Emerson, S. S., & Fleming, T. R. (1990). Parameter estimation following group sequential hypothesis testing. *Biometrika, 77*, 875–892.

Emerson, S. S., & Kittelson, J. M. (1997). A computationally simpler algorithm for the UMVUE of a normal mean following a sequential trial. *Biometrics, 53*, 365–369.

Emerson, S. S., Levin, G. P., & Emerson, S. C. (2011). Comments on 'adaptive increase in sample size when interim results are promising: A practical guide with examples'. *Statistics in Medicine, 30*, 3285–3301.

Englert, S., & Kieser, M. (2012). Improving the flexibility and efficiency of phase II designs for oncology trials. *Biometrics, 68*, 886–892.

Englert, S., & Kieser, M. (2015). Methods for proper handling of overrunning and underrunning in phase II designs for oncology trials. *Statistics in Medicine, 34*, 2128–2137.

Fairbanks, K., & Madsen, R. (1982). P values for tests using a repeated significance test design. *Biometrika, 69*, 69–74.

Faldum, A., & Hommel, G. (2007). Strategies for including patients recruited during interim analysis of clinical trials. *Journal of Biopharmaceutical Statistics, 17*, 1211–1225.

Farrington, C. P., & Manning, G. (1990). Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in Medicine, 9*, 1447–1454.

FDA. (2010). *Draft guidance for industry. Adaptive design clinical trials for drugs and biologics*. Food and Drug Administration. Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER), Rockville, MD.

FDA. (2015). *Draft guidance for industry and food and drug administration staff. Adaptive designs for medical device clinical studies*. Food and Drug Administration. Center for Devices and Radiological Health (CDRH) and Center for Biologics Evaluation and Research (CBER), Rockville, MD.

Finner, H., Roters, M., & Strassburger, K. (2015). On the Simes test under dependence. *Statistical Papers*, published online.

Fisher, A., Rosenblum, M., & Jaffee, H. (2014). interAdapt. http://cran.r-project.org/web/packages/interAdapt. R package version 0.1.

Fisher, R. A. (1932). *Statistical methods for research workers* (4th ed.). London: Oliver and Boyd.

Fleiss, J. L. (1986). *The design and analysis of clinical experiments*. New York: Wiley.

Fleming, T. R. (1982). One-sample multiple testing procedure for phase II clinical trials. *Biometrics, 38*, 143–151.

Fleming, T. R., Harrington, D. P., & O'Brien, P. C. (1984). Designs for group sequential trials. *Controlled Clinical Trials, 5*, 348–361.

Follmann, D. A., Proschan, M. A., & Geller, N. L. (1994). Monitoring pairwise comparisons in multi-armed clinical trials. *Biometrics, 50*, 325–336.

Friede, T., & Kieser, M. (2001). A comparison of methods for adaptive sample size adjustment. *Statistics in Medicine, 20*, 3861–3873.

Friede, T., & Kieser, M. (2006). Sample size recalculation in internal pilot study designs: A review. *Biometrical Journal, 48*, 537–555.

Friede, T., Parsons, N., & Stallard, N. (2012). A conditional error function approach for subgroup selection in adaptive clinical trials. *Statistics in Medicine, 31*, 4309–4320 (Correction in 2014 *Statistics in Medicine, 32*, 2513–2514).

Friede, T., Parsons, N., Stallard, N., Todd, S., Valdes-Marquez, E., Chataway, J., & Nicholas, R. (2011). Designing a seamless phase II/III clinical trial using early outcomes for treatment selection: An application in multiple sclerosis. *Statistics in Medicine, 30*, 1528–1540.

Friede, T., & Stallard, N. (2008). A comparison of methods for adaptive treatment selection. *Biometrical Journal, 50*, 767–781.

Gallo, P., DeMets, D. L., & LaVange, L. (2014). Considerations for interim analyses in adaptive trials, and perspectives on the use of DMCs. In W. He, J. Pinheiro, & O. M. Kuznetsova (Eds.), *Practical considerations for adaptive trial design and implementation* (pp. 259–272). New York: Springer, Science and Business Media.

Gao, P., Liu, L., & Mehta, C. R. (2013a). Adaptive designs for noninferiority trials. *Biometrical Journal, 55*, 310–321.

Gao, P., Liu, L., & Mehta, C. R. (2013b). Exact inference for adaptive group sequential designs. *Statistics in Medicine, 32*, 3991–4005.

Gao, P., Liu, L., & Mehta, C. R. (2014). Adaptive sequential testing for multiple comparisons. *Journal of Biopharmaceutical Statistics, 24*, 1035–1058.

Gart, J. J., & Nam, J. -M. (1988). Approximate interval estimation of the ratio of binomial parameters: A review and correction for skewness. *Biometrics, 44*, 323–338.

Geller, N. L. (1994). Discussion of 'Interim analysis: The alpha spending approach'. *Statistics in Medicine, 13*, 1353–1356.

Genz, A., & Bretz, F. (2009). *Computation of multivariate normal and t probabilities* (Vol. 45. p. 247, 279). New York: Springer.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Bornkamp, B., Maechler, M., & Hothorn, T. (2014). mvtnorm: Multivariate normal and t distributions. http://cran.r-project.org/web/packages/mvtnorm. R package version 1.0-2.

Ghosh, B. K. (1970). *Sequential tests of statistical hypotheses*. Reading, MA: Addison-Wesley.

Ghosh, B. K. (1991). A brief history of sequential analysis. In B. K. Ghosh & P. K. Sen (Eds.), *Handbook of sequential analysis* (pp. 1–19). New York: Marcel Dekker.

Glimm, E. (2012). Comments on Adaptive increase in sample size when interim results are promising: A practical guide with examples by CR Mehta and SJ Pocock. *Statistics in Medicine, 31*, 98–99.

Götte, H., Donica, M., & Mordenti, G. (2015). Improving probabilities of correct interim decision in population enrichment designs. *Journal of Biopharmaceutical Statistics, 25*, 1020–1038.

Gould, A. L. (1992). Interim analysis for monitoring clinical trials that do not materially affect the type I error rate. *Statistics in Medicine, 11*, 53–66.

Gould, A. L. (1995). Planning and revising the sample size for a trial. *Statistics in Medicine, 14*, 1039–1051.

Gould, A. L., & Pecore, V. J. (1982). Group sequential methods for clinical trials allowing early acceptance of H_0 and incorporating costs. *Biometrika, 69*, 75–80.

Graf, A. C., & Bauer, P. (2011). Maximum inflation of the type 1 error rate when sample size and allocation rate are adapted in a pre-planned interim look. *Statistics in Medicine, 30*, 1637–1647.

Graf, A. C., Bauer, P., Glimm, E., & König, F. (2014). Maximum type 1 error rate inflation in multiarmed clinical trials with adaptive interim sample size modifications. *Biometrical Journal, 56*, 614–630.

Graf, A. C., Posch, M., & König, F. (2015). Adaptive designs for subpopulation analysis optimizing utility functions. *Biometrical Journal, 57*, 76–89.

Gugerli, U. S., Maurer, W., & Mellein, B. (1993). Internally adaptive designs for parallel group trials. *Drug Information Journal, 27*, 721–732.

Gutjahr, G., Brannath, W., & Bauer, P. (2011). An approach to the conditional error rate principle with nuisance parameters. *Biometrics, 67*, 1039–1046.

Hack, N., Brannath, W., & Brückner, M. (2013). AGSDest: Estimation in adaptive group sequential trials. http://cran.r-project.org/web/packages/AGSDest. R package version 2.1.

Halperin, M., Lan, K., Ware, J. H., Johnson, N. J., & DeMets, D. L. (1982). An aid to data monitoring in long-term clinical trials. *Controlled Clinical Trials, 3*, 311–323.

Hampson, L. V., & Jennison, C. (2013). Group sequential tests for delayed responses (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology, 75*, 3–54.

Hampson, L. V., & Jennison, C. (2015). Optimizing the data combination rule for seamless phase II/III clinical trials. *Statistics in Medicine, 34*, 39–58.

Harrington, D. P., & Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika, 64*, 553–566.

Hartung, J. (2001). A self-designing rule for clinical trials with arbitrary variables. *Controlled Clinical Trials, 22*, 111–116.

Hartung, J., & Knapp, G. (2003). A new class of completely self-designing clinical trials. *Biometrical Journal, 45*, 3–19.

Hauschke, D., Kieser, M., Diletti, E., & Burke, M. (1999). Sample size determination for proving equivalence based in the ratio of two means for normally distributed data. *Statistics in Medicine, 18*, 93–105.

Haybittle, J. L. (1971). Repeated assessments of results in clinical trials of cancer treatment. *British Journal Radiology, 44*, 793–797.

He, W., Pinheiro, J., & Kuznetsova, O. M. (Eds.). (2014). *Practical considerations for adaptive trial design and implementation*. New York: Springer, Science and Business Media.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic: New York.

Hellmich, M. (2001). Monitoring clinical trials with multiple arms. *Biometrics, 57*, 892–898.

Heritier, S., Lô, S. N., & Morgan, C. C. (2011). An adaptive confirmatory trial with treatment selection: practical experiences and unbalanced randomization. *Statistics in Medicine, 30*, 1541–1554.

Herson, J. (2009). *Data and safety monitoring committees in clinical trials*. Boca Raton: CRC Press.

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. New York: Wiley.

Hommel, G. (2001). Adaptive modifications of hypotheses after an interim analysis. *Biometrical Journal, 43*, 581–589.

Hommel, G., & Kropf, S. (2001). Clinical trials with an adaptive choice of hypotheses. *Drug Information Journal, 35*, 1423–1429.

Hung, H. M. J., Wang, S. -J., & O'Neill, R. T. (2011). Flexible design clinical trial methodology in regulatory applications. *Statistics in Medicine, 30*, 1519–1527.

Hung, H. M. J., Wang, S. -J., Tsong, Y., Lawrence, J., & O'Neill, R. T. (2003). Some fundamental issues with non-inferiority testing in active controlled trials. *Statistics in Medicine, 22*, 213–225.

Hünseler, C., Balling, G., Röhlig, C., Blickheuser, R., Trieschmann, U., Lieser, U., Dohna-Schwake, C., Gebauer, C., Möller, O., Hering, F., T., H., Schubert, S., Hentschel, R., Huth, R. G., Müller, A., Müller, C., Wassmer, G., Hahn, M., Harnischmacher, U., Behr, J., & Roth, B. (2014). Continuous infusion of clonidine in ventilated newborns and infants: A randomized controlled trial. *Pediatric Critical Care Medicine, 15*, 511–522.

Huque, M. F. (2016). Validity of the Hochberg procedure revisited for clinical trial applications. *Statistics in Medicine, 35*, 5–20.

Hwang, I. K., Shih, W. J., & DeCani, J. S. (1990). Group sequential designs using a family of Type I error probability spending functions. *Statistics in Medicine, 9*, 1439–1445.

Irle, S., & Schäfer, H. (2014). Interim design modifications in time-to-event studies. *Journal of the American Statistical Association, 107*, 341–348.

Izmirlian, G. (2014). Pwrgsd: Power in a group sequential design. http://cran.r-project.org/web/packages/PwrGSD. R package version 2.0.

Jahn-Eimermacher, A., & Ingel, K. (2009). Adaptive trial design: A general methodology for censored time to event data. *Contemporary Clinical Trials, 30*, 171–177.

Jenkins, M., Stone, A., & Jennison, C. (2011). An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics, 10*, 347–356.

Jennison, C. (1987). Efficient group sequential tests with unpredictable group sizes. *Biometrika, 74*, 155–165.

Jennison, C., & Turnbull, B. W. (1983). Confidence intervals for a binomial parameter following a multistage test with application to MIL–STD 105D and medical trials. *Technometrics, 25*, 49–58.

Jennison, C., & Turnbull, B. W. (1984). Repeated confidence intervals for group sequential clinical trials. *Controlled Clinical Trials, 5*, 33–45.

Jennison, C., & Turnbull, B. W. (1989). Interim analysis: The repeated confidence interval approach. *Journal of the Royal Statistical Society B, 51*, 305–361.

Jennison, C., & Turnbull, B. W. (1991a). Exact calculations for sequential t, χ^2 and F tests. *Biometrika, 78*, 133–141.

Jennison, C., Turnbull, B. W. (1991b). Group sequential tests and repeated confidence intervals. In B. K. Ghosh & P. K. Sen (Eds.), *Handbook of sequential analysis* (pp. 283–311). New York: Marcel Dekker.

Jennison, C., & Turnbull, B. W. (1997a). Distribution theory of group sequential t, χ^2 and F tests for general linear models. *Sequential Analysis, 16*, 295–317.

Jennison, C., & Turnbull, B. W. (1997b). Group sequential analysis incorporating covariate information. *Journal of the American Statistical Association, 92*, 1330–1341.

Jennison, C., & Turnbull, B. W. (2000). *Group sequential methods with applications to clinical trials*. Boca Raton: Chapman & Hall/CRC.

Jennison, C., & Turnbull, B. W. (2003). Mid-course sample size modification in clinical trial. *Statistics in Medicine, 22*, 971–993.

Jennison, C., & Turnbull, B. W. (2006). Adaptive and nonadaptive group sequential tests. *Biometrika, 93*, 1–21.

Jennison, C., & Turnbull, B. W. (2007). Adaptive seamless designs: Selection and prospective testing of hypotheses. *Journal of Biopharmaceutical Statistics, 17*, 1135–1161.

Jennison, C., & Turnbull, B. W. (2015). Adaptive sample size modification in clinical trials: Start small then ask for more? *Statistics in Medicine, 34*, 3793–3810.

Johnson, N. L., & Kotz, S. (1970). *Continuous univariate distributions - 1.* New York: Wiley.

Jones, D., & Whitehead, J. (1979). Sequential forms of the log-rank and modified Wilcoxon tests for censored data. *Biometrika, 66*, 105–133.

Keiding, N. (2006). Event history analysis and the cross-section. *Statistics in Medicine, 25*, 2343–2364.

Keiding, N., Bayer, T., & Watt-Boolsen, S. (1987). Confirmatory analysis of survival data using left truncation of the life times of primary survivors. *Statistics in Medicine, 6*, 939–944.

Kelly, P. J., Stallard, N., & Todd, S. (2005). An adaptive group sequential design for phase II/III clinical trials that select a single treatment from several. *Journal of Biopharmaceutical Statistics, 15*, 641–658.

Kieser, M. (2005). A note on adaptively changing the hierarchy of hypotheses in clinical trials with flexible design. *Drug Information Journal, 39*, 2215–2222.

Kieser, M., Bauer, P., & Lehmacher, W. (1999). Inference on multiple endpoints in clinical trials with adaptive interim analyses. *Biometrical Journal, 41*, 261–277.

Kieser, M., & Friede, T. (2000). Re-calculating the sample size in internal pilot study designs with control of the type I error rate. *Statistics in Medicine, 19*, 901–911.

Kieser, M., Schneider, B., & Friede, T. (2002). A bootstrap procedure for adaptive selection of the test statistic in flexible two-stage designs. *Biometrical Journal, 44*, 641–652.

Kieser, M., & Wassmer, G. (1997). On the use of the upper confidence limit on the variance from a pilot sample for sample size determination. *Biometrical Journal, 38*, 941–949.

Kim, K. (1988). Improved approximation for estimation following closed sequential tests. *Biometrika, 75*, 121–128.

Kim, K., Boucher, H., & Tsiatis, A. A. (1995). Design and analysis of group sequential log-rank tests in maximum duration versus information trials. *Biometrics, 51*, 988–1000.

Kim, K., & DeMets, D. L. (1987a). Confidence intervals following group sequential tests in clinical trials. *Biometrics, 43*, 857–864.

Kim, K., & DeMets, D. L. (1987b). Design and analysis of group sequential tests based on the Type I error spending rate function. *Biometrika, 74*, 149–154.

Kim, K., & Tsiatis, A. A. (1990). Study duration for clinical trials with survival response and early stopping rule. *Biometrics, 46*, 81–92.

Klinglmüller, F., Posch, M., & König, F. (2014). Adaptive graph-based multiple testing procedures. *Pharmaceutical Statistics, 13*, 345–346.

König, F., Brannath, W., Bretz, F., & Posch, M. (2008). Adaptive Dunnett tests for treatment selection. *Statistics in Medicine, 27*, 1612–1625.

Koopman, P. A. R. (1984). Confidence intervals for the ratio of two binomial proportions. *Biometrics, 40*, 513–517.

Köpcke, W. (1984). *Zwischenauswertungen und vorzeitiger Abbruch von Therapiestudien*. Berlin: Springer.

Köpcke, W. (1989). Analyses of group sequential clinical trials. *Controlled Clinical Trials, 10*, 222–230.

Krisam, J., & Kieser, M. (2014). Decision rules for subgroup selection based on a predictive biomarker. *Journal of Biopharmaceutical Statistics, 24*, 188–202.

Kropf, S., Hommel, G., Schmidt, U., Brickwedel, J., & Jepsen, M. S. (2000). Multiple comparison of treatments with stable multivariate tests in a two-stage adaptive design, including a test for non-inferiority. *Biometrical Journal, 42*, 951–965.

Lachin, J. M. (2005). A review of methods for futility stopping based on conditional power. *Statistics in Medicine, 24*, 2747–2764.

Lai, T. L. (1984). Incorporating scientific, ethical and economic considerations into the design of clinical trials in the pharmaceutical industry: A sequential approach. *Communications in Statistics - Theory and Methods, 13*, 2355–2368.

Lan, K. K. G., & DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika, 70*, 659–663.

Lan, K. K. G., & DeMets, D. L. (1989a). Changing frequency of interim analysis in sequential monitoring. *Biometrics, 45*, 1017–1020.

Lan, K. K. G., & DeMets, D. L. (1989b). Group sequential procedures: Calender versus information time. *Statistics in Medicine, 8*, 1191–1198.

Lan, K. K. G., Hu, P., & Proschan, M. A. (2009). A conditional power approach to the evaluation of predictive power. *Statistics in Biopharmaceutical Research, 1*, 131–136.

Lan, K. K. G., & Lachin, J. M. (1990). Implementation of group sequential log-rank tests in a maximum duration trial. *Biometrics, 46*, 759–770.

Lan, K. K. G., Reboussin, D. M., & DeMets, D. L. (1994). Information and information fractions for design and sequential monitoring of clinical trials. *Communications in Statistics - Theory and Methods, 23*, 403–420.

Lan, K. K. G., Simon, R., & Halperin, M. (1982). Stochastically curtailed tests in long-term clinical trials. *Communications in Statistics C, 1*, 207–219.

Lan, K. K. G., & Trost, D. C. (1997). Estimation of parameters and sample size re-estimation. In *Proceedings-Biopharmaceutical Section American Statistical Association* (pp. 48–51). American Statistical Association.

Lang, T., Auterith, A., & Bauer, P. (2000). Trendtests with adaptive scoring. *Biometrical Journal, 42*, 1007–1020.

Lawrence, D., & Bretz, F. (2014). Approaches for optimal dose selection for adaptive design trials. In W. He, J. Pinheiro & O. M. Kuznetsova (Eds.), *Practical considerations for adaptive trial design and implementation* (pp. 125–137). New York: Springer, Science and Business Media.

Lawrence, D., Bretz, F., & Pocock, S. (2014). Inhance: An adaptive confirmatory study with dose selection at interim. In A. Trifilieff (Ed.), *Indacaterol - the first once-daily long-acting Beta2 Agonist for COPD* (pp. 77–92). New York: Springer, Science and Business Media.

Lawrence, J. (2002). Strategies for changing the test statistic during a clinical trial. *Journal of Biopharmaceutical Statistics, 12*, 193–205.

Lawrence, J., & Hung, H. (2003). Estimation and confidence intervals after adjusting the maximum information. *Biometrical Journal, 45*, 143–152.

Lehmacher, W., Kieser, M., & Hothorn, L. (2000). Sequential and multiple testing for dose-response analysis. *Drug Information Journal, 34*, 591–597.

Lehmacher, W., & Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics, 55*, 1286–1290.

Lehmacher, W., Wassmer, G., & Reitmeir, P. (1991). Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. *Biometrics, 47*, 511–521.

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation* (3rd ed.). New York: Springer.

Levin, G. P., Emerson, S. C., & Emerson, S. S. (2013). Adaptive clinical trial designs with pre-specified rules for modifying the sample size: Understanding efficient types of adaptation. *Statistics in Medicine, 32*, 1259–1275.

Li, Z., & Geller, N. L. (1991). On the choice of times for data analysis in group sequential trials. *Biometrics, 47*, 745–750.

Ligges, S. (2012). *Schätzung des Hazard-Ratios in zweiarmigen Überlebenszeitstudien*. Ph.D. thesis, University of Dortmund.

Lin, D. Y., Wei, L. J., & DeMets, D. L. (1991). Exact statistical inference for group sequential trials. *Biometrics, 47*, 1399–1408.

Liu, A., & Hall, W. J. (1999). Unbiased estimation following a group sequential test. *Biometrika, 86*, 71–78.

Liu, Q., & Anderson, K. M. (2008). On adaptive extension of group sequential trials for clinical inverstigations. *Journal of the American Statistical Association, 103*, 1621–1630.

Liu, Q., & Chi, G. Y. H. (2001). On sample size and inference for two-stage adaptive designs. *Biometrics, 57*, 172–177.

Liu, Q., Proschan, M. A., & Pledger, G. W. (2002). A unified theory of two-stage adaptive designs. *Journal of the American Statistical Association, 97*, 1034–1041.

Maca, J., Bhattacharya, S., Dragalin, V., Gallo, P., & Krams, M. (2006). Adaptive seamless phase II/III designs — background, operational aspects, and examples. *Drug Information Journal, 40*, 463–473.

Machin, D., & Campbell, M. J. (1987). *Statistical tables for the design of clinical trials*. Oxford: Blackwell Scientific Publications.

Magirr, D., Jaki, T., König, F., & Posch, M. (2014a). Adaptive survival trials. arXiv preprint arXiv:1405.1569.

Magirr, D., Jaki, T., Posch, M., & Klinglmüller, F. (2013). Simultaneous confidence intervals that are compatible with closed testing in adaptive designs. *Biometrika, 100*, 985–996.

Magirr, D., Stallard, N., & Jaki, T. (2014b). Flexible sequential designs for multi-arm clinical trials. *Statistics in Medicine, 33*, 3269–3279.

Marcus, R., Peritz, E., & Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika, 63*, 655–660.

Maurer, W., Branson, M., & Posch, M. (2010). Adaptive designs and confirmatory hypotheses testing. In A. Dmitrienko, A. C. Tamhane, & F. Bretz (Eds.), *Multiple testing problems in pharmaceutical statistics* (pp. 193–237). Boca Raton: CRC Press.

Mazumbdar, M., & Bang, H. (2008). Sequential and group sequential designs in clinical trials: Guidelines for practitioners. In C. R. Rao, J. P. Miller, & D. C. Rao (Eds.), *Handbook of statistics* (Vol. 27). Amsterdam: Elsevier.

McPherson, C. K., & Armitage, P. (1971). Repeated significance tests on accumulating data when the null hypothesis is not true. *Journal of the Royal Statistical Society A, 134*, 15–25.

McPherson, K. (1982). On choosing the number of interim analyses in clinical trials. *Statistics in Medicine, 1*, 25–36.

McPherson, K. (1990). Sequential stopping rules in clinical trials. *Statistics in Medicine, 9*, 595–600.

Mehta, C. R., Bauer, P., Posch, M., & Brannath, W. (2007). Repeated confidence intervals for adaptive group sequential trials. *Statistics in Medicine, 26*, 5422–5433.

Mehta, C. R., & Gao, P. (2011). Population enrichment designs: Case study of a large multinational trial. *Journal of Biopharmaceutical Statistics, 21*, 831–845.

Mehta, C. R., Gao, P., Bhatt, D. L., Harrington, R. A., Skerjanec, S., & Ware, J. H. (2009). Optimizing trial design sequential, adaptive, and enrichment strategies. *Circulation, 119*, 597–605.

Mehta, C. R., & Pocock, S. J. (2011). Adaptive increase in sample size when interim results are promising: A practical guide with examples. *Statistics in Medicine, 30*, 3267–3284.

Mehta, C. R., Schäfer, H., Daniel, H., & Irle, S. (2014). Biomarker driven population enrichment for adaptive oncology trials with time to event endpoints. *Statistics in Medicine, 33*, 4515–4531.

Miettinen, O., & Nurminen, M. (1985). Comparative analysis of rates. *Statistics in Medicine, 4*, 213–226.

Morgan, C. C., Huyck, S., Jenkins, M., Chen, L., Bedding, A., Coffey, C. S., Gaydos, B., & Wathen, J. K. (2014). Adaptive design: Results of 2012 survey on perception and use. *Therapeutic Innovation & Regulatory Science, 48*, 473–481.

Moyé, L. (2006). *Statistical monitoring of clinical trials. Fundamentals for inverstigators*. New York: Springer, Science and Business Media.

Müller, H. -H., & Schäfer, H. (1999). Optimization of testing times and critical values in sequential equivalence testing. *Statistics in Medicine, 18*, 1769–1788.

Müller, H. -H., & Schäfer, H. (2001). Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential approaches. *Biometrics, 57*, 886–891.

Müller, H. H., & Schäfer, H. (2004). A general statistical principle for changing a design any time during the course of a trial. *Statistics in Medicine, 23*, 2497–2508.

Nam, J. -M. (1995). Confidence limits for the ratio of two binomial proportions based on likelihood scores: Non-iterative method. *Biometrical Journal, 37*, 375–379.

Neuhäuser, M. (2001). An adaptive location-scale test. *Biometrical Journal, 43*, 809–819.

Newcombe, R. G. (1998a). Interval estimation for the difference between independent proportions: Comparison of eleven methods. *Statistics in Medicine, 17*, 873–890.

Newcombe, R. G. (1998b). Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine, 17*, 857–872.

Newcombe, R. G. (2013). *Confidence intervals for proportions and related measures of effect size*. Boca Raton: CRC Press.

O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics, 40*, 1079–1087.

O'Brien, P. C. (1998). Data and safety monitoring. In P. Armitage & T. Colton (Eds.), *Encyclopedia of biostatistics* (pp. 1058–1066). Chichester: Wiley.

O'Brien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics, 35*, 549–556.

Oellrich, S., Freischläger, F., Benner, A., & Kieser, M. (1997). Sample size determination on survival time data - a review. *Informatik, Biometrie und Epidemiolgie in Medizin und Biologie, 28*, 64–85.

Olschewski, M., & Schumacher, M. (1986). Sequential analysis of survival times in clinical trials. *Biometrical Journal, 28*, 273–293.

Ondra, T., Dmitrienko, A., Friede, T., Graf, A., Miller, F., Stallard, N., & Posch, M. (2016). Methods for identification and confirmation of targeted subgroups in clinical trials: A systematic review. *Journal of Biopharmaceutical Statistics, 26*, 99–119.

Pahl, R. (2014). GroupSeq: A GUI-based program to compute probabilities regarding group sequential designs. http://cran.r-project.org/web/packages/GroupSeq. R package version 1.3.2.

Pampallona, S., & Tsiatis, A. A. (1994). Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. *Journal of Statistical Planning and Inference, 42*, 19–35.

Pampallona, S., Tsiatis, A. A., & Kim, K. (2001). Interim monitoring of group sequential trials using spending functions for the Type I and Type II error probabilities. *Drug Information Journal, 35*, 1113–1121.

Parner, E. T., & Keiding, N. (2001). Misspecified proportional hazard models and confirmatory analysis of survival data. *Biometrika, 88*, 459–468.

Parsons, N. (2013). Asd: Simulations for adaptive seamless designs. http://cran.r-project.org/web/packages/asd. R package version 2.0.

Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J., & Smith, P. G. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *British Journal of Cancer, 34*, 585–612.

Pigeot, I., Schäfer, J., Röhmel, J., & Hauschke, D. (2003). Assessing non-inferiority of a new treatment in a three-arm clinical trial including a placebo. *Statistics in Medicine, 22*, 883–899.

Pinheiro, J. C., & DeMets, D. L. (1997). Estimating and reducing bias in group sequential designs with Gaussian independent increment structure. *Biometrika, 84*, 831–845.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika, 64*, 191–199.

Pocock, S. J. (1982). Interim analyses for randomized clinical trials: The group sequential approach. *Biometrics, 38*, 153–162.

Pong, A., & Chow, S. C. (Eds.). (2011). Adaptive design in pharmaceutical and clinical development. Boca Raton: Chapman and Hall/CRC.

Posch, M., & Bauer, P. (1999). Adaptive two stage designs and the conditional error function. *Biometrical Journal, 41*, 689–696.

Posch, M., & Bauer, P. (2000). Interim analysis and sample size assessment. *Biometrics, 56*, 1170–1176.

Posch, M., Bauer, P., & Brannath, W. (2003). Issues in designing flexible trials. *Statistics in Medicine, 22*, 953–969.

Posch, M., König, F., Branson, M., Brannath, W., Dunger-Baldauf, C., & Bauer, P. (2005). Testing and estimating in flexible group sequential designs with adaptive treatment selection. *Statistics in Medicine, 24*, 3697–3714.

Posch, M., Timmesfeld, N., König, F., & Müller, H. -H. (2004). Conditional rejection probabilities of Student's *t*-test and design adaptation. *Biometrical Journal, 46*, 389–403.

Posch, M., Wassmer, G., & Brannath, W. (2008). A note on repeated p-values for group sequential designs. *Biometrika, 95*, 253–256.

Proschan, M. A. (1999). Properties of spending function boundaries. *Biometrika, 86*, 466–473.

Proschan, M. A. (2003). The geometry of two-stage tests. *Statistica Sinica, 13*, 163–177.

Proschan, M. A., Follmann, D. A., & Geller, N. L. (1994). Monitoring multi-armed trials. *Statistics in Medicine, 13*, 1441–1452.

Proschan, M. A., Follmann, D. A., & Waclawiw, M. A. (1992). Effects on assumption violations on type I error rate in group sequential monitoring. *Biometrics, 48*, 1131–1143.

Proschan, M. A., & Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics, 51*, 1315–1324.

Proschan, M. A., Lan, K. K. G., & Wittes, J. T. (2006). *Statistical monitoring of clinical trials*. New York: Springer, Science and Business Media.

Proschan, M. A., Liu, Q., & Hunsberger, S. (2003). Practical midcourse sample size modification in clinical trials. *Controlled Clinical Trials, 24*, 4–15.

Röhmel, J., & Pigeot, I. (2010). A comparison of multiple testing procedures for the gold standard non-inferiority trial. *Journal of Biopharmaceutical Statistics, 20*, 911–926.

Rohmeyer, K., & Klinglmüller, F. (2014). gMCP: Graph based multiple comparison procedures. http://cran.r-project.org/web/packages/gMCP. R package version 0.8-7.

Rosenblum, M. (2015). Adaptive randomized trial designs that cannot be dominated by any standard design at the same total sample size. *Biometrika, 102*, 191–202.

Rosenblum, M., & van der Laan, M. J. (2011). Optimizing randomized trial designs to distinguish which subpopulations benefit from treatment. *Biometrika, 98*, 845–860.

Rosner, G. L., & Tsiatis, A. A. (1988). Exact confidence intervals following a group sequential trial: a comparison of methods. *Biometrika, 75*, 723–729.

Sankoh, A. J. (1999). Interim analysis: An update of an FDA reviewer's experience and perspective. *Drug Information Journal, 33*, 165–176.

Sarkar, S. K., & Chang, C. -K. (1997). The Simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association, 92*, 1601–1608.

SAS Institute Inc. (1995). *SAS/IML software: Changes and enhancements through release 6.11*. Cary, NC: SAS Institute Inc.

SAS Institute Inc. (2009). *What's new in SAS 9.2*. Cary, NC: SAS Institute Inc.

Schäfer, H., & Müller, H. -H. (2001). Modification of the sample size and the schedule of interim analyses in survival trials based on data inspections. *Statistics in Medicine, 20*, 3741–3751.

Scharfstein, D. O., Tsiatis, A. A., & Robins, J. M. (1997). Semiparametric efficiency and its implication on the design and analysis of group-sequential studies. *Journal of the American Statistical Association, 92*, 1342–1350.

Schlömer, P., & Brannath, W. (2013). Group sequential designs for three-arm gold standard non-inferiority trials with fixed margin. *Statistics in Medicine, 32*, 4875–4889.

Schmidli, H., Bretz, F., Racine, A., & Maurer, W. (2006). Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: Applications and practical considerations. *Biometrical Journal, 48*, 635–643.

Schmitz, N. (1993). *Optimal sequentially planned decision procedures*. Berlin: Springer.

Schoenfeld, D. A. (1981). The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika, 68*, 316–319.

Schoenfeld, D. A. (2001). A simple algorithm for designing group sequential clinical trials. *Biometrics, 57*, 972–974.

Schoenfeld, D. A. (2012). Seqmon: Sequential monitoring of clinical trials. http://cran.r-project.org/web/packages/seqmon. R package version 0.2.

Schultz, J. R., Nichol, F. R., Elfring, G. L., & Weed, S. D. (1973). Multiple-stage procedures for drug screening. *Biometrics, 29*, 293–300.

Schumacher, M., & Schulgen, G. (2002). *Methodik klinischer Studien*. Berlin: Springer.

Sellke, T., & Siegmund, S. (1982). Sequential analysis of the proportional hazards model. *Biometrika, 70*, 315–326.

Senn, S., & Bretz, F. (2007). Power and sample size when multiple endpoints are considered. *Pharmaceutical Statistics, 6*, 161–170.

Shen, Y., & Cai, J. (2003). Sample size re-estimation for clinical trials with censored survival data. *Journal of the American Statistical Association, 98*, 418–426.

Shen, Y., & Fisher, L. (1999). Statistical inference for self-designing clinical trials with a one-sided hypothesis. *Biometrics, 55*, 190–197.

Shewhart, W. A. (1931). *Economic control of quality of manufactured products*. New York: Van Nostrand.

Shun, Z., Yuan, W., Brady, W. E., & Hsu, H. (2001). Type I error in sample size re-estimations based on observed treatment difference. *Statistics in Medicine, 20*, 497–513.

Sidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association, 62*, 626–633.

Siegmund, D. (1978). Estimation following sequential test. *Biometrika, 65*, 341–349.

Siegmund, D. (1985). *Sequential analysis*. New York: Springer.

Slepian, D. (1962). The one-sided barrier problem for Gaussian noise. *Bell System Technical Journal, 41*, 463–501.

Slud, E., & Wei, L. J. (1982). Two-sample repeated significance tests based on the modified Wilcoxon statistic. *Journal of the American Statistical Association, 77*, 862–868.

Sonnemann, E. (1991). Kombination unabhängiger Tests. In J. Vollmar (Ed.), *Biometrie in der chemisch–pharmazeutischen Industrie, 4*, 91–112.

Spiegelhalter, D. J., Abrams, K. R., & Myles, J. P. (2004). *Bayesian approaches to clinical trials and health-care evaluation*. New York: Wiley.

Spiegelhalter, D. J., Freedman, L. S., & Blackburn, P. R. (1986). Monitoring clinical trials: Conditional or predictive power? *Controlled Clinical Trials, 7*, 8–17.

Spiessens, B., & Debois, M. (2010). Adjusted significance levels for subgroup analysis in clinical trials. *Contemporary Clinical Trials, 31*, 647–656.

Stallard, N. (2010). A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information. *Statistics in Medicine, 29*, 959–971.

Stallard, N., & Friede, T. (2008). A group-sequential design for clinical trials with treatment selection. *Statistics in Medicine, 27*, 6209–6227.

Stallard, N., Hamborg, T., Parsons, N., & Friede, T. (2014). Adaptive designs for confirmatory clinical trials with subgroup selection. *Journal of Biopharmaceutical Statistics, 24*, 168–187.

Stallard, N., & Todd, S. (2003). Sequential designs for phase III clinical trials incorporating treatment selection. *Statistics in Medicine, 22*, 689–703.

Sugitani, T., Bretz, F., & Maurer, W. (2014). A simple and flexible graphical approach for adaptive group-sequential clinical trials. *Journal of Biopharmaceutical Statistics, 55*, 341–359.

Sugitani, T., Hamasaki, T., & Hamada, C. (2013). Partition testing in confirmatory adaptive designs with structured objectives. *Biometrical Journal, 55*, 341–359.

Temple, R. (1994). Special study designs: Early escape, enrichment, studies in non-responders. *Communications in Statistics - Theory and Methods, 23*, 499–531.

Timmesfeld, N., Schäfer, H., & Müller, H. -H. (2007). Increasing the sample size during clinical trials with t-distributed test statistics without inflating the Type I error rate. *Statistics in Medicine, 26*, 2449–2464.

Todd, S. (2007). A 25-year review of sequential methodology in clinical studies. *Statistics in Medicine, 26*, 237–252.

Todd, S., & Whitehead, J. (1997). Confidence interval calculation for a sequential clinical trial of binary responses. *Biometrika, 84*, 737–743.

Todd, S., Whitehead, J., & Facey, K. M. (1996). Point and interval estimation following a sequential clinical trial. *Biometrika, 83*, 453–461.

Tournoux-Facon, C., De Ryckee, Y., & Tubert-Bitter, P. (2011a). How a new stratified adaptive phase II design could improve targeting population. *Statistics in Medicine, 30*, 1555–1562.

Tournoux-Facon, C., De Ryckee, Y., & Tubert-Bitter, P. (2011b). Targeting population entering phase III trials: A new stratified adaptive phase II design. *Statistics in Medicine, 30*, 801–811.

Troendle, J. F., & Yu, K. F. (1999). Conditional estimation following a group sequential trial. *Communications in Statistics - Theory and Methods, 28*, 1617–1634.

Tsiatis, A. A. (1981). The asymptotic joint distribution of the efficient scores test for the proprtional hazards model calculated over time. *Biometrika, 68*, 311–315.

Tsiatis, A. A. (1982). Repeated significance testing for a general class of statistics used in censored survival analysis. *Journal of the American Statistical Association, 77*, 855–861.

Tsiatis, A. A., Boucher, H., & Kim, K. (1995). Sequential methods for parametric survival models. *Biometrika, 82*, 165–173.

Tsiatis, A. A., & Mehta, C. R. (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika, 90*, 367–378.

Tsiatis, A. A., Rosner, G. L., & Mehta, C. R. (1984). Exact confidence intervals following a group sequential test. *Biometrics, 40*, 797–803.

Tsiatis, A. A., Rosner, G. L., & Tritchler, D. L. (1985). Group sequential tests with censored survival data adjusting for covariates. *Biometrika, 72*, 365–373.

Tymofyeyev, Y. (2014). A review of available software and capabilities for adaptive designs. In W. He, J. Pinheiro, & O. M. Kuznetsova (Eds.), *Practical considerations for adaptive trial design and implementation* (pp. 139–155). New York: Springer, Science and Business Media.

Vandemeulebroecke, M. (2006). An investigation of two-stage tests. *Statistica Sinica, 16*, 933–951.

Vandemeulebroecke, M. (2008). Group sequential and adaptive designs - A review of basic concepts and points of discussion. *Biometrical Journal, 50*, 541–557.

Vandemeulebroecke, M. (2009). Adapttest: Adaptive two-stage tests. http://cran.r-project.org/web/packages/adapttest. R package version 1.0.

Wald, A. (1947). *Sequential analysis*. New York: Wiley.

Wang, M. -D. (2007). Sample size reestimation by Bayesian prediction. *Biometrical Journal, 49*, 365–377.

Wang, S. -J. (2014). A commentary on the US FDA adaptive design draft guidance and EMA reflection paper from a regulatory perspective and regulatory experiences. In W. He, J. Pinheiro, & O. M. Kuznetsova (Eds.), *Practical considerations for adaptive trial design and implementation* (pp. 43–68). New York: Springer, Science and Business Media.

Wang, S. -J., Hung, H. M. J., & O'Neill, R. T. (2009). Adaptive patient enrichment designs in therapeutic trials. *Biometrical Journal, 51*, 358–374.

Wang, S.-J., Hung, H. M. J., Tsong, Y., & Cui, L. (2001). Group sequential test strategies for superiority and non–inferiority hypotheses in active controlled clinical trials. *Statistics in Medicine, 20*, 1903–1912.

Wang, S.-J., O'Neill, R. T., & Hung, H. M. J. (2007). Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharmaceutical Statistics, 6*, 227–244.

Wang, S. K., & Tsiatis, A. A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics, 43*, 193–199.

Wang, Y. G., & Leung, D. H. Y. (1997). Bias reduction via resampling for estimation following sequential tests. *Sequential Analysis, 16*, 249–267.

Wassmer, G. (1998). A comparison of two methods for adaptive interim analyses in clinical trials. *Biometrics, 54*, 696–705.

Wassmer, G. (1999a). Group sequential monitoring with arbitrary inspection times. *Biometrical Journal, 41*, 197–216.

Wassmer, G. (1999b). Multistage adaptive test procedures based on Fisher's product criterion. *Biometrical Journal, 41*, 279–293.

Wassmer, G. (1999c). *Statistische Testverfahren für gruppensequentielle und adaptive Pläne in klinischen Studien. Theoretische Konzepte und deren praktische Umsetzung mit SAS*. Köln: Verlag Alexander Mönch.

Wassmer, G. (2003). Data-driven analysis strategies for proportion studies in adaptive group sequential test designs. *Journal of Biopharmaceutical Statistics, 13*, 585–603.

Wassmer, G. (2006). Planning and analyzing adaptive group sequential survival trials. *Biometrical Journal, 48*, 714–729.

Wassmer, G. (2009). Group sequential designs. In R. B. D'Agostino, L. M. Sullivan, & J. M. Massaro (Eds.), *Encyclopedia of clinical trials*. New York: Wiley.

Wassmer, G. (2010). Adaptive interim analyses in clinical trials. In A. Pong & S. -C. Pong (Eds.), *Handbook of adaptive designs in pharmaceutical and clinical development*. Boca Raton: CRC Press.

Wassmer, G. (2011). On sample size determination in multi-armed confirmatory adaptive designs. *Journal of Biopharmaceutical Statistics, 21*, 802–817.

Wassmer, G., & Biller, C. (1998). Einführung in SAS/IML. In SAS Anwenderhandbuch im Netz. http://www.urz.uni-heidelberg.de/statistik/sas-ah/.

Wassmer, G., & Bock, W. (1999). Tables of Δ-class boundaries for group sequential trials. *Informatik, Biometrie und Epidemiologie in Medizin und Biologie, 30*, 190–194.

Wassmer, G., & Dragalin, V. (2015). Designing issues in confirmatory adaptive population enrichment trials. *Journal of Biopharmaceutical Statistics, 25*, 651–669.

Wassmer, G., Reitmeir, P., Kieser, M., & Lehmacher, W. (1999). Procedures for testing multiple endpoints in clinical trials: An overview. *Journal of Statistical Planning and Inference, 82*, 69–81.

Wetherill, G. B. (1975). *Sequential methods in statistics*. London: Chapman and Hall.

Whitehead, J. (1986). On the bias of maximum likelihood estimation following a sequential test. *Biometrika, 73*, 573–581.

Whitehead, J. (1997). *The design and analysis of sequential clinical trials* (2nd ed.). New York: Wiley.

Whitehead, J. (2001). Sequential methods. In C. K. Redmond & T. Colton (Eds.), *Biostatistics in clinical trials* (pp. 414–422). Chichester: Wiley.

Whitehead, J., Stratton, I. (1983). Group sequential clinical trials with triangular rejection regions. *Biometrics, 39*, 227–236.

Wittes, J., & Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine, 9*, 65–72.

Woodroofe, M. (1992). Estimation after sequential testing: A simple approach for a truncated sequential probability ratio test. *Biometrika, 79*, 347–353.

Zajicek, J. P., Hobart, J. C., Slade, A., Barnes, D., & Mattison, P. G., MUSEC Research Group (2012). Multiple sclerosis and extract of cannabis: Results of the MUSEC trial. *Journal of Neurology, Neurosurgery & Psychiatry, 83*, 1125–1132.

Zeymer, U., Suryapranata, H., Monassier, J. P., Opolski, G., Davies, J., Rasmanis, G., Linssen, G., Tebbe, U., Schröder, R., Tiemann, R., Machnig, T., & Neuhaus, K. L. (2001). The Na+/H+ exchange inhibitor eniporide as an adjunct to early reperfusion therapy for acute myocardial infarction. Results of the evaluation of the safety and cardioprotective effects of eniporide in acute myocardial infarction (ESCAMI) trial. *Journal of the American College of Cardiology, 38*, E1644–E1650.

Zhu, L., Ni, L., & Yao, B. (2011). Group sequential methods and software applications. *The American Statistician, 65*, 127–135.

Index

G. Wassmer, W. Brannath, *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*, Springer Series in Pharmaceutical Statistics,
DOI 10.1007/978-3-319-32562-0