Design and Analysis of Multi-Arm Trials with a Common Control Using Order Restrictions

Muna Arephin

Thesis submitted for the degree of PhD

Abstract

Trials for comparing I treatments with a control are considered, where the aim is to identify one treatment (if at least one exists) which is better than control. Tests are developed which use all of the data simultaneously, rather than combining separate tests of a single arm versus control.

The null hypothesis $H_0 : \Delta_i \leq 0$ is tested against $H_1 : \Delta_i > 0$ for at least one i, where Δ_i represents the scaled difference in response between treatment i and the control, $i = 1, \ldots, I$, and, if rejected, the best treatment is selected. A likelihood ratio test (LRT) is developed using order restricted inference, a family of tests is defined and it is shown that the LRT and Dunnett-type tests are members of this family. Tests are compared by simulation, both under normality and for binary data, an exact test being developed for the latter case.

The LRT compares favourably with other tests in terms of power and a simple loss function. Proportions of subjects on the control close to $(\sqrt{I} - 1)/(I - 1)$ are found to maximise the power and minimise the expected loss.

Two-stage adaptive designs for comparing two experimental arms with a control are developed, in which the trial is stopped early if the difference between the best treatment and the control is less than C_1; otherwise, it continues, with one arm if one experimental treatment is better than the other by at least C_2, or with both arms otherwise. Values of the constants C_1 and C_2 are compared and the adaptive design is found to be more powerful than the fixed design.

The new tests can make a contribution to improving the analysis of multi-arm clinical trials and further research in their application is recommended.

Acknowledgements

I thank Cancer Research UK for generously funding my work with a full-time studentship, without which I would have been unable to start this work. I feel fortunate to have worked in the Wolfson Institute of Preventive Medicine, where I have enjoyed not only a thriving environment for research, but also interacting with other members of the institute.

Many people have supported and encouraged me while I carried out this work.

I would like to thank Professor Peter Sasieni, my supervisor, for his help and support. I cannot thank him enough for his eagle-eyed feedback and comments which were beyond my imagination at times. I also appreciate his understanding and support in difficult personal times.

I would also like to thank Professor Jack Cuzick for his invaluable inputs during his busy schedule and for providing the initial idea for the project.

I would like to thank Professor Stephen Duffy, Professor Deborah Ashby and Dr Steve Coad for their support, without which I may not have finished this work.

I thank the Wolfson librarians and other support staff for their generous help. Thanks to Shahrul for accompanying me on many lunch and tea breaks. Thanks to Jonathan Myles for helping me get started with R. Thanks also to the research students and the Statistics group in the Maths department for making me welcome at seminars and other events.

Thanks to Helen and Neil for support beyond their duty.

As much as it has been a wonderful time for the last four years, it has also been a time of great loss for me. I have lost my precious grandmother and, more recently, I have lost my dearest father. My father, the most beautiful and empathetic soul I have ever known. He always encouraged me in everything I wanted to do and I miss him so much at this time. Finally, I would like to thank Mum for her moral support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

A clinical trial is a randomized experiment on humans for the assessment of one or more treatment regimes for a disease or condition. Clinical trials are widely used in the development of drugs before licensing and for the assessment of licensed drugs or other treatments when one treatment is to be selected for use.

The entire process of a drug's development goes through several stages. After it has been studied on animals and cell cultures in pre-clinical investigations, a drug undergoes toxicological, pharmacological and safe dose selection and then testing for efficacy and confirmation of efficacy. Clinical trials in drug development are usually classified as phase I, phase II or phase III, depending on whether the primary aim is the assessment of toxicity, finding efficacy and the most effective dose, or comparison with the best standard treatment. Phase IV is a post-marketing, rather than an experimental phase, and is used to check for long-term safety issues. However, in practice these divisions may become blurred. For example, when phase II studies are randomized, patient entry can be continued on a control and at least one experimental arm, which leads to a phase II/III study.

In so-called pragmatic trials, one or more potentially improved interventions, perhaps of quite different types, might be tested against current practice, in order to decide whether this practice should be changed. The treatments might be drugs which are already on the market, other types of medical procedure, or even such things as educational interventions.

In this project our main concern is phase III or II/III trials, or pragmatic trials, with efficacy as a primary response. A treatment has to be selected and evidence is needed to show that it is indeed effective by comparing it to a control treatment in a randomized trial.

Clinical trials have been in extensive use for the last four or five decades and statistical methods are used for all types of trials. Different methods and designs have been developed based on the specific demands of different types of trial. New methodological developments in statistical inference are constantly needed to meet the demands of new types of trial. Historically, most clinical trials have involved comparing two arms. Therefore widely developed methods are available for such comparisons. Increasingly, the number of arms is being extended to three or more. However this has been less common than comparing two treatment arms and consequently the methodology is much less developed.

In this thesis, we develop and evaluate methodologies for situations in which more than one experimental treatment is compared with a common control and the objectives of the trial are to establish efficacy of at least one experimental treatment and to select the best treatment. We demonstrate the importance of correctly defining the null hypothesis to meet these objectives and develop appropriate testing procedures. We provide a set of tools for selecting the best treatment where several treatments with a common control are compared, including large sample tests, small sample tests for binary responses, methods for finding suitable sample sizes and allocations and an introduction to sequential adaptive designs.

In the rest of this chapter we briefly discuss some of the background to the work in this thesis. In section 1.2 we describe the methodology which is available in trials with two arms and more than two arms in which the objective is to test for superiority of one or more treatments over a control. Examples of multi-arm trials, which will be used later in the thesis, are introduced. Some of the methodology of hypothesis testing which we will use is described in section 1.3. Design issues in clinical trials which will be used later in the thesis are mentioned in section 1.4, before the aims of the thesis are clarified in section 1.5.

1.2 Clinical trials

1.2.1 Trials with two arms

Two-arm trials for testing superiority have been in use for many years, especially in phase III, and widely developed statistical methods are available. Another possible objective of a two-arm trial is to demonstrate the equivalence of two arms, but this will not be discussed further in this thesis. In a two arm trial for superiority an experimental arm is typically compared with a standard (or placebo) control arm. In comparing two treatments there is a single comparison between treatments and hypothesis testing is simple and unambiguous. If Δ is a difference between treatments, a simple null hypothesis of the form $H_0 : \Delta = 0$ might be tested against the two-sided alternative $H_1 : \Delta \neq 0$.

Often, interest centres on whether or not the experimental treatment is better than the control, with no practical importance attaching to whether it is worse or simply no better. Then the simple null hypothesis $H_0 : \Delta = 0$ or a composite null hypothesis of the form $H_0 : \Delta \leq 0$ against the alternative $H_1 : \Delta > 0$, lead to identical uniformly most powerful tests for Normally distributed data, and identical approximate tests more generally, e.g. chi-squared test, Z-test or t-test depending on the nature of the response. Because the tests are equivalent, there is little or no discussion of which null hypothesis should be used in clinical trials when the alternative hypothesis is one-sided. The only errors in two-arm trials are of type I or type II. Standard methods are available in any statistical text book. Other types of trial are intended to show equivalence, or non-inferiority, of a new treatment, compared with a standard control.

1.2.2 Trials with three or more arms

Multi-arm trials are used for several different purposes. For example, in dose-finding studies, several doses could be used to find the maximum tolerable dose. Alternatively, interest could be in comparing the high dose with the medium dose and the medium dose with the low dose. Three arm trials for assessing non-inferiority of a new treatment can include an active control, usually the standard treatment, and a placebo along with an experimental arm in the so-called gold-standard design (Pigeot et al., 2007; Kieser and Friede, 2007). Non-inferiority trials to compare two new treatments with a control could also be run. The null hypothesis in this case might be that both new treatments are inferior to the control and the alternative is that at least one of the new treatments is not inferior to the control, where inferior is taken to mean a difference less than some pre-specified amount. These hypotheses are quite similar to those we will discuss in this thesis and this will be addressed briefly in Chapter 2, although this is not the main aim here. An equivalence trial would be similar, but two-sided, so will not be discussed further in this thesis.

Trials with more than two arms can have a number of possible treatment structures.

However, the treatment structure alone does not imply which comparisons should be tested. For example, in a three arm trial, the following treatment structures can arise for drugs A and B:

1. A, B and A+B;
2. A, B and standard (or placebo);
3. high dose of A, medium dose of A and low dose of A.

In different trials, different questions will arise, e.g. in 1 it might be desired to test for superiority of the combination against each of the mono-therapies.

Here we consider trials in which it is desired to test two or more experimental treatments against a control and, if efficacy is established for at least one, to select the best one. Instead of running several separate small trials it is more efficient, in terms of time and costs, to run one big trial with a common control and two or more experimental arms. Any of the different structures listed above can lead to an interest in comparing two experimental arms with a common control. All that is required is that two of the arms are experimental and one is a standard treatment or placebo, which is to be used as a control. For 1 and 3, any of the arms can be the control group, depending on the practical application.

Examples of Three arm trials

We introduce three trials which will be referred to later in the thesis. These all involve comparing two experimental treatments with a control, although there is one of each of the structures listed above.

1. The ATAC trial (The ATAC Trialists' Group, 2002), compared different adjuvant treatments for breast cancer, namely tamoxifen, anastrozole and a combination of anastrozole plus tamoxifen. The aim of the trial was to see whether anastrozole alone or anastrozole in combination with tamoxifen was better than tamoxifen alone (the standard treatment at the time). The two drugs work differently but both affect the estrogen response of the tumour, so there was no a priori reason to expect the combination treatment to be at least as good as the individual therapies and it is important to allow the direct comparison of the two mono-therapies. In the event, the treatments were antagonistic and the combined arm did worse that the

anastrozole alone arm. We now know that it is indeed the case that addition of tamoxifen (which can act as a mild estrogen) reduces the efficacy of anastrozole (which prevents the production of estrogen from fat). Had the trial tested only the combination against each of the individual therapies, which is quite typically of interest in combination drug trials, it would not have shown a significant result and a class of dugs (aromatase inhibitors) that have considerably improved the treatment of breast cancer might not have been licensed. This example shows the importance of testing the appropriate hypothesis for the clinical question of interest and not automatically assuming that a particular treatment structure should lead to a particular test.

2. The DASH trial (Appel et al., 1997) compared the effects on blood pressure of three dietary patterns, namely a control diet representing a typical American diet, a diet high in fruit and vegetables and a “combination” diet high in fruit, vegetables and low-fat dairy products. Although the published analysis considered two-sided tests of all pairwise comparisons, it is not desirable that one might fail to make a recommendation because the combination diet is no better than the fruit and vegetable diet, even though the fruit and vegetable diet is better than the control. From a public health perspective one wants to know whether there is convincing evidence (a significant result) that a change of diet will be beneficial and, if so, we make a recommendation as to what change should be made.

3. The MORE trial (Cummings et al., 1999) compared two doses of raloxifene with a placebo for efficacy in breast cancer prevention in post-menopausal women. The aim was to establish evidence that raloxifene was efficacious (through a hypothesis test) and, if it was, to recommend one of the doses. It is not critical whether the high dose is significantly better than the low dose and it is not necessary to model the dose-response relationship. In fact, several studies have failed to find a dose-response relationship. We would certainly not want to fail to reject the null hypothesis if the low dose was as good as or better than the high dose. In the original trial it was found that the low dose was better than the control, but the high dose was no better than the low dose.

1.3 Hypothesis testing

1.3.1 Fundamental concepts

Hypothesis testing is very widely used in analysing the data from clinical trials and in other applications. While in many application areas more emphasis is placed on estimation, in clinical trials regulatory requirements and the very large amounts spent on drug development require an objective methodology for evaluating new treatments and it has become common practice to require that an experimental treatment shows a statistically significant improvement over the control in order to be considered efficacious. Although other approaches, such as Bayesian decision theory, have begun to be influential, they are still less commonly used in practice. In this thesis, we will assume that a hypothesis test will be used as the main criterion of efficacy.

The fundamental idea of hypothesis testing is to state the possible values of one or more parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, i.e. specify the *model*, and then define a *null hypothesis*, usually denoted H_0, which restricts the parameter space to $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \subset \boldsymbol{\Theta}$. The *alternative hypothesis*, usually denoted H_1, is that the parameters lie within the set given by the model, but outside those given by the null hypothesis, i.e. $\boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \boldsymbol{\Theta}_0$. We might have a fully parametric model or study population parameters without fully specifying the population distribution. A hypothesis is said to be *simple* if it completely specifies the parameters of the model, i.e. $\boldsymbol{\Theta}_0$ consists of a single point. Otherwise, it is said to be *composite*. For example, as mentioned in the previous section, the model might be that the difference in response between an experimental and a control treatment is measured by Δ. Then $H_0 : \Delta = 0$ is a simple null hypothesis, $H_0 : \Delta \leq 0$ is a composite null hypothesis, $H_1 : \Delta = 3.5$ is a simple alternative hypothesis (though this is rarely used except for power calculations) and $H_1 : \Delta > 0$ is a composite alternative hypothesis.

The distinction between simple and composite null hypotheses, with a one-sided alternative, is important for the work presented here. Although with two arms there is generally no practical difference, an important point to note for the work described in this thesis is that, if the null hypothesis is $H_0 : \Delta = 0$ and the alternative is $H_1 : \Delta > 0$, then $H_0 \bigcup H_1$ defines a model which does not allow the possibility of the experimental treatment being worse than the control. Although this might sometimes be reasonable, usually it is not and, many authors use $H_0 : \Delta = 0$ when they really mean $H_0 : \Delta \leq 0$. Although it makes no real difference with two arms, the distinction is crucial for the work on multi-arm trials reported in this thesis, since the natural extensions of these different null hypotheses to more than two arms lead to different likelihood ratio test statistics.

It is very common to use hypothesis tests to compare two treatments. A test of $H_0 : \Delta = 0$ against $H_1 : \Delta \neq 0$ is called a *two-sided* test, whereas a test of $H_0 : \Delta = 0$ or $H_0 : \Delta \leq 0$ against $H_1 : \Delta > 0$ is called a *one-sided* test, i.e. the distinction is on the basis of the alternative hypothesis. In comparing more than two treatments, we will likewise refer to a two-sided alternative as one defined by two-sided inequalities, while a one-sided alternative will be defined by one-sided inequalities, i.e. "greater than" or "less than" relationships. Note that this should not be confused with the number of tails of a reference distribution the test statistic is compared with, e.g. in analysis of variance, the F test is used for a two-sided alternative, although we use only one tail of the F distribution.

Let $\mathbf{y}$ be the response data, assumed to be a realisation of a random variable $\mathbf{Y}$. A *test* is a rule of the form: reject H_0 if and only if $\mathbf{Y} \in R$, for some set R known as the *rejection region.* Rejecting H_0 when it is true is known as a *type-I error*, while failing to reject H_0 when it is false is known as a *type-II error.* Let $\gamma(\boldsymbol{\theta})$ be the probability that H_0 is rejected. The *size* of a test, α, is the maximum probability of a type I error, i.e. $\alpha = \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \gamma(\boldsymbol{\theta})$. The *power* of a test is $1 - \beta(\boldsymbol{\theta}) = \gamma(\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \boldsymbol{\Theta}_0$. A size α test is said to be uniformly most powerful (UMP) if it is at least as powerful as any other size α test for all $\boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \boldsymbol{\Theta}_0$. When no UMP test exists, or is known, different methods can be used, but a very common procedure is to use a likelihood ratio test. This rejects H_0 for large values of the test statistic $\lambda = \log \left\{ \frac{L(\hat{\boldsymbol{\theta}};\mathbf{Y})}{L(\tilde{\boldsymbol{\theta}};\mathbf{Y})} \right\}$, where $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta}; \mathbf{Y})$ and $\tilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} L(\boldsymbol{\theta}; \mathbf{Y})$.

For testing a single comparison between two treatments against a one sided alternative, if normality is assumed, a Z-test is the UMP test (and also the LRT) whether the null hypothesis is the simple one of equality, or the compound one of inequality. For the two-sided alternative, there is no UMP test and it is common to use the LRT, which is again the Z-test. For more than two-arms, to test the null distribution of equality against the two-sided alternative, no UMP exists and we usually use a global χ^2 test, which is the LRT. For comparing several arms with a control, with a one-sided alternative, again there is no known UMP. This is related to the fact that different types of differences between the treatments can lead to H_1 being true, e.g. all treatments being better than control, or only one.

Correctly rejecting H_0 when it is false, but for the wrong reason is sometimes known as a *type-III error.* This is most commonly used to describe the rejection of a null hypothesis in favour of a two-sided alternative, when the wrong directional decision is made, e.g. it is concluded that treatment 2 is better than treatment 1 when, in fact, the reverse is true. In practical terms a type-III error is as bad as a type-I error and so the size of the test should be defined to be the maximum probability of a type-I error or a type-III error. This corrected size will be discussed in Section 3.5.

1.3.2 Multiple comparison procedures

The most natural statistical approach to comparing more than one experimental arm with a control would be to compare each experimental arm individually to the control based on a test of significance (Pocock, 1983, p.229). This introduces the problem of multiple testing and the standard solution is to adjust the significance level of each test in order to control the overall significance level, i.e. the probability of rejecting at least one null hypothesis if they are all true, which is known as controlling the family-wise or experiment-wise type I error rate in the weak sense. The family-wise error rate is said to be controlled in the strong sense if the probability of rejecting at least one false null hypothesis is no greater than the significance level, irrespective of how many null hypotheses are actually true. Generally, in this thesis, the discussion of family-wise error rates will refer to controlling them in the weak sense. However, the corrected size used Section 3.5 controls the family-wise error rate more strongly than in the weak sense, but less strongly than in the strong sense. The probability of rejecting at least one true null hypothesis may be greater than the significance level, but the probability of rejecting at least one true null hypothesis and selecting a treatment which is no better than the control, is no greater than the significance level. There are several ways in which multiple testing can arise, for example comparing multiple end points, sequential analysis or others. A recent review of multiple testing is given by Dmitrienko et al. (2010). Comparing several treatment means is a particular type of multiple comparison and comparisons of means with a control is a particular type of this.

By testing more than one such hypothesis, where each test is carried out at level α, the probability of rejecting at least one null hypothesis when they are all true is greater than α. Therefore, to control the overall type I error rate, it is important to make an adjustment for multiple testing, i.e. each test is carried out at a smaller significance level to ensure that the family-wise error rate is α. The most commonly used correction is the classical Bonferroni correction. The Bonferroni correction is simple and so is widely used. However it is too conservative, especially when the test statistics are correlated. In comparing several treatments with a common control correlations arise, so the Bonferroni correction can be extremely conservative. A number of methods have been developed to improve the Bonferroni correction (Simes, 1986; Hommel, 2001).

Hochberg and Tamhane (1987), Hsu (1996) and Dmitrienko et al. (2010) discuss the different types of multiple comparisons that can arise from comparing several means, such as comparing all pairs, comparing each treatment with the best and comparing several experimental treatments with the control, which is the case we are interested in. The most widely used method is that of Dunnett (1955), which compares all experimental treatments to the control. Dunnett's procedure includes a one-sided method if a larger treatment effect than control is sought and it is required to find out which new treatments are better than the control. Originally expressed as simultaneous confidence intervals, Dunnett's method is easily used for hypothesis testing. Each treatment is compared to the control using a one-sided Z-test (or more generally t-test), with the significance level of each test adjusted by some amount to achieve a familywise error rate of α. Calculation of the appropriate significance level of each individual test is not a trivial problem and we will review some of the research in this direction in Chapter 3.

Following Dunnett there have been other developments in this area. Much of it is based on *closed testing procedures* for logically related hypotheses, in which decisions to accept or reject hypotheses must not contradict each other. For example, if the null hypothesis $H_0 : \mu_0 = \mu_1 = \mu_2$ is rejected, then at least one of the null hypotheses $H_0 : \mu_0 = \mu_1$ and $H_0 : \mu_0 = \mu_2$ must be rejected. In the context of multiple comparisons with a control, *step-down* and *step-up* procedures have been defined. In a step-down procedure, e.g. Holm (1979), the overall null hypothesis is tested first. If it is accepted, then all individual comparisons with the control are accepted. If the overall null hypothesis is rejected, then the treatment with the largest observed difference is declared to be different from the control and then a test is performed comparing all other treatments with the control and so on. A step-up procedure, e.g. Hochberg (1988), starts by testing the individual treatment with the smallest observed difference against the control. If it is rejected, then all treatments are declared to be different from the control, while if it is accepted, it steps up to the treatment with the next smallest observed difference, and so on. Since in this thesis, our concern is with the overall null hypothesis, rather than the individual treatment comparisons, step-down procedures just reduce to Dunnett's test, or whichever other test is used for the overall hypothesis. Step-up procedures can lead to new tests, but our interest is only in whether at least one of the individual null hypotheses is rejected.

There is a lot of discussion and controversy about multiple comparison methods in the literature, including several views being expressed that they have no place in the interpretation of data (Mead, 1988). In this thesis, however, we will not actually use multiple comparisons *per se*. Rather, we use them indirectly to obtain an overall hypothesis test, so that the family-wise error rate is exactly the significance level we require and there is no controversy about this. Note also, however, that this means that we are using these procedures for a blunter task than they were intended for, so that they might not be the most appropriate tool. Nevertheless, this approach is by far the most commonly used, perhaps the only one used, in practice when comparing several treatments with a common control.

1.3.3 Order restricted inference

In most regression models we make a strong assumption of a linear (or other) regression function and the parameters are estimated by maximum likelihood or least squares. If no assumption were made regarding the relationship between the response variable Y and the explanatory variable X, then at a point $X = x$, the estimate of $E(Y)$ will be the mean of all responses at $X = x$. The former gives a smooth line, whereas the latter leads to an unconstrained pattern. Between these two extremes, if the researcher has knowledge that the true regression function has a particular ordering then this information can be used to select a regression function. For example in a comparison of three sample means, if it is known that treatment 1 is bigger than treatment 2, which is bigger than treatment 3, then we can make an order restriction and assert this. Such a complete specification is known as a *simple order*. For I experimental treatments and a control, the general form is $\mu_0 \leq \mu_1 \leq \mu_2 \leq \ldots \leq \mu_I$, where μ_i is the mean from treatment i, $i \in \{1, \ldots, I\}$. Taking the order restriction into account we can increase the efficiency of analysis by reducing the expected error, given that the assumed order actually holds.

Any incomplete specification is known as a *partial order*. For example, if we are comparing several treatments with a control or standard group and we use the information that all treatment means are at least as large as the control mean, this is a type of partial order known as the *simple tree order*, i.e. $\mu_0 \leq \mu_i \ \forall i$. Note that the simple tree order $\mu_0 \geq \mu_i \ \forall i$ is dealt with in an equivalent manner and throughout this thesis, we will assume that a high response is desirable. There are other complex relations that are partial orders, e.g. a *matrix order* such as $\mu_1 \leq \mu_2 \leq \mu_4$ and $\mu_1 \leq \mu_3 \leq \mu_4$, but in this thesis we will use only the simple tree order. Detailed texts on order restricted inference, covering testing and estimation for many types of order restriction, are those of Barlow et al. (1972), Robertson et al. (1988) and Silvapulle and Sen (2005).

The set of fitted values which minimise the residual sum of squares, subject to obeying the specified order, is known as the *isotonic regression.* Isotonic regression leads to a wide range of constrained optimization problems, but all the partial orders we discuss in this thesis can be fitted using the *pool adjacent violators algorithm (PAVA)* (Robertson et al., 1988). If the sample means fail to obey the order restriction imposed, we pool observations which violate the order sequentially until the order restrictions are obeyed. Isotonic regression is most often used when there is prior knowledge that the treatments must obey some order restriction. An `R` (R Development Core Team, 2009) function `isoreg` is available to find the isotonic regression estimators for a simple order, with equal sample sizes in each group, but more general `R` functions or other software do not seem to be available and we will not make use of this function.

To illustrate the PAVA algorithm, consider a simple tree order, $\mu_i \geq \mu_0$, $i \in \{1, 2, 3\}$, where the sample means turn out to be $10, 9, 12, 6$ for $i = 0, 1, 2, 3$ and the sample size in each group is equal. Groups 1 and 3 both violate the order restriction by having smaller means than the control and we start by pooling the worst violator (group 3) with the control. This gives a pooled mean of $\bar{Y}_{03} = 8$. Now group 1 no longer violates the order restriction, so that the group 1 and 2 means do not require any pooling. The isotonic regression estimates are $\hat{\mu}_0 = 8, \hat{\mu}_1 = 9, \hat{\mu}_2 = 12$, and $\hat{\mu}_3 = 8$.

Robertson et al. (1988) showed that, for any one-parameter exponential family, as well as for normally distributed data with unknown variance, the isotonic regression estimates are also the maximum likelihood estimates. Therefore MLEs for the simple order, simple tree order and many other partial orders can be obtained from the PAVA algorithm for a wide range of distributional assumptions.

The MLEs can be used to obtain LRT statistics. For example, the null hypothesis of equality of the means $H_0 : \mu_0 = \mu_1 = \mu_2 = \cdots = \mu_I$ can be tested under an order restricted model $\mu_i \geq \mu_0, \forall i$. The LRT for this has been developed and extensively studied (Barlow et al., 1972; Robertson et al., 1988). Because of the restricted model under H_0 having the parameters on the boundary of the parameter space, it does not satisfy the conditions to be asymptotically χ^2 distributed. In fact, it can be shown to be a mixture of χ^2 distributions, but with unknown mixing probabilities (Robertson et al., 1988). Approximating these probabilities has been the focus of considerable research. This test is sometimes called the $\bar{\chi}^2$ test.

In this thesis we will use a simple tree order to define the null hypothesis, but we are not assuming an order restricted model. This leads to a LRT which has been much less studied, although the general form was given by Mukerjee et al. (1985) and Robertson et al. (1988). Again, it does not have a standard distribution and we will approximate its null distribution using simulations.

1.3.4 Exact conditional test

When the responses are binary, large sample tests based on normal approximations are used and, asymptotically, have the same properties as the normal theory tests described above. When the sample sizes are small, however, the approximations are not good enough and these tests are unreliable. For small samples, exact conditional tests are often used instead. Fisher's exact test, for example, compares two treatments to test the null hypothesis of equality against a two-sided alternative. If the control has y_0 successes, out of n_0 patients, and treatment 1 has y_1 successes out of n_1 patients, then we condition on the total number of successes $y_0 + y_1$ and can calculate the probability of r_0 successes on the control, given this total and $n_0 + n_1$, for all possible values r_0. This is simply the hypergeometric probability

$$P(Y_0 = r_0, Y_1 = y_0 + y_1 - r_0 | Y_0 + Y_1 = y_0 + y_1) = \frac{\begin{pmatrix} n_0 \\ r_0 \end{pmatrix} \begin{pmatrix} n_1 \\ y_0 + y_1 - r_0 \end{pmatrix}}{\begin{pmatrix} n_0 + n_1 \\ y_0 + y_1 \end{pmatrix}}.$$

Tests can then be carried out by calculating these probabilities for extreme differences between the success rates on the different treatments. Extensions of Fisher's exact test to multiple treatments are also commonly used - see for example Armitage et al. (2002) for details. Williams (1988) developed a one-sided exact conditional test by assuming an order restricted model - see also Silvapulle and Sen (2005). However, to the best of our knowledge, no one-sided test of this type has been developed which does not restrict the model to having an order restriction.

1.4 Design issues in clinical trials

Most of the work in this thesis addresses issues of hypothesis testing in multi-arm clinical trials, but we will also discuss some issues in the design of trials which are going to be analysed using these tests. Clearly, trials should be designed in order that the analysis is as informative as possible.

In most clinical trials, the main design questions relate to power - see, for example, Armitage et al. (2002). These can either be expressed by calculating the power for a particular sample size, or more commonly by calculating the sample size required to achieve a particular power. We will use the former to compare the properties of different procedures, but will also show how to calculate the latter for practical applications. An extra complication that arises in multi-arm trials, which is straightforward in two-arm trials, at least with one-sided tests, is that of selection of the best treatment.

The power represents the probability that H_0 is correctly rejected when it is false and is a function of the parameters, but we could correctly reject H_0 but then recommend a suboptimal treatment, or even one which is worse than control, i.e. we could make a type-III error. Studying power is therefore not enough and some authors (e.g. Horn and Vollandt (1998)) have previously considered various adjusted powers which take account of the probability of incorrect selection as well. However, even this seems insufficient, because we should take account of the impact of incorrect selection, as well as its probability, through some *loss function* (or *utility function*). Although loss functions are widely used in Bayesian analysis (see, for example, Lee (2004)), they can equally well be used in evaluating the effect of different testing and estimation methods for frequentist inference, although this is uncommon.

In two-arm trials, unless costs dictate otherwise, we would usually aim to allocate equal numbers of patients to the experimental treatment and the control. In comparing several experimental treatments with a control, it is less obvious that we should allocate equal numbers to each treatment. The special status of the control might suggest that it should be given to more patients. However, unless practical limitations dictate otherwise, we should aim for equal numbers of patients on each experimental arm. Thus, an additional design question arises in multi-arm trials, namely the proportion of patients to be allocated to the control. In practice, however, it will sometimes be regarded as administratively, or ethically, necessary to aim for equal allocation to all arms including the control, so we will also place considerable emphasis on this case.

Two further questions are whether the trial is to be of a fixed size, or sequentially designed with a possibility of early stopping, and whether the allocation is fixed at the outset, or adaptively designed with one or more interim analyses. Sequential design (Whitehead, 1997; Jennison and Turnbull, 2003) has received considerable attention recently, motivated by the possibility of reducing the cost of clinical trials by stopping early, either because enough evidence has been accumulated to reject the null hypothesis in favour of an experimental treatment, or because it becomes clear that there is almost no chance of the null hypothesis being rejected (*futility*). As mentioned above, the repeated testing involved in sequential analysis is a type of multiple testing problem and considerable research efforts are continuing in finding suitable adjustments for repeated testing, as described in the books by Whitehead (1997) and Jennison and Turnbull (2003).

Adaptive design is much less commonly used in later phase clinical trials, although a number of methods for adapting the allocation have been suggested in order to expose fewer subjects to inferior treatments. This is done, for example, by biasing the allocation to give a higher probability of each patient being allocated to the treatment which currently seems best, in so-called *biased coin designs* - see, for example, Pocock (1983), Bauer (1989) or Atkinson et al. (2007). There seems to be more scope for adaptive design, however, in multi-arm trials, since dropping treatments after interim analyses is akin to stopping early for futility in the case of two-arm trials and is really quite different from the adaptive allocations which have been suggested for two-arm trials. In this thesis we will consider two-stage adaptive designs for multi-arm trials, where treatments can be dropped at the interim analysis.

1.5 Aims of the thesis

The setting for this thesis is comparing several experimental treatments against a control in clinical trials. We assume that a single hypothesis test will be carried out to establish whether at least one treatment is better than the control. When the null hypothesis is rejected, the treatment with the best estimated response will be recommended, without further hypothesis tests. There is no prior assumption that experimental treatments are no worse than the control, so that an order restriction is defined by our null hypothesis, but the assumed model does not have any order restriction. The main focus of interest is three arm trials with a common control, but some general results for testing in multi-arm trials will be introduced.

The aims of this thesis are to develop and evaluate methods for addressing this problem based on likelihood ratio tests. We also compare them with other methods and consider how best they can be applied in clinical trials. Ultimately, the aim is to improve decision making in clinical trials and to allow them to be run more efficiently.

In Chapter 2, the notation used throughout this thesis is defined, the models and hypotheses we consider are stated and a likelihood ratio test is defined along with several other procedures. In Chapter 3, the properties of these test statistics are compared, particularly with reference to power and expected loss, with the emphasis on three-arm trials. Design issues in three-arm trials, especially the optimal allocation of patients to treatments, but also sample size calculations, are discussed in Chapter 4. In Chapter 5, the performance of the large sample approximations for binary data are assessed and an exact conditional test for small samples, based on the likelihood ratio test statistic, is developed and its properties studied. In Chapter 6, a few adaptive sequential designs, based on likelihood ratio test statistics and single contrasts, are developed and their properties assessed, although there is much more work to be done in this area. Finally, in Chapter 7, some conclusions are drawn and some suggestions for further work are made.

Chapter 2

Model and test procedures

Comparing more than two treatments is less common than comparing two treatments and consequently the statistical methodology is much less well developed. In this chapter, we look at several tests, some obtained by combining one dimensional tests, some which are based on likelihood ratio test statistics and others which are based on single contrasts.

As mentioned in Chapter 1, comparing two (or more) treatment regimes with a control can be considered as a type of multiple testing of the individual comparisons against the control and several authors have addressed the problem in this light. Many trials have compared several treatments with a control and, in this case, several authors have suggested using the famous procedure of Dunnett (1955), or modifications of it (Hsu, 1996). The main aim of these methods is to deal with the problem of adjusting significance levels of the individual tests to control the overall type-I error rate. However, the individual tests, which these methods are intended for, are not directly of interest in the applications we consider. Our aim is to establish that at least one of the new treatments is efficacious and to select the best treatment. It is not of primary interest to identify all efficacious treatments.

Other authors (Mukerjee et al., 1987; Tang and Lin, 1997; Peddada et al., 2006; Zhao, 2007) have addressed the problem more directly by using the methods of order restricted inference (Barlow et al., 1972; Robertson et al., 1988; Silvapulle and Sen, 2005). In this work a single hypothesis test is carried out and so there is no issue regarding the significance level or p-value, except that calculating it might be complicated. Much of the order restricted inference literature is devoted to likelihood based methods for the null hypothesis of equality of all the treatments against a one-sided alternative. Unlike in two-arm trials, this gives a different likelihood ratio test from the null hypothesis that no treatment is better than the control and it implicitly assumes some prior knowledge about the ordering of treatments, i.e. that the experimental arms are at least as good as the control. This might be appropriate in some circumstances but not for the applications considered in this thesis. More generally, it has been noted (Mukerjee et al., 1987) that methods based on order restricted inference are little used in clinical trials, probably due to lack of awareness of their existence.

In this chapter we define our model, the hypotheses we will test and several test statistics. The model and the notation we use throughout this thesis are defined in Section 2.1. The rationale for using a different null hypothesis from that tested by many of the procedures available in the literature is discussed in Section 2.2. The test statistics being considered are defined in Section 2.3 and the rationales for the test statistics used are discussed. Important relationships between the test statistics are derived, with the detailed proofs being given in the appendix to this chapter in Section 2.5. The relevant literature is reviewed as appropriate in each of these Sections. Finally, some conclusions are drawn in Section 2.4

2.1 Model and notation

We consider a trial with I experimental arms and a common "control". Here the control can be a standard therapy, not necessarily "no therapy" (or placebo). Let N be the total number of subjects in all arms and assume that N is large. We present results for the case where there are $n_i = \delta N$ subjects in each experimental arm and we let $n_0 = (1 - I\delta)N$ be the number of subjects in the control arm. Let Z_i for $i = 1, \ldots, I$ be a normalised statistic comparing treatment arm i with the control. For example, Z_i could be the standardised difference between the mean responses for continuous data, the log odds ratio for binary data, or the log hazard ratio for survival data. Since we have large samples, it follows, as usual, that $\mathbf{Z} = (Z_1, \ldots, Z_I)'$ is multivariate normal with $Var(Z_i) = 1$ and $\rho = Cov(Z_i, Z_j) = \delta/\{1 - (I-1)\delta\}$, $i \neq j$. The Z_is are correlated due to their dependence on a common control arm. The expectation of Z_i is a measure of the difference between treatment i and the control and we define $E(Z_i) = \Delta_i/\sigma$, where $\sigma^2 = \{1 - (I-1)\delta\}/\{\delta(1 - I\delta)\}$. For $I = 2$, for example,

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \Delta_1/\sigma \\ \Delta_2/\sigma \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right).$$

With equal allocation, i.e. $\delta = 1/3$, the correlation is $\rho = 1/2$.

Such Z_i could, for instance, be contrasts based on normally distributed observations, with known, constant variance. Let Y_{ij} be the (scaled) response from subject j on treatment $i = 0, 1, \ldots, I$, where $i = 0$ is the control arm. All observations are assumed to be independent. After rescaling to get unit variance, we have $Y_{0j} \sim N(\mu_0, 1)$ and $Y_{ij} \sim N(\mu_i, 1)$, $i = 1, \ldots, I$. Then $\Delta_i = \sqrt{N}(\mu_i - \mu_0)$ and $Z_i = \sqrt{N}\{(\bar{Y_i} - \bar{Y_0})/\sigma\}$. In the more typical case in which $Y_{ij} \sim N(\mu_i, \tau^2)$, with unknown variance, one could use the t-statistic for Z_i and asymptotically the assumed normal distribution of Z_i would still be correct.

With binary responses, assume $Y_{ij} \sim Bernoulli(\pi_i)$ and let

$$\psi_i = \frac{\pi_i/(1-\pi_i)}{\pi_0/(1-\pi_0)},$$

$i = 0, 1, \ldots, I$, where π_i is the success rate in treatment i. For large N, $\log \hat{\psi}_i \sim N(\log \psi_i, Var(\log \hat{\psi}_i))$ (Armitage et al., 2002, p.127), where

$$\hat{\psi}_i = \frac{\hat{\pi}_i/(1-\hat{\pi}_i)}{\hat{\pi}_0/(1-\hat{\pi}_0)}$$

and

$$Var(\log \hat{\psi}_i) = \frac{1}{n_0\pi_0(1-\pi_0)} + \frac{1}{n_i\pi_i(1-\pi_i)}.$$

Then we use

$$\widehat{Var}(\log \hat{\psi}_i) = \frac{1}{n_0\hat{\pi}_0(1-\hat{\pi}_0)} + \frac{1}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}.$$

We scale the log odds ratios to get Z_i with unit variance, so that

$$\begin{aligned} Z_i &= \frac{\log \hat{\psi}_i}{\sqrt{\widehat{Var}(\log \hat{\psi}_i)}} \\ &= \log\left(\frac{\hat{\pi}_i/(1-\hat{\pi}_i)}{\hat{\pi}_0/(1-\hat{\pi}_0)}\right)\sqrt{\frac{n_0 n_i \hat{\pi}_0 \hat{\pi}_i (1-\hat{\pi}_0)(1-\hat{\pi}_i)}{n_i\hat{\pi}_i(1-\hat{\pi}_i) + n_0\hat{\pi}_0(1-\hat{\pi}_0)}}. \end{aligned}$$

Then the Z_is follow the assumptions above.

Similarly, for survival data, if Cox's proportional hazards model is used then we can calculate our Z_i statistics as follows. Let Y_{ij} be the survival time of the jth patient on the ith treatment, being independent with some unknown distribution with hazard $h_i(t) = h_0(t)\exp(\Delta_i)$ for the ith active arm, where $h_0(t)$ is the baseline hazard at time t in the control arm. Then the hazard ratio $HR_i = h_i(t)/h_0(t) = \exp(\Delta_i)$. A natural and asymptotically normal test statistic will be based on the estimated log hazard ratio, i.e. $Z_i = \log(\widehat{HR_i})/se(\log(\widehat{HR_i}))$, which again meets the assumptions above. Alternatively, one could use the log rank statistics comparing arm i to the control arm.

Thus the methods developed and described in this thesis are applicable whenever inference regarding a single treatment compared to control would be based on a statistic that is (approximately) normally distributed. The specific examples of models just given cover most applications in clinical trials. The only restriction to the work given here is that it applies only to large sample sizes. Within this context, however, the results are very general.

It will often be useful to reorder the Z_i and we will use the notation X_j to represent the jth largest Z_i, so that $X_1 > X_2 > \cdots > X_I$. The joint likelihood of the data is

$$L(\Delta_1, ..., \Delta_I; z_1, ..., z_I) = \phi(z_1 - \Delta_1/\sigma, ..., z_I - \Delta_I/\sigma; \rho), \tag{2.1}$$

where $\phi(\cdot;\rho)$ is the probability density function of a multivariate normal random variable with mean vector zero, unit variances and a known constant covariance ρ between all pairs of variables.

2.2 Hypotheses

Given the model defined in Section 2.1, there are several null and alternative hypotheses which can be defined. We will use standard notations for these throughout this thesis, since many apparently similar methods in the literature actually test different hypotheses, or make different model assumptions. To test an individual experimental treatment i against the control, or to test each Δ_i as a separate hypothesis, we have $H^*_{0i} : \Delta_i = 0$ against $H^*_{1i} : \Delta_i \neq 0$, for a two-sided test, and $H_{0i} : \Delta_i \leq 0$ against $H_{1i} : \Delta_i > 0$, for a one-sided test. These would typically be tested using a two- or one-sided Z-test respectively, which are the likelihood ratio tests.

Considering all experimental arms together, we define the null hypothesis of equality $H^*_0 : \mathbf{\Delta} = \mathbf{0}$ against the alternative $H^*_1 : \mathbf{\Delta} \neq \mathbf{0}$, where $\mathbf{\Delta} = [\Delta_1, \ldots, \Delta_I]'$ and $\mathbf{0}$ is an I-dimensional vector of zeros. This would typically be tested using a χ^2-test. Finally, we define the hypotheses which are of interest in this thesis, $H_0 : \Delta_i \leq 0,\ \forall i = 1, \ldots, I$, against $H_1 : \Delta_i > 0$, for at least one i. In this case, it is not immediately obvious what the test should be. Various options are described in this chapter.

It might seem from the previous paragraph that, when considering all arms together, there are only two possibilities, corresponding to one-sided and two-sided tests. However, most of the literature on order restricted inference concentrates on testing the equality null H^*_0 against the one-sided alternative H_1. This implicitly assumes that we are working under the order restricted model, i.e. under the assumption that no experimental treatment can be worse than the control.

As can be seen, there are several different hypotheses which can be tested. Careful thought must be given to which is appropriate for a specific trial, because apparently minor differences can lead to quite different results. We assume that the aim of the trial is to demonstrate that at least one experimental treatment is better than the control and to select the best treatment. We argue that a single hypothesis test of $H_0 : \Delta_i \leq 0,\ i = 1, \ldots, I$, against $H_1 : \Delta_i > 0$, for at least one i, should be carried out to establish (or fail to establish) superiority of at least one experimental treatment over the control. The aim of the test is typically to convince regulators, potential users, senior management or the public that our trial provides sufficient evidence to demonstrate improvement over the standard treatment (the control). This is in contrast to multiple comparison methods for comparing several experimental treatments with a control, which aim to test H_{0i} against H_{1i}, for all $i \in \{1, \ldots, I\}$, i.e. each experimental treatment is tested against the control. However, in our applications, there is no need to find all treatments which are efficacious, since we assume that ultimately only one will be recommended for use. Multiple comparison tests would be appropriate if, for example, each experimental arm was being considered separately for licensing, so that we need to make I separate decisions.

When the null hypothesis is rejected we need to select a treatment from among the experimental arms. When the ordering of the estimated effects is the same as the ordering of the significance of the effects, then selection is straightforward: one simply selects the treatment corresponding to the largest estimated effect. That is the situation that we consider here because, by design, (approximately) equal numbers will be randomized to all treatment arms (except possibly the control arm). It might be argued that it is necessary to test the best treatment against each of the others to establish that it really is the best. However, in the types of trials we are discussing, this is meaningless. Consider a situation in which at least one experimental treatment is significantly better than the control, but there is no significant difference between the best two experimental arms. What then do we recommend? Usually, a single treatment will be recommended for practical use and clearly we will recommend that which we estimate to give the best response. Therefore, even if we carry out a test of each treatment against the best, it will have no impact on the final recommendation. Such tests might be useful in other situations, for example, if it was intended to withhold licenses from any treatment deemed to be inferior to the best. Note that here we discuss only efficacy, assuming that safety issues are dealt with separately.

In summary, we consider multi-arm trials, in which I experimental treatments are compared with a control. We test $H_0 : \Delta_i \leq 0 \forall i$ against $H_1 : \Delta_i > 0$ for at least one i, to establish efficacy. If H_0 is rejected, treatment i^* is recommended, where $i^* = \arg\max Z_i$. This seems to be logically consistent with how the null hypothesis of equality H_0^* is tested against the two-sided alternative H_1^* when the aim is to recommend one of several treatments, with none being identified as a control. Then, typically, we would carry out a χ^2 test to ensure that we are not just interpreting noise and then recommend the treatment which gives the best estimated response. Multiple comparison tests of all pairs of treatments do not add anything which might change our recommendation.

Of course not all three-arm trials fit into the framework described here. For example, if the treatments are different doses of a drug, a dose response model might be assumed, especially in early phase trials, or a strictly non-decreasing response (a *simple order*) might be assumed. However in several situations a threshold appears to occur so that dose response models would be inefficient. In other cases it might be known *a priori* that the experimental treatments will be no worse than the control, so that the null hypothesis of equality can be used. However there are pragmatic trials, which seek to recommend a treatment for application, and some phase-II/III trials in drug development, where there are two or more new candidate drugs, for which it is not reasonable to assume anything about the ordering of the treatments and the objectives can be met by testing our hypothesis. In these circumstances, the testing and selection procedure described here is appropriate, although it seems to be quite rarely used. In this thesis, we compare and develop methodologies required to carry out this procedure efficiently.

The three trials described briefly in Chapter 1 all fit into the framework described here. Consider the ATAC trial for comparing tamoxifen, the common control, anastrozole and a combination of anastrozole plus tamoxifen. In this situation the above hypothesis is appropriate. It will provide the evidence in favour of a new treatment over the standard and, if this is shown, either anastrazole or the combination would be recommended. In fact, anastrozole alone turned out to be beneficial and we will see the data analysis in Chapter 3.

The DASH trial compared a control diet, a diet high in fruit and vegetables and a “combination” diet high in fruit, vegetables and low-fat dairy products on blood pressure. As noted in Chapter 1, one wants to know whether there is convincing evidence (a significant result) that a change of diet will be beneficial and, if so, we make a recommendation as to what change should be made. This is exactly what our recommended procedure does. In the trial, Appel et al. (1997) actually performed all pairwise comparisons, but we see no benefit in doing this. If it is clear that a change in diet is beneficial, but there is no significant difference between the two experimental treatments, we would not wish to withhold a public health recommendation while further comparison of these two diets is carried out. Instead, we would simply recommend that which seems best and any further trials could be done with this as the new control.

In the MORE trial, the point was again to establish evidence that raloxifene was efficacious (through a hypothesis test) and, if it was, to recommend one of the doses. The dose-response relationship is not of interest in itself, although the use of our testing and selection procedure does not preclude further such secondary analyses from being carried out.

Thus the methods developed in this thesis are relevant to all of these examples. In all cases, the crucial question is whether at least one experimental treatment is better than the control. In these applications, it is necessary to carry out a hypothesis test in order to establish improved efficacy of at least one experimental treatment over the control. When efficacy is demonstrated, it is clear that the experimental treatment with the best estimated response will be recommended for use in practice. It is not necessary to demonstrate that the recommended arm is superior to all other experimental arms, nor is it necessary to establish individually whether or not each experimental arm is better than the control.

An important question is whether or not the selected experimental arm, i.e. the one with largest Z_i, has a mean response greater than that for the control arm. In addition to the usual type I and type II errors, another type of error can happen. It is also possible to correctly reject the null hypothesis but to select the wrong experimental arm, e.g. for I=2, to reject H_0 and select experimental arm 2 when $\Delta_2 \leq 0$ but $\Delta_1 > 0$. For practical purposes this can be treated like a type I error, but is associated with parameter values that are within the set of alternatives. It is logically analogous to carrying out a two-sided test in a two arm trial, rejecting the null hypothesis with a positive value of the test statistic and then wrongly concluding that the experimental treatment is better than (rather than simply different from) the control. Both of these are type-III errors, as described by Mosteller (1948) and formally defined by Harter (1957). Most of the literature which refers to type-III errors concerns two-sided tests in which a wrong directional decision is made. As we will see in Chapter 3, however, some authors have also discussed type-III errors in contexts much closer to ours.

We can formally take account of the probability of a type-III error in defining the size of the test. The corrected size is defined as

$$\alpha^{\dagger} = \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left(\alpha, P(\text{type-III error}) \right),$$

where $\boldsymbol{\theta}$ is a vector of parameters and $\boldsymbol{\Theta}$ is the set of all possible $\boldsymbol{\theta}$. $\alpha^\dagger$ is the probability of rejecting H_0 in favour of $i^* \mid i^*$ is no better than control.

In two arm trials, a two-sided test has $\alpha^\dagger = \alpha$. However, in multi-arm trials with one-sided tests this is not necessarily true. The probabilities of type-III errors are discussed further in Section 3.5.

For $0 < \Delta_2 < \Delta_1$, the related outcome of selecting treatment 2 will be called a type-IV error, i.e. we select a treatment which is suboptimal, but is still better than the control. Type-III errors could have very deleterious consequences, but will be quite rare, whereas type-IV errors will be much more common, but have consequences which are less serious, since they still lead to an improvement in treatment. Different authors have included type-III or -IV errors in their power comparisons in different ways and we will describe some of these in more detail in Chapter 3.

In two-arm trials these complications do not arise with one-sided tests, since there is a direct correspondence between a one-sided test and selection of a treatment. In multi-arm trials we have to consider the possibility of correctly rejecting the null hypothesis, but selecting a sub-optimal treatment, and particularly the possibility of selecting a treatment that is inferior to control.

2.3 Test procedures

Since both the null hypothesis H_0 and the alternative H_1 are composite hypotheses, a uniformly most powerful test is not known, so we now describe several possible test procedures. We also discuss relationships among the different test procedures and how they are related to other procedures in the literature. The performances of the most promising procedures are compared in the next chapter.

2.3.1 Families of test statistics

First, we define three families of test statistics, which have not appeared in the literature in these forms. In later subsections, we will show that several procedures which have appeared in the literature can be rewritten in terms of particular members of these families of test statistics and this allows us to see relationships between them, as well as simplifying the terminology and notation.

First, we consider test statistics which arise naturally in two-sided tests and define

$$S_k^0 = \left(\sum_{i=1}^{I} Z_i^k \right)^{\frac{1}{k}}$$

for some integer $k > 0$. Because we take positive powers, the larger k is, the more weight is put on effects which are further from zero. However, this family of test statistics does not seem sensible for the one-sided alternative, unless there is prior knowledge that no treatment can be worse than the control, since negative values of Z_i can dominate.

It seems logical to only include contributions from Z_i which take positive values, so a natural family of test statistics is

$$T_k^0 = \left(\sum_{i=1}^{I} Z_i^{+k} \right)^{\frac{1}{k}}, \tag{2.2}$$

where $A^+ = \max(0, A)$ and $k > 0$. Here larger values of k give more weight to treatments which give larger responses, as long as these are greater than the control. If the Z_i were independent, then there would be no reason to adjust the contributions from different Z_i, so that using T_k^0 would be logical. However, the correlation between them means that the knowledge that any one Z_i is large makes it more likely that the others will be large and including the full contribution from each exaggerates the evidence against H_0. An unusually small estimate for the control will make all other arms appear to be good.

We adjust for the correlation among Z_i using the general results for conditional multivariate normal random variables (Krzanowski and Marriott, 1994, p.25) and evaluating expectations at $\Delta_i = 0, \ \forall i = 1, \ldots, I$. Then

$$E\left(Z_i | Z_1, \ldots, Z_{i-1}\right) = \frac{\rho}{\{1 + (i-2)\rho\}} \sum_{j=1}^{i-1} Z_j \tag{2.3}$$

and

$$Var\left(Z_i | Z_1, \ldots, Z_{i-1}\right) = \frac{\{1 + (i-1)\rho\}(1-\rho)}{\{1 + (i-2)\rho\}}, \tag{2.4}$$

for $i \in \{1, \ldots, I\}$.

Considering the ordered univariate statistics, $X_1 > \cdots > X_I$, it is clear that if $X_1 < 0$, there is no evidence against H_0 and any sensible test statistic should take value zero. Hence the first contribution to a test statistic should come from X_1^+. Using the conditional expectation and variance results, X_i should only contribute to the test statistic if it is greater than

$$E_i = \frac{\rho}{\{1 + (i-2)\rho\}} \sum_{j=1}^{i-1} X_j, \tag{2.5}$$

so that the contribution should be based on $(X_i - E_i)^+$. This is scaled to have unit variance, so that the ith term in the test statistic should be $(X_i - E_i)^+/\sqrt{V_i}$, where

$$V_i = \frac{\{1 + (i-1)\,\rho\}\,(1-\rho)}{\{1 + (i-2)\,\rho\}}. \tag{2.6}$$

It remains only to decide on which scale these terms should jointly contribute to the test statistic. We can define a family of test statistics

$$T_k = \left[X_1^{+k} + \sum_{i=2}^{I} \left\{ \frac{(X_i - E_i)^+}{\sqrt{V_i}} \right\}^k \right]^{\frac{1}{k}}, \tag{2.7}$$

for $k > 0$. Clearly larger values of k will put more emphasis on the treatment with the largest effect. As k is decreased more weight is given to the other treatments which have positive effects.

In the following sections we will consider some members of these families, in particular with $k = 1, 2, \infty$.

2.3.2 Combining one-dimensional tests

Dunnett-type tests

Since we want to choose the best treatment, a natural and simple approach is to consider each Δ_i as a separate hypothesis, test the one-sided alternative using the uniformly most powerful test and then combine the results. Thus if one computes the usual test statistic for $H_{0i} : \Delta_i \leq 0$ against $H_{1i} : \Delta_i > 0$, for each i, and takes the maximum, one obtains $\max(Z_i)$, where $i = 1, ..., I$. We will actually use

$$T_\infty = \max(0, Z_1, \dots, Z_I),$$

since this is a member of the family of test statistics defined in (2.7), although this makes no difference for any sensible significance level since the critical value will be greater than zero. Note also that $T_\infty^0 = T_\infty$. The multiple comparisons being made must be taken into account when calculating the rejection boundaries, which is usually done using the well known procedure of Dunnett (1955), originally developed for obtaining confidence intervals for each comparison with the control, while controlling the overall coverage. Note that Dunnett's procedure is defined using t-statistics (assuming unknown variance), so that T_∞ can be obtained from the asymptotic case as the degrees of freedom tend to ∞.

Assume that Dunnett's procedure is used to obtain one-sided confidence intervals for

$\Delta_i,\ i = 1, \dots, I$ of the form $(\hat{\Delta}_i - C, \infty)$, for some positive constant C such that

$$P\left(\bigcap_{i=1}^{I}\{\hat{\Delta}_i - C < \Delta_i\}\right) = 1 - \alpha$$

and that we reject H_0 if at least one confidence interval excludes 0. Then, we reject H_0 if $0 \notin (\hat{\Delta}_i - C, \infty)$ for at least one i $\Rightarrow$ reject H_0 if $0 < \hat{\Delta}_i - C$ for at least one i $\Rightarrow$ reject H_0 if $\max(\hat{\Delta}_i) > C$, which is equivalent to T_∞, since $C > 0$.

Note also that $P(\text{Do not reject } H_0) = P(\hat{\Delta}_i < C\ \forall i) = P\left(\bigcap_{i=1}^{I}\{\hat{\Delta}_i - C < 0\}\right)$. Hence, by considering $\Delta_i = 0\ \forall i$, we see that the test has size α and C is the cutpoint of the rejection region for T_∞. Considering $\Delta_i \geq 0$, with at least one strict inequality, it is clear that maximising C is equivalent to minimising the power of T_∞ for a test of fixed size.

Hence, published results on Dunnett's procedure apply immediately to T_∞. Much research effort has been expended on finding good approximations to the cutpoint C. Most authors who discuss testing H_0, as defined above, have used Dunnett's test and most other tests in the literature are similar to this test. This approach makes no use of the size of the treatment differences other than the maximum. Although it is commonly used in practice and is known to have high power if only one treatment is better than the control, there is no reason to expect it to have good properties, such as high power, in all situations.

Hochberg's procedure

The step-up procedure of Hochberg (1988) for multiple testing gives a simple modification of T_∞, which allows us to reject $H_0 : \Delta_i \leq 0\ \forall i$, if several treatment differences are big but not quite big enough to cross the rejection boundary for T_∞. To test I hypotheses, order the p-values for the independent hypothesis tests $p_{(1)} < p_{(2)} < \cdots < p_{(I)}$, corresponding to reordered null hypotheses $H_{0(1)}, H_{0(2)}, \dots, H_{0(I)}$ where $H_{0(i)}$ is $H_{0j} : \Delta_j \leq 0$ for j such that $p_{(i)}$ corresponds to H_{0j}. If $p_{(k)} < \alpha/(I - k + 1)$ for any $k = 1, ..., I$, then reject $H_{0(i)}$ at the $100\alpha\%$ level for all $i \leq k$.

In our application to multi-arm clinical trials we have test statistics $Z_1, \dots, Z_I$. Order these to get $X_1 > \cdots > X_I$, corresponding to null hypotheses $H_{0(i)} : \Delta_{(i)} \leq 0,\ i = 1, ..., I$. Letting z_α be the upper $100\alpha\%$ point of the standard normal distribution, reject $H_{0(i)}$ at the $100\alpha\%$ level for all $i \leq j$, if $X_j \geq z_{\alpha/(I-j+1)}$ for any $j = 1, ..., I$. These individual tests can then be combined to assess H_0.

For example, for $I = 2$, $X_1 = \max(Z_1, Z_2)$ and $X_2 = \min(Z_1, Z_2)$. Then

- if $X_2 \geq z_\alpha$, reject $H_{0(1)}$ and $H_{0(2)}$ ($\Rightarrow$ reject H_0);

- if $X_2 < z_\alpha$, but $X_1 \geq z_{\alpha/2}$, reject $H_{0(1)}$, but do not reject $H_{0(2)}$ ($\Rightarrow$ reject H_0);
- if $X_2 < z_\alpha$ and $X_1 < z_{\alpha/2}$, do not reject $H_{0(1)}$ or $H_{0(2)}$ ($\Rightarrow$ do not reject H_0).

Thus the rejection region is

$$\left\{ Z_1 > z_{\alpha/2} \bigcup Z_2 > z_{\alpha/2} \bigcup \left(Z_1 > z_\alpha \bigcap Z_2 > z_\alpha \right) \right\} = A$$

and

$$\begin{aligned} P(A) &= P\left(Z_1 > z_\alpha \bigcap Z_2 > z_\alpha\right) + P\left(Z_1 > z_{\alpha/2} \bigcap Z_2 < z_\alpha\right) \\ &\quad + P\left(Z_2 > z_{\alpha/2} \bigcap Z_1 < z_\alpha\right). \end{aligned}$$

In the case of $I = 2$, Hochberg's procedure is equivalent to those of Simes (1986) and Dunnett and Tamhane (1992) and so our results for this case apply equally to these procedures. Simes (1986) proved that if the tests are independent, the overall size of the test is exactly α. Hochberg's test is based on the assumption that for non-independent tests, this is conservative. Hochberg and Rom (1995) showed that this is true for many cases, but not always, and Shaffer (1995) argued that the use of the Hochberg procedure should be backed up by simulations or theoretical results. Simes (1986) did simulations for many multivariate normal cases, showing conservativeness in each case. Since for $I = 2$ Hochberg's procedure is conservative with correlated multivariate normal data (Simes, 1986), we present the results of a modified Hochberg procedure which uses a nominal significance level, α^*, in order to achieve a size of exactly α.

For $I > 2$, Simes' and Hochberg's procedures differ and each requires its own nominal α^* in order to ensure size α. Although Simes' and Dunnett and Tamhane's procedures are known to be more powerful than Hochberg's for the same nominal level of significance, this is not true when the significance levels are adjusted to give exact size α. Here we decided to pursue just one of these and have used the Hochberg procedure.

2.3.3 Likelihood ratio based test statistics

A more natural and direct approach to evaluating hypotheses of the type considered here is via order restricted testing (Barlow et al., 1972; Robertson et al., 1988; Silvapulle and Sen, 2005), rather than through multiple one-dimensional tests. There has been little application of order restricted inference in medical statistics (Mukerjee et al., 1987).

In a two-arm trial, the likelihood ratio test (LRT) statistic for the one-sided alternative hypothesis $H_{1i} : \Delta_i > 0$ is identical whether the null hypothesis is $H_{0i} : \Delta_i \leq 0$ or

$H^*_{0i} : \Delta_i = 0$, i.e. whether the model allows the possibility of the experimental treatment being worse than the control or not. Therefore, for the test statistics described in Section 2.3.2, which are based on multiple one-dimensional tests, it is natural to use the same procedure whether the null hypothesis is $H_0 : \Delta_i \leq 0 \; \forall i$ or $H^*_0 : \Delta_i = 0 \; \forall i$. However, in the case of several experimental treatments against a control, the likelihood ratio test statistics are different for testing H_0 or H^*_0 against the one-sided alternative H_1.

Likelihood ratio test of equality H^*_0 in order restricted model

Most of the literature (Barlow et al., 1972; Robertson et al., 1988; Silvapulle and Sen, 2005; Peddada et al., 2006; Zhao, 2007) on order restricted inference concentrates on testing the simple null hypothesis $H^*_0 : \boldsymbol{\Delta} = \mathbf{0}$ against the alternative H_1 under the order restricted model, i.e. under the assumption that no experimental treatment can be worse than the control. The LRT of H^*_0 versus H_1 leads to a test statistic given by Barlow et al. (1972), which for three arms takes the form $\{(X_1 - \rho X_2)^{+2}/(1-\rho^2)\} + X_2^{+2}$, where $X_1 = \max(Z_1, Z_2)$ and $X_2 = \min(Z_1, Z_2)$, when $I = 2$ and $A^+ = \max(A, 0)$. Several authors (Mukerjee et al., 1985, 1987; Zhao, 2007) have compared such tests with Dunnett's test. Although it is developed from the wrong model, this test could conceivably be useful for our problem.

This test statistic is decreasing in X_2 for $X_1 > 0,\ X_2 \leq \rho X_1$, and for $X_1 < 0,\ X_2 < \rho^{-1} X_1$ and therefore not monotone increasing in X_2. Consider for example, the cases for $\rho = 0.5$ with $X_1 = 2,\ X_2 = 0$ and $X_1 = 2,\ X_2 = 1$. In the former case, the test statistic takes value $5\frac{1}{3}$ and in the latter case it takes value 4. Thus at some level of significance we would reject H_0 in the first case, but not the second. This is clearly undesirable, since the evidence against H_0 is stronger in the second case.

Related test statistics

Tang and Lin (1997) proposed a test that is similar to the LRT of Robertson but replaces the maximum likelihood estimates with "more easily computed" approximate estimates. This test has similar properties to Robertson's and suffers from the same inconsistency.

Peddada et al. (2006) used a version of T_∞, but with the estimate of the control mean replaced by that of Hwang and Peddada (1994). This replaces the simple tree order with an arbitrary simple order and then uses the isotonic regression estimates obtained from this simple order. Despite seeming very unnatural, Hwang and Peddada (1994) showed that this procedure gives good estimation of μ_0, the control mean. However, the test of Peddada et al. (2006) does not seem to perform better than the LRT for this situation.

Zhao (2007) developed another version of T_∞, but replacing the global estimates of each Δ_i with MLEs obtained under the assumption of a simple tree order. This is appropriate only if a simple tree order can be assumed initially. If used for our problem, this statistic also suffers from the problem of being decreasing over certain values of X_2.

The approaches described in this and the previous subsubsections do not allow for negative Δ_i, which is an inappropriate restriction for our problem, and it is not surprising that it leads to inappropriate test statistics. Note also that the issue of type III errors does not arise in this set up.

Generalized likelihood ratio test for unrestricted model

We saw in the previous subsection that the LRT statistic developed for H_0^* has undesirable properties if used for H_0. Now we show the development of the correct likelihood ratio test statistic under H_0. The likelihood ratio test statistic is given by

$$\lambda = 2 \log \left\{ \frac{L(\mathbf{y}; \hat{\boldsymbol{\Delta}})}{L(\mathbf{y}; \tilde{\boldsymbol{\Delta}})} \right\}, \tag{2.8}$$

where $\boldsymbol{\Delta} = (\Delta_1 \ \cdots \ \Delta_I)'$, the likelihood is given in equation (2.1), $\hat{\boldsymbol{\Delta}}$ is the maximum likelihood estimator (MLE) of $\boldsymbol{\Delta}$ under the model defined in Section 2.1 and $\tilde{\boldsymbol{\Delta}}$ is the restricted maximum likelihood estimator of $\boldsymbol{\Delta}$ under $H_0 : \Delta_i \leq 0, \ \forall i \in \{1, \ldots, I\}$.

The unrestricted MLEs are the ordinary maximum likelihood estimators, e.g. in the case of normal data, the mean of each arm, $\mu_i = \mu_0 + (\Delta_i/\sqrt{N})$, is estimated by its sample mean. The restriction that each experimental arm is no better than the control, which is our null hypothesis, defines a simple tree order (Barlow et al., 1972; Robertson et al., 1988). It is shown by Barlow et al. (1972) that the MLEs are obtained from the isotonic regression for this partial order. Barlow et al. (1972) showed that the isotonic regression estimators $\tilde{\boldsymbol{\Delta}}$ can be obtained from the pool adjacent violators algorithm (PAVA) in the case of normally distributed data. Barlow et al. (1972) gave the description of how to calculate MLEs using PAVA, but did not give any general form of the MLEs.

We now present a theorem which gives a general expression for the MLEs under H_0 in a simple form. This applies not just to the case of normally distributed data Y_{ij}, but for any normally distributed univariate statistics Z_i. We start with some preliminaries.

We define a set of orthogonalized transformed random variables,

$$U_1 = Z_1 \tag{2.9}$$

and

$$U_i = (Z_i - E_i^*) / \sqrt{V_i^*}, \tag{2.10}$$

$i = 2, \ldots, I$, where $E_i^* = E(Z_i | Z_1, \ldots, Z_{i-1})$ and $V_i^* = Var(Z_i | Z_1, \ldots, Z_{i-1})$ are the conditional expectation and variance of Z_i respectively, as given in (2.3) and (2.4). Then $U_1, \ldots, U_I$ are independent normally distributed random variables with unit variance, as shown in section 2.5.1 in the appendix.

The joint likelihood of the U_i is

$$L = \left(\frac{1}{\sqrt{2\pi}}\right)^I \exp\left[-\frac{1}{2}\sum_{i=1}^{I}\{U_i - E(U_i)\}^2\right]. \tag{2.11}$$

Then the log likelihood, ignoring constants, can be written as

$$\log L = -\frac{1}{2\sigma^2}\sum_{i=1}^{I}\frac{g(i)^2}{f(i)},$$

where

$$f(i) = \{1 + (i-2)\rho\}\{1 + (i-1)\rho\}(1-\rho), \tag{2.12}$$

$$g(1) = \sigma(1-\rho)U_1 - (1-\rho)\Delta_1 \tag{2.13}$$

and

$$g(i) = \sigma\sqrt{f(i)}U_i - \{1 + (i-2)\rho\}\Delta_i + \rho\sum_{j=1}^{i-1}\Delta_j, \tag{2.14}$$

for $i \in \{2, \ldots, I\}$.

The unrestricted MLEs are $\hat{\Delta}_i = \sigma Z_i$ $(i = 1, \ldots, I)$, since $E(Z_i) = \Delta_i/\sigma$. Let $\tilde{\Delta}_i$ be the MLEs of the restricted model under the null hypothesis. Assume that $\hat{\Delta}_1 \geq \hat{\Delta}_2 \geq \cdots \geq \hat{\Delta}_I$ (other orders follow by symmetry). The calculations depend on the number of violations there are. If $\hat{\Delta}_1 \leq 0$, there are no violations, while if it is necessary to pool J treatments with the control, there are J violations. We need to calculate $\tilde{\Delta}_i$ for the following cases:

- Case 0: $\hat{\Delta}_i \leq 0 \; \forall i = 1, ..., I$;
- Case J, $J = 1, \ldots, I$: $\hat{\Delta}_1 > 0, \hat{\Delta}_2 > \rho\hat{\Delta}_1, \ldots, \hat{\Delta}_J > \frac{\rho\sum_{j=1}^{J-1}\hat{\Delta}_j}{1+(J-2)\rho}$, but $\hat{\Delta}_{J+1} \leq \frac{\rho\sum_{j=1}^{J}\hat{\Delta}_j}{1+(J-1)\rho}$.

For Case 0, the restricted and unrestricted MLEs are the same, i.e. $\tilde{\Delta}_i = \hat{\Delta}_i \; \forall i$. For Case J we show the form of the restricted MLEs by using the following results from Thompson (1962), restated in our notation.

Lemma 1 (Thompson, 1962) *If* $\log L(\mathbf{\Delta})$ *is differentiable at* $\tilde{\mathbf{\Delta}}$ *then a necessary condition for* $\tilde{\mathbf{\Delta}}$ *to maximize* $\log L(\mathbf{\Delta})$*, subject to* $\Delta_i \leq 0,\ i = 1, \dots, I$*, is that for each* $i \in \{1, \dots, I\}$ *, either:*

(a) $\tilde{\Delta}_i = 0$ *and* $\frac{\partial \log L(\mathbf{\Delta})}{\partial \Delta_i}\Big|_{\mathbf{\Delta}=\tilde{\mathbf{\Delta}}} \geq 0$*; or*

(b) $\tilde{\Delta}_i < 0$ *and* $\frac{\partial \log L(\mathbf{\Delta})}{\partial \Delta_i}\Big|_{\mathbf{\Delta}=\tilde{\mathbf{\Delta}}} = 0.$

Corollary 2 (Thompson, 1962) *If* $\log L(\mathbf{\Delta})$ *is strictly concave, then* $\tilde{\mathbf{\Delta}}$ *is a unique maximum.*

Thus, since our log-likelihood is strictly concave (multivariate normal), the restricted MLEs under H_0 will be the unique values $\tilde{\mathbf{\Delta}}$ which satisfy all conditions (a) and (b) in Lemma 1. These are given by the following theorem.

Theorem 3 *Assume that the unrestricted MLEs* $\hat{\Delta}_i$ *are such that* $\hat{\Delta}_1 \geq \hat{\Delta}_2 \geq \cdots \geq \hat{\Delta}_I$ *(other orders follow by symmetry). If, for some* $1 \leq J \leq I$*,*

$$\hat{\Delta}_1 > 0, \hat{\Delta}_2 > \rho\hat{\Delta}_1, \dots, \hat{\Delta}_J > \frac{\rho}{1 + (J-2)\rho} \sum_{j=1}^{J-1} \hat{\Delta}_j,$$

but

$$\hat{\Delta}_{J+1} \leq \frac{\rho}{1 + (J-1)\rho} \sum_{j=1}^{J} \hat{\Delta}_j,$$

then the order restricted MLEs $\tilde{\Delta}_i$ *are given by*

$$\tilde{\Delta}_i = \begin{cases} 0, & i = 1, \dots, J; \\ \sigma\left\{Z_i - \frac{\rho}{1+(J-1)\rho} \sum_{j=1}^{J} Z_j\right\}, & i = J+1, \dots, I. \end{cases} \tag{2.15}$$

The proof is given in Section 2.5.2 in the appendix. Most of this is done by simple but tedious algebraic manipulation. We need to show that (2.15) provide a solution to (a) $\frac{\partial \log L}{\partial \Delta_k} = 0$, $k = J+1, \dots, I$, and then show that $\tilde{\Delta}_i$ in (2.15) give (b) $\frac{\partial \log L}{\partial \Delta_k}\Big|_{\Delta_k = \tilde{\Delta}_k} \geq 0$ for $k = 1, \dots, J$. We need to show (a) for the cases: (i) $k = J+1$ (Lemma 5); and (ii) $k \geq J+2$ (Lemma 6). Similarly we show (b) for the cases: (i) $k = J$ (Lemma 7); (ii) $k = J-1$ (Lemma 8); and (iii) $k \leq J-2$ (Lemma 10). We present this series of lemmas in Section 2.5.2 in the appendix.

Mukerjee et al. (1985) and Robertson et al. (1988) worked out the likelihood ratio test statistics for our H_0 against H_1, in the case of normally distributed data, and studied some of its properties. This is extended to any univariate normal statistics Z_i and related to our family of test statistics in the following theorem.

Theorem 4 *The likelihood ratio test statistic of H_0 versus H_1 can be written as $\lambda = T_2^2$, where*

$$T_2 = \sqrt{X_1^{+2} + \sum_{j=2}^{I} \frac{(X_j - E_j)^{+2}}{V_j}},$$

as defined in equation (2.7).

This Theorem is proved in Section 2.5.3 in the appendix by substitution of $\hat{\boldsymbol{\Delta}} = \sigma\mathbf{Z}$ and $\tilde{\boldsymbol{\Delta}}$, as defined in (2.15), into (2.8). We also carried out extensive numerical checks to check that the result is correct.

For two arms $E_2 = \rho X_1$, $V_i = 1 - \rho^2$ and the test statistic simplifies to

$$T_2 = \sqrt{X_1^{+2} + \frac{(X_2 - \rho X_1)^{+2}}{1 - \rho^2}}.$$

Mukerjee et al. (1985) and Robertson et al. (1988) mainly used this test, in a different form and with signs reversed, for checking the assumption of the order restriction before testing H_0^* against H_1 under that assumption. It was also described in a general form by Silvapulle and Sen (2005). There seems to have been little or no direct use of this LRT to test our hypothesis of interest and none of these authors dealt with the problem of type-III errors or considered the corrected size defined in Section 2.2. Note the difference between this statistic and the LRT for H_0^* versus H_1 given at the begining of this section. The difference is that X_1 and X_2 have been interchanged. Whereas previously the statistic was not monotone in X_2, T_2 is monotone in X_2.

A simpler test statistic

For $\rho = 0$, e.g. if separate trials were conducted with separate control groups for each experimental treatment, T_2 reduces to $T_2^0 = \sqrt{\sum_{i=1}^{I} Z_i^{+2}}$. This is another possible test statistic which we will study, since it might be that ignoring the correlation does little harm and this test statistic is easier to compute than T_2. T_2^0 can also be considered as a simple modification of the χ^2 test statistic S_2^0 for the two-sided alternative hypothesis, with differences in the negative direction replaced by zero. We do not consider S_2^0, since it is clearly inappropriate to give so much weight to large negative Z_is.

2.3.4 Single contrast and related tests

Single contrast test

When all of the Δ_i are known to be equal, then the labelling of experimental arms is irrelevant and, in the case of normally distributed observations, a test based on the single contrast $\left(\sum_{i=1}^{I} \bar{Y}_i/I\right) - \bar{Y}_0$ is most powerful. Here we consider the test statistic

$$S_1^0 = \frac{I\sqrt{N}}{\sigma}\left(\sum_{i=1}^{I}\frac{\bar{Y}_i}{I} - \bar{Y}_0\right) = \sum_{i=1}^{I} Z_i,$$

which is a special case of a test suggested by Abelson and Tukey (1963) and Schaafsma and Smid (1966) for testing H_0^* against H_1, when prior knowledge is available about the relative sizes of the alternatives. This is sometimes known as the Abelson-Tukey-Schaafsma-Smid contrast test, but we will continue to refer to it as S_1^0. However, this test is not even consistent against all outcomes in H_1 and a large negative value of one mean can cancel the positive contribution from a favorable treatment.

Mukerjee et al. (1987) suggested a family of test statistics for the simple null hypothesis H_0^*, which can be expressed as a weighted average of T_∞ and S_1^0, with the weights chosen to make the two parts orthogonal. Their main motivation was to obtain a test statistic whose null distribution could be obtained analytically and they noted that "the LRT seems very difficult to beat provided ... that its null ... distribution can be ... reasonably approximated."

A modified test

A modification of S_1^0, which avoids the inconsistency, is obtained by ignoring negative contributions. We define

$$T_1^0 = \sum_{i=1}^{I} Z_i^+.$$

The simplicity of this test statistic is its main attraction and we study it, along with others, in the next chapter.

A further refinement

The above test statistic can be further refined by giving more weight to X_1 to give

$$T_1 = X_1^+ + \sum_{j=2}^{I}\frac{(X_j - E_j)^+}{\sqrt{V_j}},$$

where E_j and V_j are defined in (2.5) and (2.6) respectively. For two arms this simplifies to

$$T_1 = X_1^+ + \frac{(X_2 - \rho X_1)^+}{\sqrt{1-\rho^2}}.$$

It is easy to see that this test is consistent in the sense that H_0 will be rejected with probability 1 as some $\Delta_i \to \infty$, regardless of the values of other Δ_j.

2.4 Conclusions

In this chapter, we have defined several test statistics, which could be used to test $H_0 : \Delta_i \leq 0 \ \forall i$ against $H_1 : \Delta_i > 0$ for at least one i. It is not known which is best and it is clear that there is no uniformly most powerful test statistic. It is therefore necessary to compare the properties of these test statistics, such as their power, and this is done in the next chapter.

The extension of these tests to non-inferiority trials is conceptually straightforward. A typical scenario might be that there are several experimental treatments, one of which could potentially replace, or compete with, the standard treatment if it can be shown to be not inferior in terms of efficacy. The argument in favour of one-sided non-inferiority trials over equivalence trials is similar to the argument we have made above in favour of one-sided trials, i.e. an experimental treatment should not be rejected for being superior to the control. In a non-inferiority trial, the null hypothesis becomes $H_0^\dagger : \Delta_i \leq -\delta^\dagger \forall i$, where $\delta^\dagger$ is the non-inferiority margin, and the alternative is $H_1^\dagger : \Delta_i > -\delta^\dagger$ for at least one i. All of the test statistics defined above can be used, with Z_i replaced by $Z_i + \delta^\dagger$ in their definitions. Hence all of the methods described here can be used immediately, although we will not emphasise this potential application in this thesis.

2.5 Appendix: proofs

2.5.1 Preliminaries

We defined the orthogonalized transformed random variables, $U_1, \ldots, U_I$, in equations (2.9) and (2.10). Substituting (2.3) and (2.4) into these equations, we obtain

$$U_i = \frac{\{1 + (i-2)\rho\} Z_i - \rho \sum_{j=1}^{i-1} Z_j}{\sqrt{\{1+(i-2)\rho\}\{1+(i-1)\rho\}(1-\rho)}}, \tag{2.16}$$

where, throughout this appendix, $\sum_{j=1}^{0} a(j)$ is defined to be zero, no matter what function $a(\cdot)$ is used. Then

$$E(U_i) = \frac{\{1+(i-2)\rho\}\,\Delta_i - \rho\sum_{j=1}^{i-1}\Delta_j}{\sigma\sqrt{\{1+(i-2)\rho\}\,\{1+(i-1)\rho\}\,(1-\rho)}}.$$

The variances are $Var(U_1) = Var(Z_1) = 1$ and, for $i = 2, \ldots, I$,

$$\begin{aligned}
Var(U_i) &= \frac{Var(Z_i) + Var(E_i^*) - 2Cov(Z_i, E_i^*)}{V_i^*} \\
&= \frac{1}{V_i^*}\left[1 + \frac{\rho^2}{\{1+(i-2)\,\rho\}^2}\left\{\sum_{j=1}^{i-1} Var(Z_j) + 2\sum_{j=1}^{i-2}\sum_{k=j+1}^{i-1} Cov(Z_j, Z_k)\right\}\right. \\
&\quad \left. -2\frac{\rho}{1+(i-2)\rho}\sum_{j=1}^{i-1} Cov(Z_i, Z_j)\right] \\
&= \frac{1}{V_i^*}\left[1 + \frac{\rho^2}{\{1+(i-2)\,\rho\}^2}\left\{i-1+2\frac{(i-1)\,(i-2)}{2}\rho\right\} - \frac{2\,(i-1)\,\rho^2}{1+(i-2)\rho}\right] \\
&= \frac{1+(i-2)\,\rho}{\{1+(i-1)\,\rho\}\,(1-\rho)}\,\frac{1+(i-2)\,\rho-(i-1)\,\rho^2-2\,(i-1)\,\rho^2}{1+(i-2)\,\rho} \\
&= 1.
\end{aligned}$$

We obtain covariances as follows. For $i = 2, \ldots, I$

$$\begin{aligned}
Cov(U_1, U_i) &= Cov\left(Z_1, \frac{Z_i - \frac{\rho}{1+(i-2)\rho}\sum_{j=1}^{i-1} Z_j}{V_i^*}\right) \\
&= \frac{1}{V_i^*}\rho - \frac{\rho}{V_i^*\,\{1+(i-2)\,\rho\}}\,\{1+(i-2)\,\rho\} \\
&= 0.
\end{aligned}$$

For $1 < i < j \leq I$,

$$\begin{aligned}
Cov(U_i, U_j) &= Cov\left(\frac{Z_i - \frac{\rho}{1+(i-2)\rho}\sum_{k=1}^{i-1} Z_k}{V_i^*}, \frac{Z_j - \frac{\rho}{1+(j-2)\rho}\sum_{k=1}^{j-1} Z_k}{V_j^*}\right) \\
&= \frac{1}{V_i^* V_j^*}\left(\rho - \frac{\rho\,(i-1)\,\rho}{1+(i-2)\,\rho} - \frac{\rho\,\{1+(j-2)\,\rho\}}{1+(j-2)\,\rho}\right. \\
&\quad \left. + \frac{\rho^2\,[(i-1)+\{(i-1)\,(j-1)-(i-1)\}\,\rho]}{\{1+(i-2)\,\rho\}\,\{1+(j-2)\,\rho\}}\right) \\
&= \frac{1}{V_i^* V_j^*}\left[\frac{\rho^2\,(i-1)\,\{1+(j-2)\,\rho\}}{\{1+(i-2)\,\rho\}\,\{1+(j-2)\,\rho\}} - \frac{\rho^2\,(i-1)}{1+(i-2)\,\rho}\right] \\
&= 0.
\end{aligned}$$

Thus $U_1 \ldots U_I$ are independent normal variables with unit variance.

Substituting the values of U_i from (2.16) into the joint likelihood of the U_i given in (2.11), taking $\log L$ and ignoring the constant term we have

$$\begin{aligned}\log L &= -\frac{1}{2\sigma^2}\left((\sigma U_1 - \Delta_1)^2 + \right.\\ &\left.\sum_{i=2}^{I} \frac{\left[\sigma\sqrt{\{1+(i-2)\rho\}\{1+(i-1)\rho\}(1-\rho)}U_i - \{1+(i-2)\rho\}\Delta_i + \rho\sum_{j=1}^{i-1}\Delta_j\right]^2}{\{1+(i-2)\rho\}\{1+(i-1)\rho\}(1-\rho)}\right).\\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^{I}\frac{g(i)^2}{f(i)}, \qquad (2.17)\end{aligned}$$

where $f(i)$ and $g(i)$ are defined in (2.12), (2.13) and (2.14).

Differentiating, we have

$$\frac{\partial \log L}{\partial \Delta_k} = -\frac{1}{2\sigma^2}\sum_{i=1}^{I}\frac{\frac{\partial g(i)^2}{\partial \Delta_k}}{f(i)}. \qquad (2.18)$$

Since

$$\frac{\partial}{\partial \Delta_k}\left\{g(i)^2\right\} = 2g(i)\frac{\partial g(i)}{\partial \Delta_k}$$

and

$$\frac{\partial g(i)}{\partial \Delta_k} = -\{1+(i-2)\rho\}\frac{\partial \Delta_i}{\partial \Delta_k} + \rho\sum_{j=1}^{i-1}\frac{\partial \Delta_j}{\partial \Delta_k},$$

we have

$$\frac{\partial g(i)}{\partial \Delta_k} = \begin{cases} \rho & \text{if } k < i; \\ -\{1+(i-2)\rho\} & \text{if } k = i; \\ 0 & \text{if } k > i. \end{cases}$$

Substituting $g(i)$ from (2.14), we get

$$\frac{\partial\left\{g(i)^2\right\}}{\partial \Delta_k} = \begin{cases} 2\rho\left[\sigma\sqrt{f(i)}U_i - \{1+(i-2)\rho\}\Delta_i + \rho\sum_{j=1}^{i-1}\Delta_j\right] & \text{if } k < i; \\ -2\{1+(i-2)\rho\}\left[\sigma\sqrt{f(i)}U_i - \{1+(i-2)\rho\}\Delta_i + \rho\sum_{j=1}^{i-1}\Delta_j\right] & \text{if } k = i; \\ 0 & \text{if } k > i. \end{cases}$$

Substituting U_i from (2.10) and simplifying, we obtain

$$\frac{\partial\left\{g(i)^2\right\}}{\partial \Delta_k} = \begin{cases} 2\rho\left(\sigma\left[\{1+(i-2)\rho\}Z_i - \rho\sum_{j=1}^{i-1}Z_j\right]\right. & \\ \quad\left. - \{1+(i-2)\rho\}\Delta_i + \rho\sum_{j=1}^{i-1}\Delta_j\right) & \text{if } k < i; \\ -2\{1+(i-2)\rho\}\left(\sigma\left[\{1+(i-2)\rho\}Z_i - \rho\sum_{j=1}^{i-1}Z_j\right]\right. & \\ \quad\left. - \{1+(i-2)\rho\}\Delta_i + \rho\sum_{j=1}^{i-1}\Delta_j\right) & \text{if } k = i; \\ 0 & \text{if } k > i. \end{cases}$$

Splitting the summation over i in (2.18) into the cases $i < k$, $i = k$ and $i > k$ gives

$$\frac{\partial \log L}{\delta \Delta_k} = -\frac{1}{2\sigma^2}\left[\sum_{i=1}^{k-1}\frac{\partial\left\{g(i)^2\right\}/\partial \Delta_k}{f(i)} + \frac{\partial\left\{g(k)^2\right\}/\partial \Delta_k}{f(k)} + \sum_{i=k+1}^{I}\frac{\partial\left\{g(i)^2\right\}/\partial \Delta_k}{f(i)}\right].$$

Since, for $i < k$, $\partial\left\{g(i)^2\right\}/\partial\Delta_k = 0$, it follows that

$$\begin{aligned}\frac{\partial \log L}{\partial \Delta_k} &= -\frac{1}{2\sigma^2}\left[\frac{\partial\left\{g(i)^2\right\}/\partial\Delta_k}{f(k)} + \sum_{i=k+1}^{I}\frac{\partial\left\{g(i)^2\right\}/\partial\Delta_k}{f(i)}\right] \\ &= -\frac{1}{\sigma^2}\left(-\frac{\{1+(k-2)\rho\}}{f(k)}\left[\{1+(k-2)\rho\}(\sigma Z_k - \Delta_k) - \rho\left(\sigma\sum_{j=1}^{k-1} Z_j - \sum_{j=1}^{k-1}\Delta_j\right)\right]\right. \\ &\quad \left. +\rho\sum_{i=k+1}^{I}\frac{1}{f(i)}\left[\{1+(i-2)\rho\}(\sigma Z_i - \Delta_i) - \rho\sum_{j=1}^{i-1}(\sigma Z_j - \Delta_j)\right]\right). \qquad (2.19)\end{aligned}$$

This form will be used in the proof of Theorem 3 in Section 2.5.2.

2.5.2 Proof of Theorem 3

We show that

$$\tilde{\Delta}_s = \begin{cases} 0 & \text{if } s = 1, \ldots, J; \\ \sigma\left\{Z_s - \frac{\rho}{1+(J-1)\rho}\sum_{j=1}^{J} Z_j\right\} & \text{if } s = J+1, \ldots, I \end{cases} \qquad (2.20)$$

provide a solution to (a) $\frac{\partial \log L}{\partial \Delta_k} = 0$, $k = J+1, \ldots, I$, and then show that $\tilde{\Delta}_s$ give (b) $\frac{\partial \log L}{\partial \Delta_k} \geq 0$ for $k = 1, \ldots, J$. We need to show (a) for the cases: (i) $k = J+1$; and (ii) $k \geq J+2$. Similarly we show (b) for the cases: (i) $k = J$; (ii) $k = J-1$; and (iii) $k \leq J-2$. We present these as a series of lemmas.

Lemma 5
$$\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} = 0,$$
for $k = J+1$.

Proof. Replacing k by $J+1$ in (2.19), we get

$$\begin{aligned}\frac{\partial \log L}{\partial \Delta_k} &= -\frac{1}{\sigma^2}\left(-\frac{\{1+(J-1)\rho\}}{f(J+1)}\left[\{1+(J-1)\rho\}(\sigma Z_{J+1} - \Delta_{J+1}) - \rho\sum_{j=1}^{J}(\sigma Z_j - \Delta_j)\right]\right. \\ &\quad \left. +\rho\sum_{i=J+2}^{I}\frac{1}{f(i)}\left[\{1+(i-2)\rho\}(\sigma Z_i - \Delta_i) - \rho\sum_{j=1}^{i-1}(\sigma Z_j - \Delta_j)\right]\right).\end{aligned}$$

Evaluating at $\boldsymbol{\Delta} = \tilde{\boldsymbol{\Delta}}$, multiplying out the summations in j and substituting $\sum_{j=1}^{J}\tilde{\Delta}_j = 0$, we obtain

$$\begin{aligned}\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} &= -\frac{1}{\sigma^2}\left(-\frac{\{1+(J-1)\rho\}}{f(J+1)}\left[\{1+(J-1)\rho\}\left(\sigma Z_{J+1} - \tilde{\Delta}_{J+1}\right) - \rho\sum_{j=1}^{J}\sigma Z_j\right]\right. \\ &\quad \left. +\rho\sum_{i=J+2}^{I}\frac{1}{f(i)}\left[\{1+(i-2)\rho\}\left(\sigma Z_i - \tilde{\Delta}_i\right) - \rho\sum_{j=1}^{i-1}\sigma Z_j + \rho\sum_{j=1}^{i-1}\tilde{\Delta}_j\right]\right).\end{aligned}$$

Splitting up the summations over $j = 1, \ldots, i-1$ into $\sum_{j=1}^{J}$ plus $\sum_{j=J+1}^{i-1}$, where $\sum_{j=1}^{J} \tilde{\Delta}_j = 0$, gives

$$\begin{aligned}
\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\mathbf{\Delta}=\tilde{\mathbf{\Delta}}} &= -\frac{1}{\sigma^2}\left(-\frac{\{1+(J-1)\rho\}}{f(J+1)}\left[\{1+(J-1)\rho\}\left(\sigma Z_{J+1}-\tilde{\Delta}_{J+1}\right)-\rho\sum_{j=1}^{J}\sigma Z_j\right]\right. \\
&\quad +\rho\sum_{i=J+2}^{I}\frac{1}{f(i)}\left[\{1+(i-2)\rho\}\left(\sigma Z_i-\tilde{\Delta}_i\right)-\rho\sum_{j=1}^{J}\sigma Z_j\right. \\
&\quad \left.\left.-\rho\sum_{j=J+1}^{i-1}\sigma Z_j+\rho\sum_{j=J+1}^{i-1}\tilde{\Delta}_j\right]\right).
\end{aligned}$$

Substituting $\tilde{\Delta}_j$, for $j = J+1, \ldots, i-1$, from (2.20), we obtain

$$\begin{aligned}
\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\mathbf{\Delta}=\tilde{\mathbf{\Delta}}} &= -\frac{1}{\sigma^2}\left(-\frac{1+(J-1)}{f(J+1)}\right. \\
&\quad \left[\{1+(J-1)\rho\}\sigma\left\{Z_{J+1}-Z_{J+1}+\frac{\rho}{1+(J-1)\rho}\sum_{j=1}^{J}Z_j\right\}-\rho\sigma\sum_{j=1}^{J}Z_j\right] \\
&\quad +\rho\sum_{i=J+2}^{I}\frac{1}{f(i)}\left[\{1+(i-2)\rho\}\sigma\left\{Z_i-Z_i+\frac{\rho}{1+(J-1)\rho}\sum_{j=1}^{J}Z_j\right\}\right. \\
&\quad \left.\left.-\rho\sigma\sum_{j=1}^{J}Z_j-\rho\sigma\sum_{j=J+1}^{i-1}Z_j+\rho\sigma\sum_{j=J+1}^{i-1}\left\{Z_j-\frac{\rho}{1+(J-1)\rho}\sum_{l=1}^{J}Z_l\right\}\right]\right).
\end{aligned}$$

Cancelling out the terms inside the first square brackets and cancelling Z_is we get

$$\begin{aligned}
\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\mathbf{\Delta}=\tilde{\mathbf{\Delta}}} &= -\frac{1}{\sigma^2}\left(\rho\sum_{i=J+2}^{I}\frac{1}{f(i)}\left[\frac{\{1+(i-2)\rho\}\rho\sigma}{1+(J-1)\rho}\sum_{j=1}^{J}Z_j-\rho\sigma\sum_{j=1}^{J}Z_j\right.\right. \\
&\quad \left.\left.-\rho\sigma\sum_{j=J+1}^{i-1}Z_j+\rho\sigma\sum_{j=J+1}^{i-1}\left\{Z_j-\frac{\rho}{1+(J-1)\rho}\sum_{l=1}^{J}Z_l\right\}\right]\right) \\
&= -\frac{\rho\sigma}{\sigma^2}\left[\rho\sum_{i=J+2}^{I}\frac{1}{f(i)}\left\{\frac{1+(i-2)\rho}{1+(J-1)\rho}\sum_{j=1}^{J}Z_j-\sum_{j=1}^{J}Z_j\right.\right. \\
&\quad \left.\left.-\sum_{j=J+1}^{i-1}\frac{\rho}{1+(J-1)\rho}\sum_{l=1}^{J}Z_l\right\}\right].
\end{aligned}$$

Since $\sum_{j=J+1}^{i-1}\left(\sum_{l=1}^{J} Z_l\right) = (i-1-J)\sum_{l=1}^{J} Z_l$, we get

$$\begin{aligned}
\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} &= -\frac{\rho^2}{\sigma}\left[\sum_{i=J+2}^{I}\frac{1}{f(i)}\left\{\frac{1+(i-2)\,\rho}{1+(J-1)\,\rho}\sum_{j=1}^{J} Z_j - \sum_{j=1}^{J} Z_j\right.\right. \\
&\qquad \left.\left.-\frac{(i-1-J)\,\rho}{1+(J-1)\,\rho}\sum_{j=1}^{J} Z_j\right\}\right]. \\
&= -\frac{\rho^2}{\sigma}\sum_{j=1}^{J} Z_j\left[\sum_{i=J+2}^{I}\frac{1}{f(i)}\left\{\frac{(i-1-J)\,\rho-(i-1-J)\,\rho}{1+(J-1)\,\rho}\right\}\right] \\
&= 0.
\end{aligned}$$

■

Lemma 6

$$\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} = 0,$$

for $k \geq J+2$.

Proof. In (2.19), splitting the summations

$$\sum_{j=1}^{k-1}(\sigma Z_j - \Delta_j) = \sum_{j=1}^{J}(\sigma Z_j - \Delta_j) + \sum_{j=J+1}^{k-1}(\sigma Z_j - \Delta_j)$$

and

$$\sum_{j=1}^{i-1}(\sigma Z_j - \Delta_j) = \sum_{j=1}^{J}(\sigma Z_j - \Delta_j) + \sum_{j=J+1}^{i-1}(\sigma Z_j - \Delta_j),$$

we have

$$\begin{aligned}
\frac{\partial \log L}{\partial \Delta_k} &= -\frac{1}{\sigma^2}\left(-\frac{1+(k-2)\,\rho}{f(k)}\left[\{1+(k-2)\,\rho\}(\sigma Z_k-\Delta_k) - \rho\sum_{j=1}^{J}(\sigma Z_j-\Delta_j)\right.\right. \\
&\qquad \left. -\rho\sum_{j=J+1}^{k-1}(\sigma Z_j-\Delta_j)\right] \\
&\qquad +\rho\sum_{i=k+1}^{I}\frac{1}{f(i)}\left[\{1+(i-2)\,\rho\}(\sigma Z_i-\Delta_i) - \rho\sum_{j=1}^{J}(\sigma Z_j-\Delta_j)\right. \\
&\qquad \left.\left. -\rho\sum_{j=J+1}^{i-1}(\sigma Z_j-\Delta_j)\right]\right).
\end{aligned}$$

Substituting $\tilde{\Delta}_s$, multiplying out the brackets with $\sum_{j=1}^{J}\left(\sigma Z_j - \Delta_j\right)$ and substituting $\sum_{j=1}^{J}\tilde{\Delta}_j = 0$, we have

$$\begin{aligned}
\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} &= -\frac{1}{\sigma^2}\Bigg(-\frac{1+(k-2)\,\rho}{f(k)}\Bigg[\{1+(k-2)\,\rho\}\left\{\sigma Z_k - \sigma Z_k + \frac{\sigma\rho}{1+(J-1)\,\rho}\sum_{j=1}^{J} Z_j\right\} \\
&\quad -\rho\sigma\sum_{j=1}^{J} Z_j - \rho\sum_{j=J+1}^{k-1}\left\{\sigma Z_j - \sigma Z_j + \frac{\rho\sigma}{1+(J-1)\,\rho}\sum_{l=1}^{J} Z_l\right\}\Bigg] \\
&\quad +\rho\sum_{i=k+1}^{I}\frac{1}{f(i)}\Bigg[\{1+(i-2)\,\rho\}\left\{\sigma Z_i - \sigma Z_i + \frac{\rho\sigma}{1+(J-1)\,\rho}\sum_{j=1}^{J} Z_j\right\} \\
&\quad -\rho\sigma\sum_{j=1}^{J} Z_j - \rho\sum_{j=J+1}^{i-1}\left\{\sigma Z_j - \sigma Z_j - \frac{\rho\sigma}{1+(J-1)\,\rho}\sum_{l=1}^{J} Z_l\right\}\Bigg]\Bigg).
\end{aligned}$$

Since $\sum_{j=J+1}^{i-1}\left(\sum_{l=1}^{J} Z_l\right) = (i-1-J)\sum_{l=1}^{J} Z_l$, we obtain

$$\begin{aligned}
\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} &= -\frac{\rho}{\sigma}\Bigg[-\frac{1+(k-2)\,\rho}{f(k)}\left\{\frac{1+(k-2)\,\rho}{1+(J-1)\,\rho}\sum_{j=1}^{J} Z_j - \sum_{j=1}^{J} Z_j - \frac{\rho\,(k-1-J)}{1+(J-1)\,\rho}\sum_{l=1}^{J} Z_l\right\} \\
&\quad +\rho\sum_{i=k+1}^{I}\frac{1}{f(i)}\left\{\frac{1+(i-2)\,\rho}{1+(J-1)\,\rho}\sum_{j=1}^{J} Z_j - \sum_{j=1}^{J} Z_j - \frac{\rho\,(i-1-J)}{1+(J-1)\,\rho}\sum_{l=1}^{J} Z_l\right\}\Bigg] \\
&= -\frac{\rho}{\{1+(J-1)\,\rho\}\,\sigma}\Bigg(-\frac{\{1+(k-2)\,\rho\}\sum_{j=1}^{J} Z_j}{f(k)}\,[\{1+(k-2)\,\rho\} \\
&\quad -\{1+(J-1)\,\rho\} - \rho\,(k-1-J)] + \rho\sum_{i=k+1}^{I}\frac{\sum_{j=1}^{J} Z_j}{f(i)}\,[\{1+(i-2)\,\rho\} \\
&\quad -\{1+(J-1)\,\rho\} - \rho\,(i-1-J)]\Bigg) \\
&= 0.
\end{aligned}$$

■

Lemma 7

$$\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} \geq 0$$

for $k = J$.

Proof. Replacing $k = J$ and substituting $\tilde{\Delta}_s$ in (2.19) we have

$$\begin{aligned}
\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} &= -\frac{1}{\sigma^2}\Bigg\{-\frac{1+(J-2)\,\rho}{f(J)}\left[\{1+(J-2)\,\rho\}\,(\sigma Z_J - 0) - \rho\sigma\left(\sum_{j=1}^{J} Z_j - 0\right)\right] \\
&\quad +\rho\sum_{i=J+1}^{I}\frac{1}{f(i)}\Bigg(\{1+(i-2)\,\rho\}\left[\sigma Z_i - \sigma\left\{Z_i - \frac{\rho}{1+(J-1)\,\rho}\sum_{j=1}^{J} Z_j\right\}\right] \\
&\quad -\rho\sum_{j=1}^{i-1}\left(\sigma Z_j - \Delta_j\right)\Bigg)\Bigg\}.
\end{aligned}$$

Splitting up the summation of $\sigma Z_j - \Delta_j$ over $j = 1, \ldots, i-1$ into $\sum_{j=1}^{J}$ plus $\sum_{j=J+1}^{i-1}$ in the last brackets, where $\sum_{j=1}^{J} \tilde{\Delta}_j = 0$, gives

$$\begin{aligned}
\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} &= -\frac{1}{\sigma^2}\left(-\frac{1+(J-2)\rho}{f(J)}\left[\{1+(J-2)\rho\}\sigma Z_J - \rho\sigma\sum_{j=1}^{J} Z_j\right]\right. \\
&\quad +\rho\sum_{i=J+1}^{I}\frac{1}{f(i)}\left[\frac{\{1+(i-2)\rho\}\sigma\rho}{1+(J-1)\rho}\sum_{j=1}^{J} Z_j - \rho\sigma\sum_{j=1}^{J} Z_j\right. \\
&\quad \left.\left. -\rho\sum_{j=J+1}^{i-1}\left\{\frac{\rho\sigma}{1+(J-1)\rho}\sum_{k=1}^{J} Z_k\right\}\right]\right).
\end{aligned}$$

Substituting $\sum_{j=J+1}^{i-1}\left(\sum_{k=1}^{J} Z_k\right) = (i-1-J)\sum_{j=1}^{J} Z_j$, we obtain

$$\begin{aligned}
\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} &= -\frac{1}{\sigma^2}\left(-\frac{1+(J-2)\rho}{f(J)}\left[\{1+(J-2)\rho\}\sigma Z_J - \rho\sigma\sum_{j=1}^{J} Z_j\right]\right. \\
&\quad +\rho\sum_{i=J+1}^{I}\frac{1}{f(i)}\left[\frac{\{1+(i-2)\rho\}\sigma\rho}{1+(J-1)\rho}\sum_{j=1}^{J} Z_j - \rho\sigma\sum_{j=1}^{J} Z_j\right. \\
&\quad \left.\left. -\frac{\rho^2\sigma(i-1-J)}{1+(J-1)\rho}\sum_{j=1}^{J} Z_j\right]\right).
\end{aligned}$$

Taking $\sum_{j=1}^{J} Z_j$ as a common factor and simplifying, we obtain

$$\begin{aligned}
\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} &= -\frac{1}{\sigma^2}\left[-\frac{1+(J-2)\rho}{f(J)}\left\{\sigma Z_J(1+J\rho-2\rho+\rho) - \rho\sigma\sum_{j=1}^{J} Z_j\right\}\right. \\
&\quad \left. +\rho^2\sigma\sum_{j=1}^{J} Z_j\left\{\sum_{i=J+1}^{I}\frac{1}{f(i)}\frac{1+(i-2)\rho-1-(J-1)\rho-(i-1-J)\rho}{1+(J-1)\rho}\right\}\right] \\
&= \frac{1}{\sigma^2}\left(\frac{1+(J-2)\rho}{f(J)}\left[\sigma Z_J\{1+(J-1)\rho\} - \rho\sigma\sum_{j=1}^{J} Z_j\right]\right).
\end{aligned}$$

Replacing σZ_k with $\hat{\Delta}_k$,

$$\begin{aligned}
\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} &= \frac{1+(J-2)\rho}{\sigma^2 f(J)}\left[\hat{\Delta}_J\{1+(J-1)\rho\} - \rho\sum_{j=1}^{J-1}\hat{\Delta}_j\right] \\
&> \frac{1+(J-2)\rho}{\sigma^2 f(J)}\left[\hat{\Delta}_J\{1+(J-1)\rho\} - \hat{\Delta}_J\right],
\end{aligned}$$

since

$$\begin{aligned}
\hat{\Delta}_J &> \frac{\rho}{1+(J-2)\rho}\sum_{j=1}^{J-1}\hat{\Delta}_j \\
\Rightarrow -\rho\sum_{j=1}^{J-1}\hat{\Delta}_j &> -\{1+(J-2)\rho\}\hat{\Delta}_J.
\end{aligned}$$

Hence,

$$\frac{\partial \log L}{\partial \Delta_k} > \frac{1+(J-2)\,\rho}{\sigma^2 f(J)}\hat{\Delta}_J \rho \geq 0.$$

■

Lemma 8

$$\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} \geq 0,$$

for $k = J-1$.

Proof. Replacing k by $J-1$ in (2.19), the Δ_k and Δ_j in the first square bracket become zero, the summation before the second square bracket is over $i = J, \ldots, I$ and, splitting it into 3 parts for $i = J$, $i = J+1$ and $i = J+2, \ldots, I$, we have

$$\begin{aligned}
\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} &= -\frac{1}{\sigma^2}\left(-\frac{1+(J-3)\,\rho}{f(J-1)}\left[\{1+(J-3)\,\rho\}\,\sigma Z_{J-1} - \rho\sigma\sum_{j=1}^{J-2} Z_j\right]\right. \\
&+\rho\frac{1}{f(J)}\left[\{1+(J-2)\,\rho\}\,(\sigma Z_J - \Delta_J) - \rho\sum_{j=1}^{J-1}(\sigma Z_j - \Delta_j)\right] \\
&+\rho\frac{1}{f(J+1)}\left[\{1+(J-1)\,\rho\}\,(\sigma Z_{J+1} - \Delta_{J+1}) - \rho\sum_{j=1}^{J}(\sigma Z_j - \Delta_j)\right] \\
&\left.+\rho\sum_{i=J+2}^{I}\frac{1}{f(i)}\left[\{1+(i-2)\,\rho\}\,(\sigma Z_i - \Delta_i) - \rho\sum_{j=1}^{i-1}(\sigma Z_j - \Delta_j)\right]\right).
\end{aligned}$$

Substituting $\tilde{\Delta}_J = 0$, and $\sum_{j=1}^{J-1}\tilde{\Delta}_j = 0$ in the second square bracket, the value of $\tilde{\Delta}_{J+1}$ and $\sum_{j=1}^{J}\tilde{\Delta}_j = 0$ in the third square bracket and the value of $\tilde{\Delta}_i$ in the fourth square bracket, we split the summation $\sum_{j=1}^{i-1}(\sigma Z_j - \Delta_j)$ into $\sum_{j=1}^{J}(\sigma Z_j - \Delta_j)+\sum_{j=J+1}^{i-1}(\sigma Z_j - \Delta_j)$ (because the value of Δ_j depends on whether $j = 1, \ldots, J$ or $j = J+1, \ldots, i-1$) and substitute the value of $\sum_{j=1}^{J}\tilde{\Delta}_j = 0$ and the value of $\tilde{\Delta}_j$ for $j \geq J+1$, to obtain

$$\begin{aligned}
\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} &= -\frac{1}{\sigma^2}\Bigg(-\frac{1+(J-3)\rho}{f(J-1)}\left[\{1+(J-3)\rho\}\sigma Z_{J-1} - \rho\sigma\sum_{j=1}^{J-2} Z_j\right] \\
&+\rho\frac{1}{f(J)}\left[\{1+(J-2)\rho\}\sigma Z_J - \rho\sum_{j=1}^{J-1}\sigma Z_j\right] \\
&+\rho\frac{1}{f(J+1)}\Bigg[\{1+(J-1)\rho\}\sigma\left\{Z_{J+1} - Z_{J+1} + \frac{\rho}{1+(J-1)\rho}\sum_{j=1}^{J} Z_j\right\} \\
&-\rho\sum_{j=1}^{J}\sigma Z_j\Bigg] \\
&+\rho\sum_{i=J+2}^{I}\frac{1}{f(i)}\Bigg[\{1+(i-1)\rho\}\sigma\left\{Z_i - Z_i + \frac{\rho}{1+(J-1)\rho}\sum_{j=1}^{J} Z_j\right\} \\
&-\rho\sum_{j=1}^{J}\sigma Z_j - \rho\sum_{j=J+1}^{i-1}\sigma\left\{Z_j - Z_j + \frac{\rho}{1+(J-1)\rho}\sum_{l=1}^{J} Z_l\right\}\Bigg]\Bigg).
\end{aligned}$$

Simplifying the third square bracket, it becomes zero and, in the fourth square bracket, simplifying and substituting $\sum_{j=J+1}^{i-1}\left(\sum_{l=1}^{J} Z_l\right) = (i-1-J)\sum_{j=1}^{J} Z_j$, we obtain

$$\begin{aligned}
\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} &= -\frac{1}{\sigma^2}\Bigg(-\frac{1+(J-3)\rho}{f(J-1)}\left[\{1+(J-3)\rho\}\sigma Z_{J-1} - \rho\sigma\sum_{j=1}^{J-2} Z_j\right] \\
&+\rho\frac{1}{f(J)}\left[\{1+(J-2)\rho\}\sigma Z_J - \rho\sum_{j=1}^{J-1}\sigma Z_j\right] \\
&+\rho\sum_{i=J+2}^{I}\frac{1}{f(i)}\Bigg[\{1+(i-2)\rho\}\frac{\sigma\rho}{1+(J-1)\rho}\sum_{j=1}^{J} Z_j - \rho\sigma\sum_{j=1}^{J} Z_j \\
&-\frac{\rho^2\sigma(i-1-J)}{1+(J-1)\rho}\sum_{j=1}^{J} Z_j\Bigg]\Bigg).
\end{aligned}$$

Taking out $\rho\sigma\sum_{j=1}^{J} Z_j$ as a common factor in the third square bracket and simplifying, it becomes zero, so that

$$\begin{aligned}
\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} &= -\frac{1}{\sigma^2}\Bigg(-\frac{1+(J-3)\rho}{f(J-1)}\left[\{1+(J-3)\rho\}\sigma Z_{J-1} - \rho\sigma\sum_{j=1}^{J-2} Z_j\right] \\
&+\rho\frac{1}{f(J)}\left[\{1+(J-2)\rho\}\sigma Z_J - \rho\sum_{j=1}^{J-1}\sigma Z_j\right]\Bigg).
\end{aligned}$$

Now substituting in the values of $f(J-1)$ and $f(J)$, we obtain

$$\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} = -\frac{1}{\sigma^2}\left(\frac{-\left\{1+(J-3)\rho\right\}\left[\left\{1+(J-3)\rho\right\}\sigma Z_{J-1}-\rho\sigma\sum_{j=1}^{J-2}Z_j\right]}{\left\{1+(J-3)\rho\right\}\left\{1+(J-2)\rho\right\}(1-\rho)}\right.$$
$$\left.+\frac{\rho\left[\left\{1+(J-2)\rho\right\}\sigma Z_J-\rho\sum_{j=1}^{J-1}\sigma Z_j\right]}{\left\{1+(J-2)\rho\right\}\left\{1+(J-1)\rho\right\}(1-\rho)}\right).$$

Taking over a common denominator and simplifying, we get

$$\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} = -\frac{1}{\sigma^2 f(J)}\left[-\left\{1+(J-1)\rho\right\}\left\{1+(J-3)\rho\right\}\sigma Z_{J-1}\right.$$
$$\left.+\left\{1+(J-1)\rho\right\}\rho\sigma\sum_{j=1}^{J-2}Z_j+\left\{1+(J-1)\rho\right\}\sigma\rho Z_J-\rho^2\sigma\sum_{j=1}^{J-1}Z_j\right].$$

Rewriting $\sum_{j=1}^{J-1}Z_j=\sum_{j=1}^{J-2}Z_j+Z_{J-1}$ in the last term, this becomes

$$\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} = -\frac{1}{\sigma^2 f(J)}\left[-\left\{1+(J-1)\rho\right\}\left\{1+(J-3)\rho\right\}\sigma Z_{J-1}\right.$$
$$+\left\{1+(J-1)\rho\right\}\rho\sigma\sum_{j=1}^{J-2}Z_j+\left\{1+(J-2)\rho\right\}\sigma\rho Z_J-\rho^2\sigma\sum_{j=1}^{J-2}Z_j$$
$$\left.-\rho^2\sigma Z_{J-1}\right].$$

Gathering σZ_{J-1} and $\rho\sigma\sum_{j=1}^{J-2}Z_j$ terms, we have

$$\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} = \frac{-1}{\sigma^2 f(J)}\left(\left[-\left\{1+(J-1)\rho\right\}\left\{1+(J-3)\rho\right\}-\rho^2\right]\sigma Z_{J-1}\right.$$
$$\left.+\left[\left\{1+(J-1)\rho\right\}-\rho\right]\rho\sigma\sum_{j=1}^{J-2}Z_j+\left\{1+(J-2)\rho\right\}\rho\sigma Z_J\right).$$

Multiplying out the two braces in the first square bracket and multiplying throughout by -1 we get

$$\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} = \frac{1}{\sigma^2 f(J)}\left[\left\{1+(J-1)\rho+(J-3)\rho+(J-1)(J-3)\rho^2+\rho^2\right\}\sigma Z_{J-1}\right.$$
$$\left.-\left\{1+(J-2)\rho\right\}\rho\sigma\sum_{j=1}^{J-2}Z_j-\left\{1+(J-2)\rho\right\}\rho\sigma Z_J\right].$$

Collecting ρ and ρ^2 terms, substituting $\sigma Z_k=\hat{\Delta}_k$ and simplifying gives

$$\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} = \frac{1}{\sigma^2 f(J)}\left[\left\{1+2(J-2)\rho+(J-2)^2\rho^2\right\}\hat{\Delta}_{J-1}\right.$$
$$\left.-\left\{1+(J-2)\rho\right\}\rho\sum_{j=1}^{J-2}\hat{\Delta}_j-\left\{1+(J-2)\rho\right\}\rho\hat{\Delta}_J\right].$$

Writing the first braces as a perfect square of $\{1+(J-2)\,\rho\}^2$ and simplifying gives

$$\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} = \frac{1+(J-2)\,\rho}{\sigma^2 f(J)} \left[\{1+(J-2)\,\rho\}\,\hat{\Delta}_{J-1} - \rho\sum_{j=1}^{J-2}\hat{\Delta}_j - \rho\hat{\Delta}_J\right].$$

Since

$$\begin{aligned} \hat{\Delta}_{J-1} &> \frac{\rho}{1+(J-3)\rho}\sum_{j=1}^{J-2}\hat{\Delta}_j \\ \Rightarrow -\{1+(J-3)\rho\}\hat{\Delta}_{J-1} &< -\rho\sum_{j=1}^{J-2}\hat{\Delta}_j, \end{aligned}$$

we get the inequality

$$\begin{aligned} \left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} &> \frac{1+(J-2)\,\rho}{\sigma^2 f(J)}\rho(\hat{\Delta}_{J-1}-\hat{\Delta}_J) \\ &\geq 0, \end{aligned}$$

since $\hat{\Delta}_{J-1} > \hat{\Delta}_J$. ■

To prove the final result, we need the following Lemma.

Lemma 9 *For $r = 0, \ldots, J-k-2$,*

$$\begin{aligned} &\frac{\{1+(k+r)\,\rho\}\,\hat{\Delta}_k - \rho\sum_{j=1}^{k+r+1}\hat{\Delta}_j}{1+(k+r)\,\rho} - \rho\frac{\{1+(k+r)\,\rho\}\,\hat{\Delta}_{k+r+2} - \rho\sum_{j=1}^{k+r+1}\hat{\Delta}_j}{\{1+(k+r)\,\rho\}\,\{1+(k+r+1)\,\rho\}} \\ &\qquad = \frac{\{1+(k+r)\,\rho\}\,\hat{\Delta}_k - \rho\sum_{j=1}^{k+r+2}\hat{\Delta}_j}{1+(k+r+1)\,\rho} \end{aligned}$$

Proof.

$$\begin{aligned} &\frac{\{1+(k+r)\,\rho\}\,\hat{\Delta}_k - \rho\sum_{j=1}^{k+r+1}\hat{\Delta}_j}{1+(k+r)\,\rho} - \rho\frac{\{1+(k+r)\,\rho\}\,\hat{\Delta}_{k+r+2} - \rho\sum_{j=1}^{k+r+1}\hat{\Delta}_j}{\{1+(k+r)\,\rho\}\,\{1+(k+r+1)\,\rho\}} \\ = \;& \frac{\{1+(k+r)\,\rho\}\,\{1+(k+r+1)\,\rho\}\,\hat{\Delta}_k - \rho\,\{1+(k+r+1)\,\rho-\rho\}\sum_{j=1}^{k+r+1}\hat{\Delta}_j - \rho\,\{1+(k+r)\,\rho\}\,\hat{\Delta}_{k+r+2}}{\{1+(k+r)\,\rho\}\,\{1+(k+r+1)\,\rho\}} \\ = \;& \frac{\{1+(k+r+1)\,\rho\}\,\hat{\Delta}_k - \rho\sum_{j=1}^{k+r+2}\hat{\Delta}_j}{1+(k+r+1)\,\rho}. \end{aligned}$$

■

Lemma 10

$$\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} \geq 0,$$

for $k \leq J-2$.

Proof. In (2.19), for $k \leq J-2$, $\tilde{\Delta}_k$ and $\sum_{j=1}^{k}\Delta_j$ become zero in the first large braces. Splitting the summation before the second braces over $i = k+1, \ldots, I$ into $i = k+1, \ldots, J$

plus $i = J + 1$ plus $i = J + 2, \ldots, I$, we get

$$
\begin{aligned}
\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} = & -\frac{1}{\sigma^2}\Bigg\{-\frac{1+(k-2)\,\rho}{f(k)}\Bigg[\{1+(k-2)\,\rho\}\,\sigma Z_k - \rho\sigma\sum_{j=1}^{k} Z_j\Bigg] \\
& + \Bigg(\rho\sum_{i=k+1}^{J}\frac{1}{f(i)}\Bigg[\{1+(i-2)\,\rho\}\left(\sigma Z_i - \tilde{\Delta}_i\right) - \rho\sum_{j=1}^{i-1}\left(\sigma Z_j - \tilde{\Delta}_j\right)\Bigg] \\
& + \rho\frac{1}{f(J+1)}\Bigg[\{1+(J-1)\,\rho\}\left(\sigma Z_{J+1} - \tilde{\Delta}_{J+1}\right) - \rho\sum_{j=1}^{J}\left(\sigma Z_j - \tilde{\Delta}_j\right)\Bigg] \\
& + \rho\sum_{i=J+2}^{I}\frac{1}{f(i)}\Bigg[\{1+(i-2)\,\rho\}\left(\sigma Z_i - \tilde{\Delta}_i\right) - \rho\sum_{j=1}^{i-1}\left(\sigma Z_j - \tilde{\Delta}_j\right)\Bigg]\Bigg)\Bigg\}.
\end{aligned}
$$

The summation before the second square bracket is $i = k + 1, \ldots, J$, so both $\tilde{\Delta}_i$ and $\tilde{\Delta}_j$ become zero. In the third square bracket, substituting the value of $\tilde{\Delta}_{J+1}$ and zero for $\tilde{\Delta}_j$, since $j = 1, \ldots, J$, and in the fourth square bracket substituting the value of $\tilde{\Delta}_i$ and splitting $\sum_{j=1}^{i-1}(\sigma Z_j - \Delta_j)$ into the sums over $j = 1, \ldots, J$ (where $\tilde{\Delta}_j = 0$) and over $j = J + 1, \ldots, i - 1$ and substituting the value of $\tilde{\Delta}_j$, we obtain

$$
\begin{aligned}
\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} = & -\frac{1}{\sigma^2}\Bigg(-\frac{1+(k-2)\,\rho}{f(k)}\Bigg[\{1+(k-2)\,\rho\}\,\sigma Z_k - \rho\sigma\sum_{j=1}^{k} Z_j\Bigg] \\
& + \rho\sum_{i=k+1}^{J}\frac{1}{f(i)}\Bigg[\{1+(i-2)\,\rho\}\,\sigma Z_i - \rho\sum_{j=1}^{i-1}\sigma Z_j\Bigg] \\
& + \rho\frac{1}{f(J+1)}\Bigg[\{1+(J-1)\,\rho\}\,\sigma\Bigg\{Z_{J+1} - Z_{J+1} + \frac{\rho}{1+(J-1)\,\rho}\sum_{j=1}^{J} Z_j\Bigg\} \\
& - \rho\sum_{j=1}^{J}\sigma Z_j\Bigg] \\
& + \rho\sum_{i=J+2}^{I}\frac{1}{f(i)}\Bigg[\{1+(i-2)\,\rho\}\,\sigma\Bigg\{Z_i - Z_i + \frac{\rho}{1+(J-1)\,\rho}\sum_{j=1}^{J} Z_j\Bigg\} \\
& - \rho\sum_{j=1}^{J}\sigma Z_j - \rho\sum_{j=J+1}^{i-1}\sigma\Bigg\{Z_j - Z_j + \frac{\rho}{1+(J-1)\,\rho}\sum_{l=1}^{J} Z_l\Bigg\}\Bigg]\Bigg).
\end{aligned}
$$

After simplification the third square bracket becomes zero. In the fourth square bracket $\sum_{j=J+1}^{i-1}\sum_{l=1}^{J} Z_l = (i - 1 - J)\sum_{j=1}^{J} Z_l$, since the second summation is independent of the first summation, so we get

$$\begin{aligned}\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} &= -\frac{1}{\sigma^2}\left(-\frac{1+(k-2)\,\rho}{f(k)}\left[\{1+(k-2)\,\rho\}\,\sigma Z_k - \rho\sigma\sum_{j=1}^{k} Z_j\right]\right. \\ &\quad +\rho\sum_{i=k+1}^{J}\frac{1}{f(i)}\left[\{1+(i-2)\,\rho\}\,\sigma Z_i - \rho\sum_{j=1}^{i-1}\sigma Z_j\right] \\ &\quad +\rho\sum_{i=J+2}^{I}\frac{1}{f(i)}\left[\{1+(i-2)\,\rho\}\,\frac{\rho}{1+(J-1)\,\rho}\sum_{j=1}^{J}\sigma Z_j - \rho\sum_{j=1}^{J}\sigma Z_j\right. \\ &\quad \left.\left. -\frac{\rho^2\,(i-1-J)}{1+(J-1)\,\rho}\sum_{j=1}^{J}\sigma Z_j\right]\right).\end{aligned}$$

In the third square bracket, taking out $\rho\sigma\sum_{j=1}^{J} Z_j$ as a common factor and simplifying, we obtain

$$\begin{aligned}\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} &= -\frac{1}{\sigma^2}\left(-\frac{1+(k-2)\,\rho}{f(k)}\left[\{1+(k-2)\,\rho\}\,\sigma Z_k - \rho\sigma\sum_{j=1}^{k} Z_j\right]\right. \\ &\quad +\rho\sum_{i=k+1}^{J}\frac{1}{f(i)}\left[\{1+(i-2)\,\rho\}\,\sigma Z_i - \rho\sum_{j=1}^{i-1}\sigma Z_j\right] \\ &\quad \left. +\rho^2\sigma\sum_{j=1}^{J} Z_j\sum_{i=J+2}^{I}\frac{1}{f(i)}\left\{\frac{1+(i-2)\,\rho-1-(J-1)\,\rho-(i-1-J)\,\rho}{1+(J-1)\,\rho}\right\}\right).\end{aligned}$$

Substituting $f(k)$ and $f(i)$ and noting that, after simplification, the third square bracket becomes zero, we get

$$\begin{aligned}\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} &= \frac{1}{\sigma^2}\left(\frac{\{1+(k-2)\,\rho\}\,\sigma Z_k}{\{1+(k-1)\,\rho\}\,(1-\rho)} - \frac{\rho\sigma\sum_{j=1}^{k} Z_j}{\{1+(k-1)\,\rho\}\,(1-\rho)}\right. \\ &\quad \left. -\rho\sum_{i=k+1}^{J}\left[\frac{\sigma Z_i}{\{1+(i-1)\,\rho\}\,(1-\rho)} - \frac{\rho\sum_{j=1}^{i-1}\sigma Z_j}{\{1+(i-2)\,\rho\}\,\{1+(i-1)\,\rho\}\,(1-\rho)}\right]\right).\end{aligned}$$

Substituting $\sigma Z_j = \hat{\Delta}_j$, this simplifies to

$$\begin{aligned}\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} &= \frac{1}{\sigma^2}\left[\frac{\{1+(k-2)\,\rho\}\,\hat{\Delta}_k - \rho\sum_{j=1}^{k}\hat{\Delta}_j}{\{1+(k-1)\,\rho\}\,(1-\rho)}\right. \\ &\quad \left. -\rho\sum_{i=k+1}^{J}\frac{\{1+(i-2)\,\rho\}\,\hat{\Delta}_i - \rho\sum_{j=1}^{i-1}\hat{\Delta}_j}{\{1+(i-2)\,\rho\}\,\{1+(i-1)\,\rho\}\,(1-\rho)}\right].\end{aligned}$$

Splitting the summation over $i = k+1, \ldots, J$ into parts for $k+1$ and over $i = k+2, \ldots, J$ and then simplifying, we obtain

$$\begin{aligned}\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} &= \frac{1}{\sigma^2\,(1-\rho)}\left[\frac{\{1+(k-2)\,\rho\}\,(1+k\rho)\,\hat{\Delta}_k - \rho\,(1+k\rho)\sum_{j=1}^{k}\hat{\Delta}_j - \rho\,\{1+(k-1)\,\rho\}\,\hat{\Delta}_{k+1} + \rho^2\sum_{j=1}^{k}\hat{\Delta}_j}{\{1+(k-1)\,\rho\}\,(1+k\rho)}\right. \\ &\quad \left. -\rho\sum_{i=k+2}^{J}\frac{\{1+(i-2)\,\rho\}\,\hat{\Delta}_i - \rho\sum_{j=1}^{i-1}\hat{\Delta}_j}{\{1+(i-2)\,\rho\}\,\{1+(i-1)\,\rho\}}\right].\end{aligned}$$

Gathering $\hat{\Delta}_k$ terms and $\sum_{j=1}^{k}\hat{\Delta}_j$ terms, we get

$$\begin{aligned}
\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} &= \frac{1}{\sigma^2(1-\rho)}\left[\frac{\left\{1+(k-2)\rho+k\rho+k(k-2)\rho^2+\rho^2\right\}\hat{\Delta}_k+\left(\rho^2-\rho-k\rho^2\right)\sum_{j=1}^{k}\hat{\Delta}_j-\rho\{1+(k-1)\rho\}\hat{\Delta}_{k+1}}{\{1+(k-1)\rho\}(1+k\rho)}\right. \\
&\quad \left.-\rho\sum_{i=k+2}^{J}\frac{\{1+(i-2)\rho\}\hat{\Delta}_i-\rho\sum_{j=1}^{i-1}\hat{\Delta}_j}{\{1+(i-2)\rho\}\{1+(i-1)\rho\}}\right] \\
&= \frac{1}{\sigma^2(1-\rho)}\left[\frac{\{1+(k-1)\rho\}^2\hat{\Delta}_k-\rho\{1+(k-1)\rho\}\sum_{j=1}^{k}\hat{\Delta}_j-\rho\{1+(k-1)\rho\}\hat{\Delta}_{k+1}}{\{1+(k-1)\rho\}(1+k\rho)}\right. \\
&\quad \left.-\rho\sum_{i=k+2}^{J}\frac{\{1+(i-2)\rho\}\hat{\Delta}_i-\rho\sum_{j=1}^{i-1}\hat{\Delta}_j}{\{1+(i-2)\rho\}\{1+(i-1)\rho\}}\right].
\end{aligned}$$

Taking out $\{1+(k-1)\rho\}$ as a common factor in the first term inside the square bracket and simplifying, we obtain

$$\begin{aligned}
\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} &= \frac{1}{\sigma^2(1-\rho)}\left[\frac{\{1+(k-1)\rho\}\hat{\Delta}_k-\rho\sum_{j=1}^{k}\hat{\Delta}_j-\rho\hat{\Delta}_{k+1}}{1+k\rho}\right. \\
&\quad \left.-\rho\sum_{i=k+2}^{J}\frac{\{1+(i-2)\rho\}\hat{\Delta}_i-\rho\sum_{j=1}^{i-1}\hat{\Delta}_j}{\{1+(i-2)\rho\}\{1+(i-1)\rho\}}\right].
\end{aligned}$$

Rewriting $-\rho\sum_{j=1}^{k}\hat{\Delta}_j-\rho\hat{\Delta}_{k+1}=-\rho\sum_{j=1}^{k+1}\hat{\Delta}_j+\rho\hat{\Delta}_k$, this becomes

$$\begin{aligned}
\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} &= \frac{1}{\sigma^2(1-\rho)}\left[\frac{\{1+(k-1)\rho\}\hat{\Delta}_k-\rho\sum_{j=1}^{k+1}\hat{\Delta}_j+\rho\hat{\Delta}_k}{1+k\rho}\right. \\
&\quad \left.-\rho\sum_{i=k+2}^{J}\frac{\{1+(i-2)\rho\}\hat{\Delta}_i-\rho\sum_{j=1}^{i-1}\hat{\Delta}_j}{\{1+(i-2)\rho\}\{1+(i-1)\rho\}}\right] \\
&= \frac{1}{\sigma^2(1-\rho)}\left[\frac{(1+k\rho)\hat{\Delta}_k-\rho\sum_{j=1}^{k+1}\hat{\Delta}_j}{1+k\rho}\right. \\
&\quad \left.-\rho\sum_{i=k+2}^{J}\frac{\{1+(i-2)\rho\}\hat{\Delta}_i-\rho\sum_{j=1}^{i-1}\hat{\Delta}_j}{\{1+(i-2)\rho\}\{1+(i-1)\rho\}}\right].
\end{aligned}$$

Again splitting the summation over $i=k+2,\ldots,J$ into parts for $k+2$ and over $i=k+3,\ldots,J$ and then simplifying, we obtain

$$\begin{aligned}
\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} &= \frac{1}{\sigma^2(1-\rho)}\left[\frac{(1+k\rho)\hat{\Delta}_k-\rho\sum_{j=1}^{k+1}\hat{\Delta}_j}{1+k\rho}-\rho\frac{(1+k\rho)\hat{\Delta}_{k+2}-\rho\sum_{j=1}^{k+1}\hat{\Delta}_j}{(1+k\rho)\{1+(k+1)\rho\}}\right. \\
&\quad \left.-\rho\sum_{i=k+3}^{J}\frac{\{1+(i-2)\rho\}\hat{\Delta}_i-\rho\sum_{j=1}^{i-1}\hat{\Delta}_j}{\{1+(i-2)\rho\}\{1+(i-1)\rho\}}\right].
\end{aligned}$$

Now, repeatedly applying Lemma 6, and writing $\rho\sum_{j=1}^{J}\hat{\Delta}_j=\rho\sum_{j=1}^{J-1}\hat{\Delta}_j+\rho\hat{\Delta}_J$, we get

$$\begin{aligned}
\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} &= \frac{1}{\sigma^2(1-\rho)}\left[\frac{\{1+(J-1)\rho\}\hat{\Delta}_k-\rho\sum_{j=1}^{J}\hat{\Delta}_j}{1+(J-1)\rho}\right] \\
&= \frac{1}{\sigma^2(1-\rho)}\left[\frac{\{1+(J-1)\rho\}\hat{\Delta}_k-\rho\hat{\Delta}_J-\rho\sum_{j=1}^{J-1}\hat{\Delta}_j}{1+(J-1)\rho}\right].
\end{aligned}$$

Now $\hat{\Delta}_k \geq \hat{\Delta}_J$, $k < J \Rightarrow \hat{\Delta}_k \geq \hat{\Delta}_J$ so replacing the first term $\{(J-1)\rho\}\hat{\Delta}_k$ by $\{(J-1)\rho\}\hat{\Delta}_J$, we have the inequality

$$\begin{aligned}\left.\frac{\partial \log L}{\partial \Delta_k}\right|_{\boldsymbol{\Delta}=\tilde{\boldsymbol{\Delta}}} &\geq \frac{1}{\sigma^2(1-\rho)}\left[\frac{\{1+(J-1)\rho\}\hat{\Delta}_J - \rho\hat{\Delta}_J - \rho\sum_{j=1}^{J-1}\hat{\Delta}_j}{1+(J-1)\rho}\right] \\ &= \frac{\{1+(J-2)\rho\}\hat{\Delta}_J - \rho\sum_{j=1}^{J-1}\hat{\Delta}_j}{\sigma^2(1-\rho)\{1+(J-1)\rho\}}.\end{aligned}$$

Also

$$\begin{aligned}\hat{\Delta}_J &\geq \frac{\rho}{1+(J-2)\rho}\sum_{i=1}^{J-1}\hat{\Delta}_j \\ \Rightarrow \{1+(J-2)\rho\}\hat{\Delta}_j &\geq \rho\sum_{i=1}^{J-1}\hat{\Delta}_j,\end{aligned}$$

and so

$$\frac{\partial \log L}{\partial \Delta_k} \geq 0.$$

■

We have proved that (2.20) gives the restricted MLEs under H_0.

2.5.3 Proof of Theorem 4

From (2.8), (2.13), (2.14) and (2.17), twice the log-likelihood ratio is

$$\begin{aligned}\lambda &= -\frac{1}{\sigma^2}\left(\sum_{i=1}^{I}\frac{1}{f(i)}\left[\sigma\sqrt{f(i)}U_i - \{1+(i-2)\rho\}\hat{\Delta}_i + \rho\sum_{j=1}^{i-1}\hat{\Delta}_j\right]^2\right) \\ &\quad +\frac{1}{\sigma^2}\left(\sum_{i=1}^{I}\frac{1}{f(i)}\left[\sigma\sqrt{f(i)}U_i - \{1+(i-2)\rho\}\tilde{\Delta}_i + \rho\sum_{j=1}^{i-1}\tilde{\Delta}_j\right]^2\right).\end{aligned}$$

Since the first square bracket with unrestricted MLEs is equal to zero, we obtain

$$\lambda = \frac{1}{\sigma^2}\left[\sum_{i=1}^{I}\frac{1}{f(i)}\left\{\sigma\sqrt{f(i)}U_i - \{1+(i-2)\rho\}\tilde{\Delta}_i + \rho\sum_{j=1}^{i-1}\tilde{\Delta}_j\right\}^2\right]. \qquad (2.21)$$

Splitting the summation into parts over $i = 1, \ldots, J$ and $i = J+1, \ldots, I$, we get

$$\begin{aligned}\lambda &= \frac{1}{\sigma^2}\left(\sum_{i=1}^{J}\frac{1}{f(i)}\left[\sigma\sqrt{f(i)}U_i - \{1+(i-2)\rho\}\tilde{\Delta}_i + \rho\sum_{j=1}^{i-1}\tilde{\Delta}_j\right]^2\right. \\ &\quad \left. + \sum_{i=J+1}^{I}\frac{1}{f(i)}\left[\sigma\sqrt{f(i)}U_i - \{1+(i-2)\rho\}\tilde{\Delta}_i + \rho\sum_{j=1}^{i-1}\tilde{\Delta}_j\right]^2\right).\end{aligned}$$

For $i = 1, \ldots, J$, $\tilde{\Delta}_i$ is zero. Hence, substituting the values of $\tilde{\Delta}_i$ and U_i from (2.20) and (2.10) respectively and simplifying we get

$$\begin{aligned}\lambda &= \frac{1}{\sigma^2}\left(\sum_{i=1}^{J}\frac{1}{f(i)}\left[\sigma\left\{1+(i-2)\rho\right\}Z_i - \sigma\rho\sum_{k=1}^{i-1}Z_k\right]^2\right. \\ &\quad \left. + \sum_{i=J+1}^{I}\frac{1}{f(i)}\left[\sigma\left\{1+(i-2)\rho\right\}Z_i - \rho\sigma\sum_{j=1}^{i-1}Z_k - \left\{1+(i-2)\rho\right\}\tilde{\Delta}_i + \rho\sum_{j=1}^{i-1}\tilde{\Delta}_j\right]^2\right).\end{aligned}$$

Replacing $\sum_{j=1}^{i-1}\tilde{\Delta}_j$ by $\sum_{j=J+1}^{i-1}\tilde{\Delta}_j$ in the second braces, since $\tilde{\Delta}_s = 0$ for $s = 1, \ldots, J$, substituting the value of $\tilde{\Delta}_s$ for $s = J+1, \ldots, I$ and simplifying we get

$$\begin{aligned}\lambda &= \frac{1}{\sigma^2}\left(\sum_{i=1}^{J}\frac{1}{f(i)}\left[\sigma\left\{1+(i-2)\rho\right\}Z_i - \sigma\rho\sum_{k=1}^{i-1}Z_k\right]^2\right. \\ &\quad + \sum_{i=J+1}^{I}\frac{1}{f(i)}\left[-\rho\sigma\sum_{k=1}^{i-1}Z_k + \frac{\left\{1+(i-2)\rho\right\}\sigma\rho}{1+(J-1)\rho}\sum_{j=1}^{J}Z_j\right. \\ &\quad \left.\left. +\rho\sigma\sum_{j=J+1}^{i-1}\left\{Z_j - \frac{\rho}{1+(J-1)\rho}\sum_{k=1}^{J}Z_k\right\}\right]^2\right).\end{aligned}$$

Multiplying out the last braces in the second square brackets and simplifying, we get

$$\begin{aligned}\lambda &= \frac{1}{\sigma^2}\left(\sum_{i=1}^{J}\frac{1}{f(i)}\left[\sigma\left\{1+(i-2)\rho\right\}Z_i - \sigma\rho\sum_{k=1}^{i-1}Z_k\right]^2 + \right. \\ &\quad \left. \sum_{i=J+1}^{I}\frac{1}{f(i)}\left[-\rho\sigma\sum_{j=1}^{J}Z_j + \frac{\left\{1+(i-2)\rho\right\}\rho\sigma}{1+(J-1)\rho}\sum_{j=1}^{J}Z_j - \frac{(i-J-1)\sigma\rho^2}{1+(J-1)\rho}\sum_{j=1}^{J}Z_j\right]^2\right).\end{aligned}$$

Simplifying the second square brackets, we obtain

$$\begin{aligned}\lambda &= \sum_{i=1}^{J}\frac{1}{f(i)}\left[\left\{1+(i-2)\rho\right\}Z_i - \rho\sum_{j=1}^{i-1}Z_j\right]^2 \\ &\quad +\rho\sum_{j=1}^{J}Z_j\left\{\sum_{i=J+1}^{I}\frac{1}{f(i)}\frac{-1-(J-1)\rho+1+(i-2)\rho-(i-J-1)\rho}{1+(J-1)\rho}\right\}^2 \\ &= \sum_{i=1}^{J}\frac{1}{f(i)}\left[\left\{1+(i-2)\rho\right\}Z_i - \rho\sum_{j=1}^{i-1}Z_j\right]^2.\end{aligned}$$

Finally substituting $f(i)$ from (2.12), we obtain

$$\begin{aligned}\lambda &= \sum_{i=1}^{J}\frac{\left[\left\{1+(i-2)\rho\right\}Z_i - \rho\sum_{j=1}^{i-1}Z_j\right]^2}{\left\{1+(i-2)\rho\right\}\left\{1+(i-1)\rho\right\}(1-\rho)} \\ &= Z_1^2 + \sum_{j=2}^{I}\frac{(Z_j - E_j)^2}{V_j}.\end{aligned}$$

Since in Case 0, $\lambda = 0$, it follows that $\lambda = T_2^2$.

Chapter 3

Properties of Test Statistics

Since the properties of the test statistics defined in the previous chapter are not completely known, and there is no uniformly most powerful test, it is not clear which is best for practical use in clinical trials. In this chapter we study their properties. The aim of this chapter is to compare the performances of different test procedures, in terms of maximising power and correctly selecting the best treatment, in the case of equal allocation. Other allocations will be studied in Chapter 4. The emphasis is on three-arm trials, though we also briefly consider trials with more arms. Various authors (Robertson et al., 1988; Dunnett and Tamhane, 1992; Horn and Vollandt, 1998; Horn and Dunnett, 2004) have compared some of these tests in terms of power. The comparison here is more comprehensive than those.

The null distributions of the test statistics for three arms are studied in Section 3.1. We prove a theorem which allows us to obtain the critical values for rejecting the null hypothesis by simulating from a single point within the parameter space defined by H_0 : $\Delta_i \leq 0 \; \forall i$ and show that all of the test statistics we consider satisfy the conditions of the theorem. Simulation under the null hypothesis is used to determine the boundaries of the critical regions at $5, 2.5$ and 1% for $I = 2$. Various graphs are given to illustrate the null distributions and the rejection boundaries. In Section 3.2 examples of the practical implementation of the tests are given. More simulations are reported which illustrate the properties of the tests in Section 3.3. Power functions calculated from the simulation results are shown in graphs and tables. A loss function, which takes account of type III and IV errors and their impact, is defined and the expected loss is shown in graphs and tables. Some conclusions are drawn in Section 3.4. The specific problem of the probability of a type-III error being greater than the size of the test is considered in Section 3.5. We also briefly study $I = 3$ and $I = 7$ in Section 3.6 and give results in terms of power and expected loss.

3.1 Null distributions

3.1.1 A general result for calculating the null distribution

As emphasised in earlier chapters, the null hypothesis $H_0 : \Delta \leq 0\ \forall i$ we consider is composite, which means that the null distributions of the test statistics are not fully determined but depend on which values the parameters take within the set defined by H_0. Hence, the probability of a type I error depends on the specific parameter values and, in order to find critical values for a test at a fixed significance level, or to find p-values, we need to find the maximum probability of a type I error over all parameter values in the set defined by H_0. We now show that this is easily accomplished and that we need only calculate these probabilities when all treatments are equivalent to the control, i.e. when $\mathbf{\Delta} = \mathbf{0}$.

Lemma 11 *Let $\mathbf{Z} \sim N_I(\mathbf{\Delta}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = \rho\mathbf{J} + (1-\rho)\mathbf{I}$, $\mathbf{I}$ is the I-dimensional identity matrix and $\mathbf{J}$ is an $I \times I$ matrix of 1s. Let $\mathbf{\Delta}_1$ and $\mathbf{\Delta}_2$ be $I \times 1$ vectors such that $\Delta_{1i} \leq \Delta_{2i}$, $\forall i \in \{1,\ldots,I\}$. Let $A_k = \{\mathbf{z} : z_i \leq a_{ki}; i = 1,\ldots,I\}$, $k = 1,\ldots,m$, where a_{ki} are real constants.*

$$P\left(\mathbf{Z} \in \bigcup_{k=1}^{m} A_k \middle| \mathbf{\Delta} = \mathbf{\Delta}_1\right) \geq P\left(\mathbf{Z} \in \bigcup_{k=1}^{m} A_k \middle| \mathbf{\Delta} = \mathbf{\Delta}_2\right).$$

Proof. By the inclusion-exclusion theorem in probability,

$$P\left(\mathbf{Z} \in \bigcup_{k=1}^{m} A_k \middle| \mathbf{\Delta} = \mathbf{\Delta}_j\right) = \sum_{k=1}^{m} (-1)^{k-1} \sum_{K \in \mathcal{K}} P\left(\mathbf{Z} \in \bigcap_{l \in K} A_l \middle| \mathbf{\Delta} = \mathbf{\Delta}_j\right),$$

where $j = 1,2$ and $\mathcal{K}$ is the set of $K \subset \{1,\ldots,m\}$ such that $|K| = k$. Then

$$\begin{aligned} P\left(\mathbf{Z} \in \bigcup_{k=1}^{m} A_k \middle| \mathbf{\Delta} = \mathbf{\Delta}_j\right) &= \sum_{k=1}^{m} (-1)^{k-1} \sum_{K \in \mathcal{K}} P\left(Z_1 \leq \min_{l \in K}(a_{l1}), \ldots, Z_I \leq \min_{l \in K}(a_{lI}) \middle| \mathbf{\Delta} = \mathbf{\Delta}_j\right) \\ &= \sum_{k=1}^{m} (-1)^{k-1} \sum_{K \in \mathcal{K}} F_j\left(\min_{l \in k}(a_{l1}), \ldots, \min_{l \in k}(a_{lI}),\right), \end{aligned}$$

where F_j is the c.d.f. of $N(\mathbf{\Delta}_j, \mathbf{\Sigma})$, $j = 1,2$. Therefore,

$$P\left(\mathbf{Z} \in \bigcup_{k=1}^{m} A_k \middle| \mathbf{\Delta} = \mathbf{\Delta}_j\right) = \sum_{k=1}^{m} (-1)^{k-1} \sum_{K \in \mathcal{K}} \Phi\left(\min_{l \in k}(a_{l1}) - \Delta_{j1}, \ldots, \min_{l \in k}(a_{lI}) - \Delta_{jI}\right),$$

where Φ is the c.d.f. of $N(0, \mathbf{\Sigma})$.

Hence,

$$\begin{aligned}
P\left(\mathbf{Z}\in\bigcup_{j=1}^{m}A_k\middle|\,\boldsymbol{\Delta}=\boldsymbol{\Delta}_1\right) &= \sum_{k=1}^{m}(-1)^{k-1}\sum_{K\in\mathcal{K}}\Phi\left(\min_{l\in k}(a_{l1})-\Delta_{11},\ldots,\min_{l\in k}(a_{lI})-\Delta_{1I}\right)\\
&= \sum_{k=1}^{m}(-1)^{k-1}\sum_{K\in\mathcal{K}}F_2\left(\min_{l\in k}(a_{l1})-\Delta_{11}+\Delta_{21},\ldots\right.\\
&\qquad\qquad\left.\ldots,\min_{l\in k}(a_{lI})-\Delta_{1I}+\Delta_{2I}\right)\\
&= P\left(\mathbf{Z}\in\bigcup_{j=1}^{m}C_k\middle|\,\boldsymbol{\Delta}=\boldsymbol{\Delta}_2\right),
\end{aligned}$$

where $C_k=\{\mathbf{z}:z_i\le a_{ki}-\Delta_{1i}+\Delta_{2i};\ i=1,\ldots,I\}$, $k=1,\ldots,m$.

Since $\Delta_{1i}\le\Delta_{2i},\ i=1,\ldots,I$, $\bigcup_{k=1}^{m}A_k\subset\bigcup_{k=1}^{m}C_k$ and this implies that

$$P\left(\mathbf{Z}\in\bigcup_{j=1}^{m}A_k\middle|\,\boldsymbol{\Delta}=\boldsymbol{\Delta}_1\right)\ge P\left(\mathbf{Z}\in\bigcup_{j=1}^{m}A_k\middle|\,\boldsymbol{\Delta}=\boldsymbol{\Delta}_2\right).$$

■

By letting $m\to\infty$, we obtain the following corollary.

Corollary 12

$$P\left(\mathbf{Z}\in\bigcup_{k=1}^{\infty}A_j\middle|\,\boldsymbol{\Delta}=\boldsymbol{\Delta}_1\right)\ge P\left(\mathbf{Z}\in\bigcup_{k=1}^{\infty}A_j\middle|\,\boldsymbol{\Delta}=\boldsymbol{\Delta}_2\right).$$

Then the following result follows immediately.

Theorem 13 *Under H_0, P (Reject $H_0\mid\boldsymbol{\Delta}$) is maximised when $\boldsymbol{\Delta}=\mathbf{0}$ if the acceptance region of H_0 can be written as $\bigcup_{k=1}^{m}A_k$ or $\bigcup_{k=1}^{\infty}A_k$.*

Note that we do not actually require $\mathbf{Z}$ to have a normal distribution, although that is the case we use in this thesis. All that is needed is that the second and higher order moments of the distribution of $\mathbf{Z}$ do not depend on the mean vector $\boldsymbol{\Delta}$.

Corollary 14 *Lemma 11, Corollary 12 and Theorem 13 hold for any family of distributions of Z_i having cumulative distribution function F_i, where $F_i(z)=F_0(z-\Delta_i)$ with F_0 having zero expectation and unit variance.*

This seems to be a useful general result which, to the best of our knowledge, has not been previously published in this generality. Silvapulle and Sen (2005) showed the same result for a class of likelihood ratio test statistics, which includes T_2^2, but none of the other test procedures considered here.

Any real-valued test statistic which is non-decreasing if the value of any Z_i increases, has an acceptance region with the required form. This is clear, since any point $\mathbf{a}_k$ on the rejection boundary has the property that any $\mathbf{z}$, such that $z_i \leq a_{ki}$ $\forall k \in \{1, \ldots, I\}$, is in the acceptance region.

From the definitions of the T_k^0 and T_k, given in equations (2.2) and (2.7) respectively in Chapter 2, it can be seen that each Z_i contributes only if it is greater than zero and this contribution increases with Z_i. Hence, all test statistics in the families T_k^0 and T_k are non-decreasing in each Z_i and hence Theorem 13 applies to all of these procedures. This is not in general true for the family S_k^0, but it is true for S_1^0, which is the only member of this family we consider here. Hochberg's procedure does not correspond to a single real-valued test statistic, so we need to consider it separately.

From Chapter 2, Hochberg's procedure rejects H_0 if, for any $j \in \{1, \ldots, I\}$, $X_j \geq z_{\alpha^*/(I-j+1)}$ for a suitably chosen constant α^*. Hence, H_0 is accepted if, for all j, $X_j < z_{\alpha^*/(I-j+1)}$. This acceptance region can clearly be written in the form of a union of sets of the form A_k and so Hochberg's procedure satisfies Theorem 13.

All of these results can be seen graphically for $I = 2$ and we will do this in Section 3.1.4.

3.1.2 Evaluating the null distributions

We have shown in Section 3.1.1 that, for all test statistics considered here, if $P(\text{type I error}|\boldsymbol{\Delta} = \mathbf{0}) = \alpha$ then, for any $\boldsymbol{\Delta}$ within H_0, $P(\text{type I error}|\boldsymbol{\Delta}) \leq \alpha$. Hence we consider the null distributions at $\boldsymbol{\Delta} = \mathbf{0}$. All distributional statements in the rest of this section are to be understood as being for $\boldsymbol{\Delta} = \mathbf{0}$. It should be noted, however, that when H_0 is true, the probability of rejecting H_0 might be less than $100\alpha\%$.

The tests statistics and the key notations defined in Chapter 2 are shown again in Table 3.1 for $I = 2$.

It is trivial to show that $S_1^0 \sim N(0, 3)$, so standard normal tables can be used. The null distribution of T_∞ can be obtained by numerical integration of the bivariate normal density (Dunnett, 1955), which is available in several packages, including `R` (R Development Core Team, 2009). The Hochberg procedure does not use a univariate test statistic so the decision to accept or reject H_0 is based directly on the bivariate normal density function

Table 3.1: Test statistics and some key notations from Chapter 2.

$S_1^0 = Z_1 + Z_2$
$T_1^0 = Z_1^+ + Z_2^+$
$T_2^0 = \sqrt{Z_1^{+2} + Z_2^{+2}}$
$T_\infty = \max(Z_1, Z_2)^+$
$T_1 = \max(Z_1, Z_2)^+ + \frac{(\min(Z_1,Z_2)-\rho\max(Z_1,Z_2))^+}{\sqrt{1-\rho^2}}$
$T_2 = \sqrt{\max(Z_1, Z_2)^{+2} + \frac{(\min(Z_1,Z_2)-\rho\max(Z_1,Z_2))^{+2}}{1-\rho^2}}$
$A^+ = \max(0, A)$
$\rho = Cov(Z_1, Z_2) = \delta/\{1-\delta\}$
$E(Z_i) = \Delta_i/\sigma,\ i = 1, 2$
$\sigma^2 = \{1-\delta\}/\{\delta(1-2\delta)\}$

and so is easily computed, but here we use an adjusted α^*, which is not known, to achieve true size α, so that further work is required to find the rejection region. Mukerjee et al. (1985) and Robertson et al. (1988) showed that the LRT statistic T_2^2 has a mixture of χ^2 distributions but with unknown mixing probabilities. They developed approximations which are quite complex to compute and whose accuracy is not completely known. Indeed, Sen and Silvapulle (2001) suggested that these difficulties were the main reason for the lack of application of likelihood ratio tests in the general area of order restricted inference. Nothing is known about the null distribution of any of the other test statistics used here.

3.1.3 Simulations

The distributions of all the test statistics were approximated by Monte Carlo simulations, including for comparability those which can be calculated directly, in order to compare their properties. These simulations are then used to determine the boundaries of the critical regions at 5, 2.5 and 1%. The correlation was set to be $\rho = 0.5$, which corresponds to equal numbers of subjects in each arm.

The null distributions were approximated by simulating one million values of Z_1 and Z_2 with $\Delta_1 = \Delta_2 = 0$ using the `mvrnorm` function in `R`, specifying the expectation to be a zero vector and the covariance matrix to be

$$\mathbf{\Sigma} = (1-\rho)\mathbf{I} + \rho\mathbf{J},$$

where $\mathbf{I}$ is the $I \times I$ identity matrix and $\mathbf{J}$ is an $I \times I$ matrix of ones. The value of each test statistic is calculated from each of these pairs of Z_1 and Z_2. For all tests except the Hochberg procedure, the critical region at $100\alpha\%$ was defined by the biggest $100\alpha\%$ of the one million values of the test statistic, since the tests are one-sided. Although the critical points estimated were reasonably stable, in order to improve the accuracy we repeated this whole procedure 1000 times and the mean of the critical points was used. The null hypothesis should be rejected if the value of the test statistic is greater than the calculated critical value.

For Hochberg's procedure, the nominal α^* was found iteratively so as to achieve the required size in a separate batch of two million simulations. We also used the tabulated critical values, i.e. the procedure of Hochberg (1988) in its original form, and found that they are indeed conservative, as expected. Since all the other procedures studied here have exact size α (up to the accuracy of the simulations), we do not present the results from the conservative Hochberg procedure and in the remainder of this chapter references to Hochberg's procedure refer to the modified procedure with exact size α.

3.1.4 Results

The estimated mean cutpoints at which the 5%, 2.5% and 1% rejection region boundaries cross the Z_1 and Z_2 axes for each of the test statistics is shown in Table 3.2, along with their estimated standard errors. The standard errors are estimated by taking the sample variance of each of 10 batches of 100 simulated cutpoints, each from one million simulations, and pooling them to obtain an estimate of the variance of cutpoints calculated from a billion simulations. The square root of this is the standard error reported. Since each set of one million simulations takes several minutes to run, obtaining more precision would take a considerable amount of computing time, at least using the `R` (R Development Core Team, 2009) package.

The estimated standard errors show that the cutpoints are precise to at least three decimal places, and reasonably precise to four decimal places, although the critical values for 1% are slightly less well estimated than the others. Using the `pmnorm` function in `R` (R Development Core Team, 2009) for numerical integration of bivariate normal density functions we calculated, by repeated interpolation, the upper 95%, 97.5% and 99% values for T_∞ to be 1.91633, 2.21214 and 2.55782 respectively. The mean cutpoints in Table 3.2 to one additional decimal place are 1.91635, 2.21210 and 2.55772 and so are accurate enough, at least for T_∞, for all practical purposes. For S_1^0 the distribution is $N(0,3)$, from which we can easily calculate the values 2.85030, 3.39575 and 4.02275, for 5%, 2.5%

Table 3.2: Mean (s.e.) of the cutpoints of the rejection region boundaries with $\rho = 0.5$.

Test	5%	2.5%	1%
S_1^0	2.8488	3.3948	4.0291
	(0.00011)	(0.00014)	(0.00021)
T_1	2.1537	2.5289	2.9839
	(0.00008)	(0.00010)	(0.00015)
T_2	1.9545	2.2579	2.6120
	(0.00006)	(0.00008)	(0.00012)
T_∞	1.9163	2.2121	2.5577
	(0.00006)	(0.00008)	(0.00011)
T_1^0	2.8505	3.3950	4.0291
	(0.00011)	(0.00014)	(0.00021)
T_2^0	2.1671	2.5301	2.9584
	(0.00002)	(0.00010)	(0.00014)
Hoch*	1.9381	2.2265	2.5662

* Simulated separately and based on nominal α^* such that the size is α.

and 1% respectively, compared with the mean cutpoints from the simulations of 2.84883, 3.39476 and 4.02912. Although these are not as close as for T_∞ (see also the estimated standard errors in Table 3.2), they are good enough for practical purposes. For example, they are at least an order of magnitude better than using 2.58, rather than 2.5758, for the upper 0.5% point in a standard normal distribution. In Section 3.3, we will consider the impact of the variances of the cutpoints on the power of the tests.

For the modified Hochberg procedure used, the nominal α^* values are 0.02631, 0.01299 and 0.00514 for 5%, 2.5% and 1% respectively. Since this was based on a separate batch of two million simulations, these cutpoints are less precise. A rough calculation suggests that their standard errors should be about $\sqrt{500}$ times those shown, giving approximately 0.002, so the cutpoints should be precise to two, and reasonably precise to three, decimal places.

Figure 3.1 shows histograms of the null distributions for five of the test statistics at 5%, with zeros deleted to make the pattern clearer. Around one third of the simulations gave zero for all the test statistics except S_1^0, corresponding to the case where both experimental arms give worse responses than the control. The theoretical value of this is exactly 1/3, since, with equal sample sizes, it is completely random chance which of the three arms gives the highest response. Note that in this and several of the following figures, T_3 refers to T_∞. All the distributions seem from the histograms to be fairly close to normal truncated below their means, but normal probability plots (not shown) show that this is not a good enough approximation. Since the null distribution of S_1^0 is known to be N(0,3), we do not show it. Hochberg's procedure does not use a real-valued test statistic, so we do not plot its null distribution.

The rejection boundary at the 5% level of significance for each test statistic is plotted, along with the contours of the joint p.d.f. of Z_1 and Z_2 under H_0 at $\mathbf{\Delta} = \mathbf{0}$, in Figure 3.2. It can be seen that every rejection boundary crosses each other boundary at some point, since all tests are at 5%. Each test "slices off" 5% of the bivariate probability distribution, but uses a different pattern to slice off a different 5%. It is also immediately clear that all the rejection boundaries meet the conditions for Theorem 13 to hold.

We see that T_∞, T_2 and Hochberg's procedure are all fairly similar, but have noticeable differences close to $Z_1 = Z_2$. The simplest rejection boundary is that of T_∞ and this is one reason for its popularity in practice. T_2 has the seemingly attractive property that its boundary is smooth and more closely follows the contours of the null distribution in the upper right quadrant. In comparison, the boundaries of T_∞ and the Hochberg procedure change abruptly. With T_2, the null hypothesis is never rejected with a p.d.f. > 0.03, but this is not true for any of the other tests. This seems like a desirable property of T_2, that it rejects H_0 when $\mathbf{Z}$ has values which have the smallest probability density under H_0. This, of course, is a consequence of its equivalence to the likelihood ratio test. With T_1^0, for some values with p.d.f. < 0.01, we still do not reject H_0, but for some values with p.d.f. > 0.04, H_0 is rejected. This seems undesirable. It can be seen that the Hochberg test is close to T_∞ when Z_1 and Z_2 are dissimilar, but is closer to T_2 near $Z_1 = Z_2$. T_1 appears to be quite different from T_2, T_∞ and Hochberg. The disadvantage of S_1^0 can also be seen, as the rejection boundary keeps changing linearly beyond the axes, where a large positive value of Z_1, for example, can be cancelled out by a large negative value of Z_2.

Figure 3.3 shows the same rejection boundaries, but with the contours of the null distribution when $\Delta_1 = -2$ and $\Delta_2 = -1$. In this figure the contours are in steps of 0.001, in order to better show the extremes of the distribution. The results of Theorem 13 can be illustrated by comparing Figure 3.3 with Figure 3.2. They show clearly that H_0 is less likely to be rejected when $\Delta_1 < 0$ and $\Delta_2 < 0$ than when $\Delta_1 = \Delta_2 = 0$.

The distribution of Z_1 and Z_2 under H_1, with $\Delta_1 = 2$ and $\Delta_2 = 0$ is shown in Figure 3.4, along with the rejection boundaries. They show clearly that H_0 is more likely to be

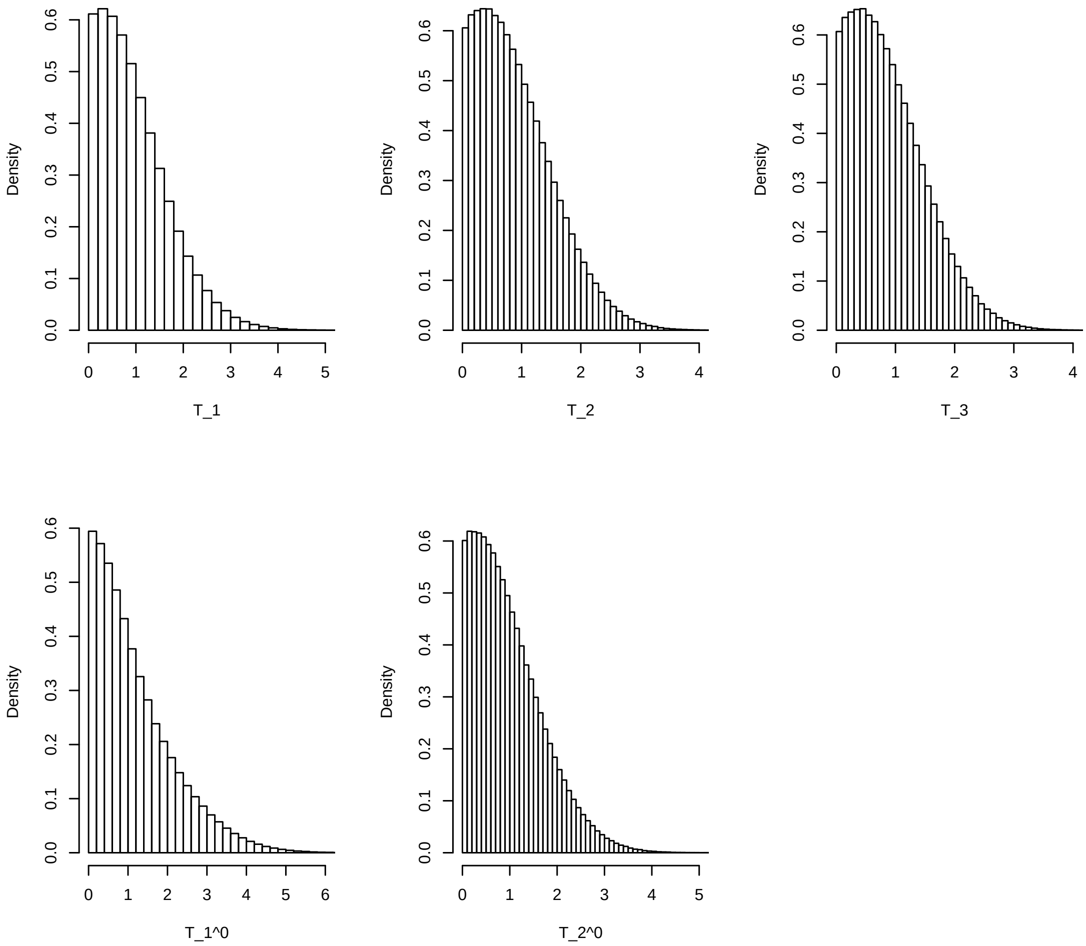

Figure 3.1: Histogram of the null distribution excluding zero for T_1, T_2, T_∞, T_1^0 and T_2^0. The percentage at zero is 33.329%.

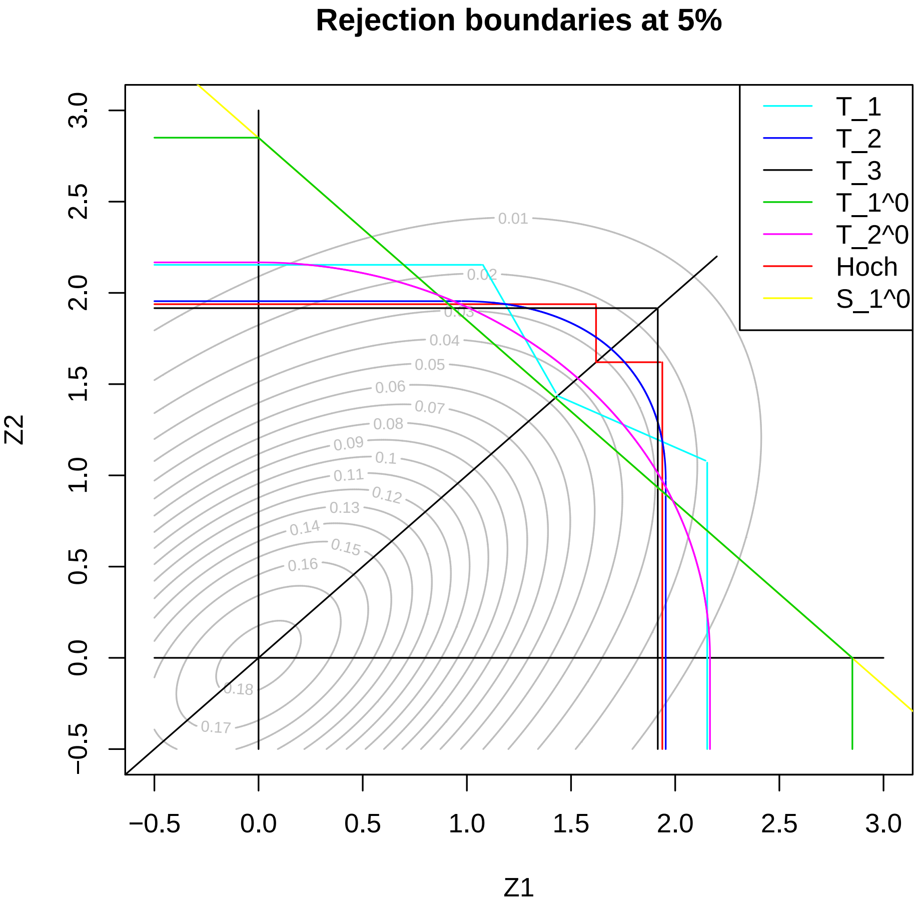

Figure 3.2: Rejection boundaries with contours of the joint null distribution of Z_1 and Z_2 with $(\Delta_1, \Delta_2) = (0, 0)$.

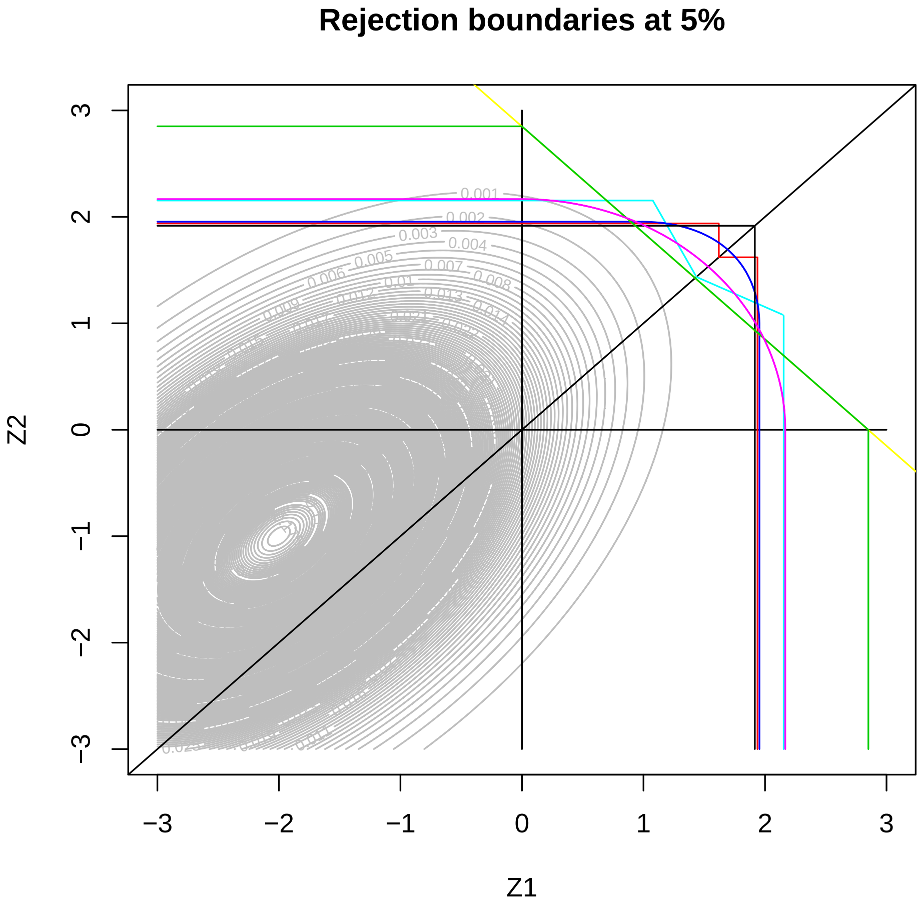

Figure 3.3: Rejection boundaries with contours of the joint null distribution of Z_1 and Z_2 with $(\Delta_1, \Delta_2) = (-2, -1)$. For key, see Figure 3.2.

rejected when it is false than when it is true. It is also easily seen that, at these particular values of Δ_1 and Δ_2, the probability of rejecting H_0, i.e. the power, is different for different test procedures. For example, it is immediately obvious that the rejection region for S_1^0 includes less than 50% of the probability distribution, whereas those for T_2, T_∞ and the Hochberg procedure include more than 50%.

3.2 Examples

We now discuss the two examples of three-arm trials mentioned in the introduction to illustrate our test statistics described in Chapters 1 and 2. We consider trials with fairly large sample sizes, so that we can assume that we can use the normal approximation in our summary statistics.

3.2.1 Example 1

The DASH (Dietary Approaches to Stop Hypertension) trial (Appel et al., 1997) was a randomized trial consisting of two experimental diets and a control diet for studying the effect of dietary pattern on blood pressure. The control diet had the fat content of a typical United States diet and was low in fruits, vegetables and dairy products. Diet 1 was a diet rich in fruits and vegetables and diet 2 was a "combination" diet rich in fruits, vegetables and low-fat dairy products and with reduced saturated and total fat.

Let D_i denote the unscaled difference in blood pressure between the control diet and experimental diet i, where $i = 1, 2$. In our notation $Z_i = \hat{D}_i/se(\hat{D}_i)$, which has unit variance. In fact, the estimated standard errors were not given in the paper, so we calculated them from the confidence intervals given. This is subject to rounding error as only one decimal place was shown, but the results should be accurate enough for illustration. The number of subjects on the control, fruit and combination diet were 154, 151 and 154 respectively. We treat these as approximately equal allocations, so that $\rho = 0.5$.

From the results in the paper, for systolic blood pressure, we get $D_1 = 2.8$, $D_2 = 5.5$ and $se(\hat{D}_i) = 0.884$. Then $Z_1 = 2.8/0.884 = 3.3472$ and $Z_2 = 5.5/0.884 = 6.5748$. Therefore, we can calculate

$$S_1^0 = 6.5748 + 3.3472 = 9.9220,$$

$$T_1 = 6.5748 + \frac{(3.3472 - 0.5 \times 6.5748)}{\sqrt{1 - 0.5^2}} = 6.6538,$$

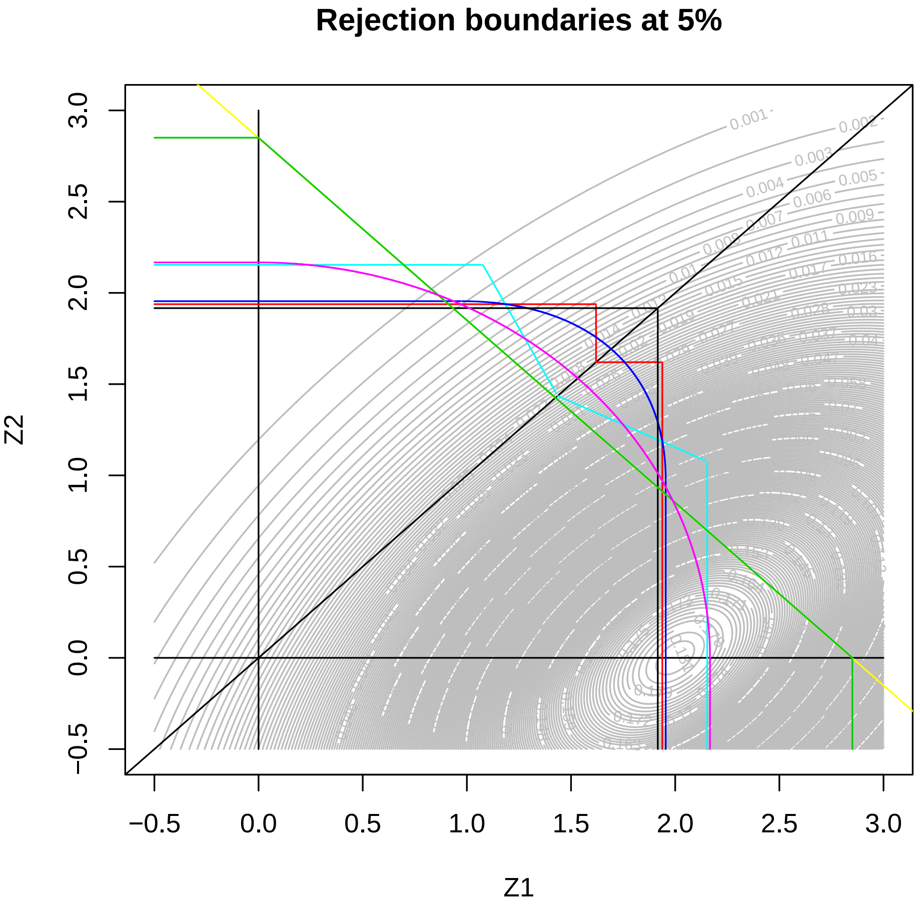

Figure 3.4: Rejection boundaries with contours of the joint null distribution of Z_1 and Z_2 with $(\Delta_1, \Delta_2) = (2, 0)$. For key, see Figure 3.2.

$$T_2 = \sqrt{6.5748^2 + \frac{(3.3472 - 0.5 \times 6.5748)^2}{1 - 0.5^2}} = 6.5751,$$

$$T_\infty = 6.5748,$$

$$T_1^0 = 6.5748 + 3.3472 = 9.9220,$$

$$T_2^0 = \sqrt{6.5748^2 + 3.3472^2} = 7.3778$$

and, compared with the cutpoints in Table 3.2, we can see that they are all clearly significant at the 1% level. Since $Z_1 > 2.5662$, this is also true for the Hochberg procedure.

For diastolic blood pressure, $\hat{D}_1 = 1.1$, $\hat{D}_2 = 3.0$ and $se(\hat{D}_i) = 0.602$. Hence $Z_1 = 1.1/0.602 = 1.8263$ and $Z_2 = 3/0.602 = 4.9809$. Note that the second best experimental diet, i.e. the fruit and vegetable diet, is not bigger than ρ times the best diet, so it does not contribute to the test statistics T_1 and T_2. Here T_1, T_2 and T_∞ all have the same value of 4.9809, $S_1^0 = T_1^0 = 4.9809 + 1.8263 = 6.8072$, $T_2^0 = \sqrt{4.9809^2 + 1.8263^2} = 5.3052$ and all are significant at the 1% level from Table 3.2. Also, since $Z_2 > 2.5662$, Hochberg's procedure rejects H_0 at the 1% level.

For both systolic and diastolic blood pressure, we conclude that at least one diet is better than control. From the estimates, we would recommend the combination diet.

3.2.2 Example 2

We use the results from the ATAC (Arimidex, Tamoxifen Alone or in Combination) randomized breast cancer trial (The ATAC Trialists' Group, 2002), comparing the efficacy of tamoxifen (T) with anastrozole (A) alone and the combination of the two (A+T) in terms of 3 years disease-free survival, which we treat as a binary outcome, and in terms of the survival time. Tamoxifen is the control treatment. Disease-free survival at 3 years was 87.4%, 89.4% and 87.2% on tamoxifen, anastrozole and the combination respectively. The number of subjects was 3125 for both anastrozole and the combination and 3116 for tamoxifen, so that a normal approximation should be good and the sample sizes are almost equal.

Let

$$\psi_i = \frac{\pi_i/(1-\pi_i)}{\pi_0/(1-\pi_0)},$$

$i = 1, 2$, be the odds ratio of disease-free survival for treatment i and control, $i = 0$, i.e. tamoxifen, where π_i is the disease-free survival for treatment i. This will be estimated by

$$\hat{\psi}_i = \frac{\hat{\pi}_i/(1-\hat{\pi}_i)}{\hat{\pi}_0/(1-\hat{\pi}_0)}.$$

For large N, $\log\hat{\psi}_i \sim N(\log\psi_i, Var(\log\hat{\psi}_i))$ (Armitage et al., 2002, p.127), where

$$Var(\log\hat{\psi}_i) = \frac{1}{n_0\pi_0(1-\pi_0)} + \frac{1}{n_i\pi_i(1-\pi_i)}.$$

Then we use

$$\widehat{Var}(\log\hat{\psi}_i) = \frac{1}{n_0\hat{\pi}_0(1-\hat{\pi}_0)} + \frac{1}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}.$$

We scale the log odds ratios to get Z_i with unit variance, so that

$$\begin{aligned} Z_i &= \frac{\log\hat{\psi}_i}{\sqrt{\widehat{Var}(\log\hat{\psi}_i)}} \\ &= \log\left(\frac{\hat{\pi}_i/(1-\hat{\pi}_i)}{\hat{\pi}_0/(1-\hat{\pi}_0)}\right)\sqrt{\frac{n_0n_i\hat{\pi}_0\hat{\pi}_i(1-\hat{\pi}_0)(1-\hat{\pi}_i)}{n_i\hat{\pi}_i(1-\hat{\pi}_i)+n_0\hat{\pi}_0(1-\hat{\pi}_0)}}. \end{aligned}$$

Using $\hat{\pi}_1 = 0.894$, the survival in anastrozole, and $\hat{\pi}_2 = 0.872$ in the combination group and comparing with the tamoxifen group, which gave $\hat{\pi}_0 = 0.874$, we have $Z_1 = 2.4644$ and $Z_2 = -0.2373$. Then T_1, T_2, T_∞, T_1^0 and T_2^0 all equal 2.4644 so that all except T_1^0 show significance at the 5% level and T_2 and T_∞ also show significance at 2.5%. Since $Z_1 > 2.2265$ Hochberg's procedure also leads to rejection at the 2.5% level. Here $S_1^0 = 2.4644 - 0.2373 = 2.2271$, so that it does not show significance even at 5%. This example illustrates clearly the weakness of S_1^0, that one inferior treatment cancels out the gain from the superior treatment so that one would fail to reject H_0. T_1^0, while less extreme, also fails to detect a difference in this case.

We also use the hazard ratios (HR) of disease-free survival times on experimental arms versus the control estimated from the log-rank test statistics. The estimated HR for A against T was 0.83 (95% CI 0.71-0.96) and for A+T against T was 1.02 (95% CI 0.89-1.18). Our asymptotically normal test statistics are based on the log HRs. Let D_i be the log HR for treatment i against control, so that $\hat{D}_1 = -\log 0.83$ and $\hat{D}_2 = -\log 1.02$ with $se(\hat{D}_1) = (\log 0.96 - \log 0.71)/4 = 0.07544$ and $se(\hat{D}_2) = (\log 1.18 - \log 0.89)/4 = 0.07051$. Then $Z_1 = \hat{D}_1/se(\hat{D}_1) = 2.4706$ and $Z_2 = \hat{D}_2/se(\hat{D}_2) = -0.2808$. Hence, the results are almost identical to those for three-year survival and, in particular, the decisions to reject H_0 or not are identical. This is as expected, since more than 85% of the observations were censored, and tests are known to be similar in these circumstances (Cuzick, 1982).

3.3 Comparing the test performances

The tests are compared in terms of power, an adjusted power function and a loss function. We use the standard definition of power, namely the probability that H_0 is rejected, given that it is false. As usual, the power depends on unknown parameters, here Δ_1 and Δ_2. Some authors use adjusted power functions and often refer to these simply as "power". In particular, in the multiple comparisons literature, e.g. Dunnett and Tamhane (1992), Horn and Vollandt (1998) and Horn and Dunnett (2004), it is common to use an adjusted power, which is the probability that at least one false $H_{0i} : \Delta_i \leq 0$ is rejected. This "any-pair power" is not quite relevant for our study, since we are interested in the overall $H_0 : \Delta_i \leq 0 \; \forall i$. In contrast, the literature on order restricted inference (Robertson et al., 1988) usually uses the standard definition of power, as we do here, since they do not consider the individual hypotheses. In fact, they typically do not consider the problem of selection, so the issues of type III and IV errors do not arise.

Here, we define the adjusted power function to be

$$\text{Adjusted power} \;=\; \text{power} \; - P(\text{type III error}) - P(\text{type IV error}),$$

where, as defined in Chapter 2, a type III error is the rejection of H_0 in favour of a treatment which is no better than the control, while a type IV error is rejection of H_0 in favour of a treatment which is better than the control, but worse than the best experimental treatment. Whereas Horn and Vollandt (1998) stated that their adjustment to the power made almost no difference, that does not follow here, since we consider type IV errors which are very common when Δ_2 is close to Δ_1, as well as type III errors which are rare when Δ_1 is reasonably large.

Although it seems reasonable to take account of the probability of incorrect selection, the adjusted power might not be the best way of doing so. It treats the selection of a treatment which is only very slightly worse than the best in the same way as selection of a treatment which is much worse than the control. It seems more appropriate to use a function which takes account of the impact of incorrect selection as well as its probability.

We define the loss as

$$\text{Loss} \;=\; \begin{cases} 0 & \text{if we select treatment 1;} \\ \Delta_1 - \Delta_2 & \text{if we select treatment 2;} \\ \Delta_1 & \text{if we fail to reject } H_0. \end{cases}$$

We compare the different tests in terms of the expected loss, calculated as

$$\begin{aligned} E(\text{loss}) \;&=\; 0 \times (\text{adjusted power}) + (\Delta_1 - \Delta_2) \times P(\text{type III/IV error}) + \Delta_1 \times (1 - \text{power}) \\ &=\; (\Delta_1 - \Delta_2) \times P(\text{type III/IV error}) + \Delta_1 \times P(\text{type II error}). \end{aligned}$$

Note that the expectation is taken over the sampling distribution of Z_1 and Z_2 and is a function of the unknown parameters Δ_1 and Δ_2. No use is being made here of prior distributions.

The simulation program generates 100,000 values of bivariate normal variables W_1 and W_2 with expectations zero, unit variances and correlation 0.5, in the same way as Z_1 and Z_2 were simulated under the null distribution, in order to create simulated data. The correlation was set to be $\rho = 0.5$ which corresponds to equal numbers of patients in each arm, as in Section 3.1. Observed values of Z_1 and Z_2 are calculated by using $w_{1i} + (\Delta_1/\sigma)$ and $w_{2i} + (\Delta_2/\sigma)$ respectively, for each simulation $1 = 1, \ldots, 100,000$, i.e. the same set of values w_1 and w_2 were used for each different pair of values of Δ_1 and Δ_2. Hence the same simulated data were used for each significance level α. We define $d_i = \Delta_i/\sigma$ for $i = 1, 2$. d_1 ranged from 0 to 4 and d_2 from 0 to d_1.

The null hypothesis is rejected if the value of the test statistic was greater than the critical value given in Section 3.1.4 for the corresponding test. The power is estimated by the proportion of 100,000 simulations which rejected H_0. For $\Delta_1 > \Delta_2$, the probability of a type III or IV error is $P(Z_1 < Z_2 \bigcap \text{ reject } H_0)$. For the values with $\Delta_1 = \Delta_2$, type IV error is not relevant and this calculation would give simply $\frac{1}{2} \times$ power. The approximation from the simulation is

$$\begin{aligned} P(\text{type III or IV error}) &= P(\text{test statistic} > C \bigcap Z_1 < Z_2) \\ &= \frac{\text{number of } (Z_1 < Z_2) \bigcap \text{reject } H_0}{100,000}. \end{aligned}$$

From this, we can easily estimate the adjusted power and the expected loss.

Ignoring simulation error in the cutpoints used, the estimated power from the simulations has standard error equal to $\sqrt{p(1-p)}/\sqrt{100,000}$, where p is the true power. Hence, when expressed as a percentage, as in the rest of this thesis, the standard errors of powers are approximately 0.158, 0.126 and 0.069 for 50%, 80% and 95% power respectively. Thus the results presented can be assumed to be fairly precise to about 0.25%. The standard error of a difference between two independently simulated powers would be $\sqrt{2}$ times the standard errors given above. However, because we have used the same simulated errors, our powers are positively correlated, which reduces the standard error of a difference. It is not easy to calculate what it should be, but we can be confident that any differences greater than about 0.4% power are probably real and not due to simulation variation.

Results for power and expected loss are shown for size 5%, 2.5% and 1%. We also show some results for the adjusted power at size 5%, although we believe that it is less useful, since it does not take account of the differences between treatments.

3.3.1 Power results

In the tables in this section, we present only configurations which, for tests at the 5% level, give powers in the range 65 – 95%. These are typically the types of powers of interest. More complete results are presented in the figures. Note that the tests used here are not unbiased. A test of size α is said to be *unbiased* if $P(\text{Reject } H_0|\mathbf{\Delta}) \geq \alpha$ for all $\mathbf{\Delta}$ such that H_1 is true. This is obviously not true here. Consider $\Delta_1 = \varepsilon_1$, for some small positive constant ε_1 and $\Delta_2 \to -\infty$. Then T_k and T_k^0 will equal Z_1 and S_1^0 will tend to $-\infty$. Hence the upper 5% point of the distribution will be $1.6465 + \varepsilon_2$, for some small constant ε_2 and from the cutpoints in Table 3.2, we will reject H_0 less than 5% of the time. However, unbiasedness of tests is not generally considered to be a very important property, since it concerns parameter values which, though in the set defined by H_1, are very close to the set defined by H_0.

The power at 5% for each of the tests considered is shown in Table 3.3 for various configurations $d_i = \Delta_i/\sigma$ for $i = 1, 2$. Further results are illustrated in Figure 3.5. Note that in this and subsequent figures, T_∞ is referred to as T_3 and Hochberg's procedure is referred to as Hoch. The first point to note is that the differences between the tests are not very large, except that S_1^0 and T_1^0 have much lower power than the others when $d_2 = 0$. In Figure 3.5, T_2, T_∞ and Hochberg are indistinguishable. In Table 3.3, we see that when $d_2 = 0$, T_∞ is slightly better than Hochberg, which is very slightly better than T_2. T_2^0 and T_1 are a little worse than these three, and about equal to each other. As seen from the top left panel in Figure 3.5, T_1^0 and S_1^0 are vastly inferior to the others when $d_2 = 0$. When $d_2 = d_1/2$, T_2 has the highest power, but is only very slightly better than Hochberg, which is very slightly better than T_∞. T_2^0 is slightly worse and T_1 slightly worse again. Again, T_1^0 and S_1^0 are somewhat worse than the other tests (upper right panel in Figure 3.5). When $d_2 = d_1$, S_1^0 is the most powerful among these tests, being very slightly better than T_1^0, which is very slightly better than T_2^0 and T_1. T_2 is slightly worse than these, but very slightly better than Hochberg, which is slightly better than T_∞ (lower panel of Figure 3.5).

Some results at 5% with negative Δ_2 are shown in Table 3.4. Comparing these with the results in Table 3.3 for $\Delta_2 = 0$, we see that for T_1, T_2, T_∞ and Hochberg's procedure, the power is almost the same. For T_1^0 and T_2^0, it is somewhat reduced and for S_1^0 it is greatly reduced. This illustrates an important advantage of the T_k family of test statistics: because they only include contributions from the second largest treatment effect when it is at least half as big as the biggest ($\rho = 1/2$), they are not much affected by changes in

Table 3.3: Power of the tests at 5% for selected values of d_1 and d_2, where $d_i = \Delta_i/\sigma$.

d_1	d_2	T_1	T_2	T_∞	T_1^0	T_2^0	S_1^0	Hoch
2.5	0	64.691	71.169	72.242	47.651	65.089	42.094	71.677
2.7	0	71.587	77.508	78.532	53.813	71.829	46.693	77.937
2.9	0	77.798	83.113	84.015	60.187	77.943	51.309	83.491
3	0	80.653	85.477	86.330	63.205	80.705	53.625	85.822
3.5	0	91.376	93.956	94.398	77.787	91.299	64.820	94.138
2.5	1.25	72.941	74.586	74.349	70.345	73.859	70.020	74.537
2.7	1.35	78.788	80.402	80.249	76.199	79.627	75.800	80.330
2.9	1.45	83.886	85.425	85.353	81.308	84.682	80.879	85.399
3	1.5	86.084	87.579	87.490	83.581	86.863	83.133	87.502
3.5	1.75	94.094	94.926	94.909	92.260	94.494	91.881	94.897
2.	2	73.814	71.518	70.028	74.893	73.824	74.917	70.982
2.5	2.5	88.688	87.151	85.937	89.463	88.782	89.466	86.714
2.7	2.7	92.540	91.300	90.259	93.136	92.547	93.148	90.950
2.9	2.9	95.253	94.381	93.613	95.657	95.322	95.661	94.106
3.	3	96.247	95.565	94.877	96.565	96.269	96.564	95.320

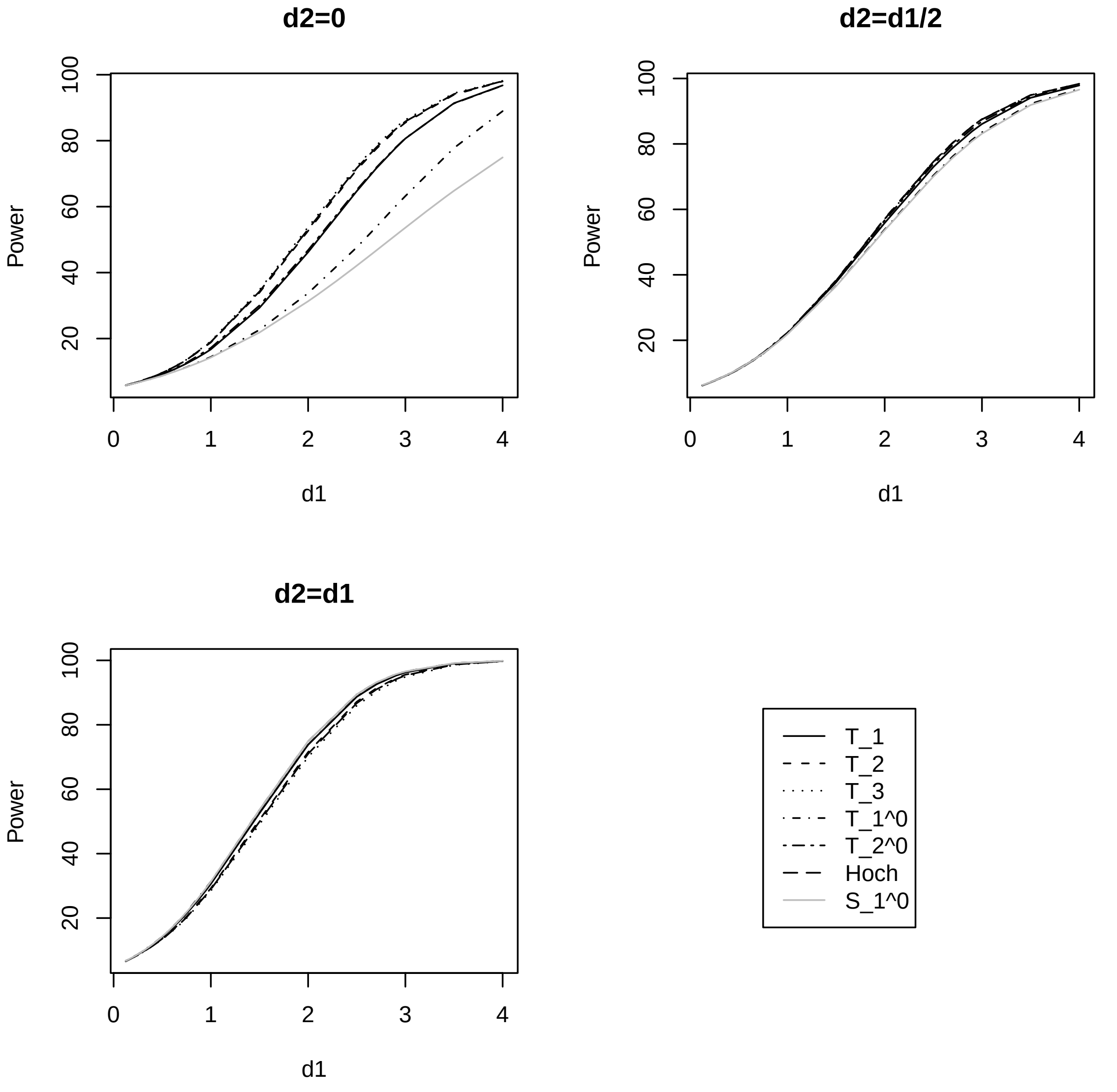

Figure 3.5: Power at 5%. $d_i = \Delta_i/\sigma$ for $i = 1, 2$.

Table 3.4: Power of the tests at 5% when $d_2 < 0$.

d_1	d_2	T_1	T_2	T_∞	T_1^0	T_2^0	S_1^0	Hoch
2.5	-0.625	63.920	70.976	72.181	40.841	63.726	28.817	71.554
3	-0.75	80.360	85.422	86.315	57.986	80.049	36.510	85.790
3.5	-0.875	91.296	93.945	94.396	74.919	91.098	44.975	94.135
2.5	-1.25	63.759	70.946	72.173	37.735	63.345	17.793	71.535
3	-1.5	80.334	85.419	86.313	56.434	79.952	21.843	85.787
3.5	-1.75	91.294	93.945	94.396	74.410	91.070	26.358	94.135

the size of the smaller effect close to zero.

The power of the tests at the 2.5% level of significance for various configurations of d_1 and d_2 are shown in Table 3.5 and plotted, for a larger set of values, in Figure 3.6. The plots show a very similar pattern to those from the tests at 5%, but with slightly greater differences between the tests. The ordering of the test statistics in terms of power are exactly the same as for the tests at 5%, although as noted from the graphs the differences are slightly greater. Some results with negative Δ_2 are shown in Table 3.6 and, as with the 5% significance level, they show that the results for the T_k family and Hochberg's test are almost the same as when $\Delta_2 = 0$, while the other test statistics are adversely affected.

The corresponding results for testing at the 1% level of significance are given in Table 3.7 and Figure 3.7. Again these results are broadly similar to those at the other significance levels. In a few cases the ordering of tests changes slightly, but this is probably just due to simulation variation, which we noted is greater at 1% than at the other significance levels. The results with negative values of Δ_2, shown in Table 3.8, again show that T_1, T_2, T_∞ and Hochberg's test are not greatly affected by whether the value of Δ_2 is zero or negative.

Our results are consistent with those of Mukerjee et al. (1985), also briefly summarised by Robertson et al. (1988), which showed that Dunnett's procedure was more powerful than the LRT when $\Delta_2 = 0$, but not towards the middle of the rejection region, i.e. close to $\Delta_2 = \Delta_1$. It is worth noting some differences between our simulation study and theirs. First, the simulation study reported here is more comprehensive, covering more testing procedures and more different values of Δ_1 and Δ_2. Secondly, we have used the same simulation program to generate the cutpoints for all the test procedures, whereas they used different numerical approximation procedures for the different test statistics. Hence,

Table 3.5: Power of the tests at 2.5%.

d_1	d_2	T_1	T_2	T_∞	T_1^0	T_2^0	S_1^0	Hoch
2.7	0	57.790	67.445	68.996	37.710	58.914	34.492	68.483
2.9	0	65.240	74.225	75.647	43.295	66.205	38.776	75.191
3	0	68.767	77.348	78.625	46.267	69.613	41.063	78.205
3.5	0	83.828	89.460	90.294	61.856	84.167	52.587	90.037
2.5	1.25	60.802	63.821	63.574	58.392	62.593	58.281	63.673
2.7	1.35	67.575	70.871	70.726	65.033	69.444	64.904	70.744
2.9	1.45	73.965	77.119	77.026	71.266	75.732	71.109	77.060
3	1.5	76.920	79.988	79.871	74.209	78.635	74.044	79.914
3.5	1.75	88.696	90.888	90.916	86.092	89.866	85.865	90.901
2	2	62.825	60.075	58.011	63.881	62.951	63.888	58.835
2.5	2.5	81.649	79.358	77.255	82.460	81.705	82.461	78.168
2.7	2.7	87.109	85.184	83.486	87.787	87.182	87.786	84.189
2.9	2.9	91.355	89.811	88.366	91.911	91.420	91.912	88.967
3	3	93.077	91.699	90.340	93.564	93.076	93.564	90.909

Table 3.6: Power of the tests at 2.5% when $d_2 < 0$.

d_1	d_2	T_1	T_2	T_∞	T_1^0	T_2^0	S_1^0	Hoch
3	-0.75	68.377	77.284	78.612	38.331	68.527	25.452	78.178
3.5	-0.875	83.696	89.441	90.292	55.940	83.704	32.954	90.030
3	-1.5	68.330	77.274	78.610	35.600	68.321	13.716	78.175
3.5	-1.75	83.688	89.441	90.292	54.522	83.654	17.143	90.030

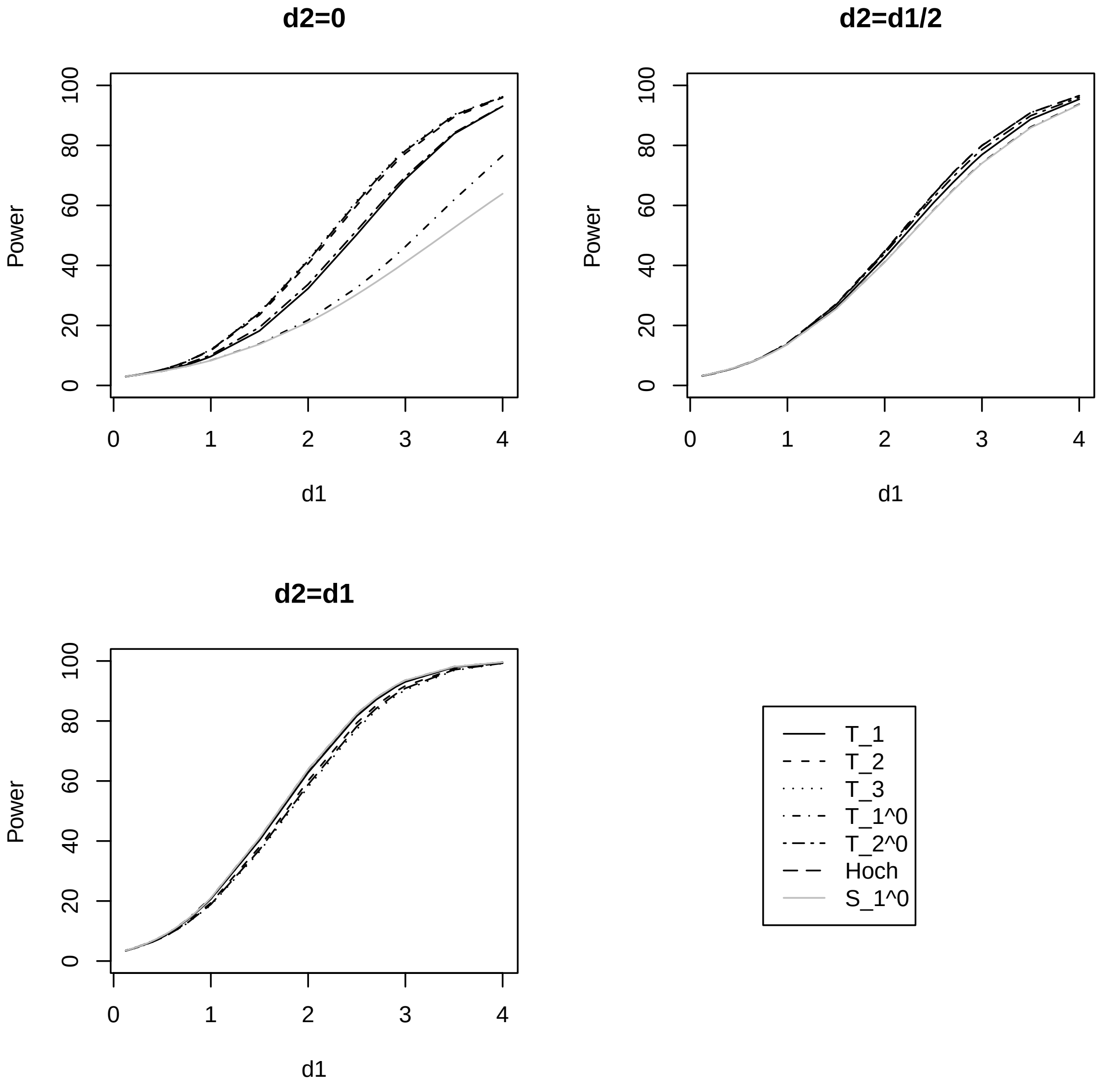

Figure 3.6: Power at 2.5%. $d_i = \Delta_i/\sigma$ for $i = 1, 2$.

Table 3.7: Power of the tests at 1%.

d_1	d_2	T_1	T_2	T_∞	T_1^0	T_2^0	S_1^0	Hoch
2.9	0	47.556	61.631	63.593	27.126	50.150	25.767	63.294
3	0	51.509	65.451	67.313	29.377	53.984	27.701	67.042
3.5	0	70.162	81.522	82.960	42.296	71.804	38.014	82.732
2.9	1.45	59.170	64.969	64.886	57.528	62.694	57.495	64.980
3	1.5	62.712	68.576	68.493	60.952	66.185	60.918	68.603
3.5	1.75	78.284	83.525	83.644	76.187	81.398	76.128	83.649
2.5	2.5	70.575	67.465	64.408	71.410	70.668	71.411	65.201
2.7	2.7	77.951	75.087	72.174	78.749	78.119	78.749	72.932
2.9	2.9	84.177	81.711	79.086	84.731	84.272	84.731	79.863
3	3	86.750	84.595	82.177	87.345	86.883	87.345	82.879
3.5	3.5	95.438	94.390	93.029	95.754	95.537	95.754	93.440

Table 3.8: Power of the tests at 1% when $d_2 < 0$.

d_1	d_2	T_1	T_2	T_∞	T_1^0	T_2^0	S_1^0	Hoch
3	-0.75	50.950	65.349	67.302	20.033	52.348	15.194	67.019
3.5	-0.875	69.951	81.500	82.957	33.164	70.943	20.875	82.727
3	-1.5	50.894	65.347	67.302	16.310	51.950	7.220	67.019
3.5	-1.75	69.937	81.498	82.957	30.547	70.812	9.394	82.726

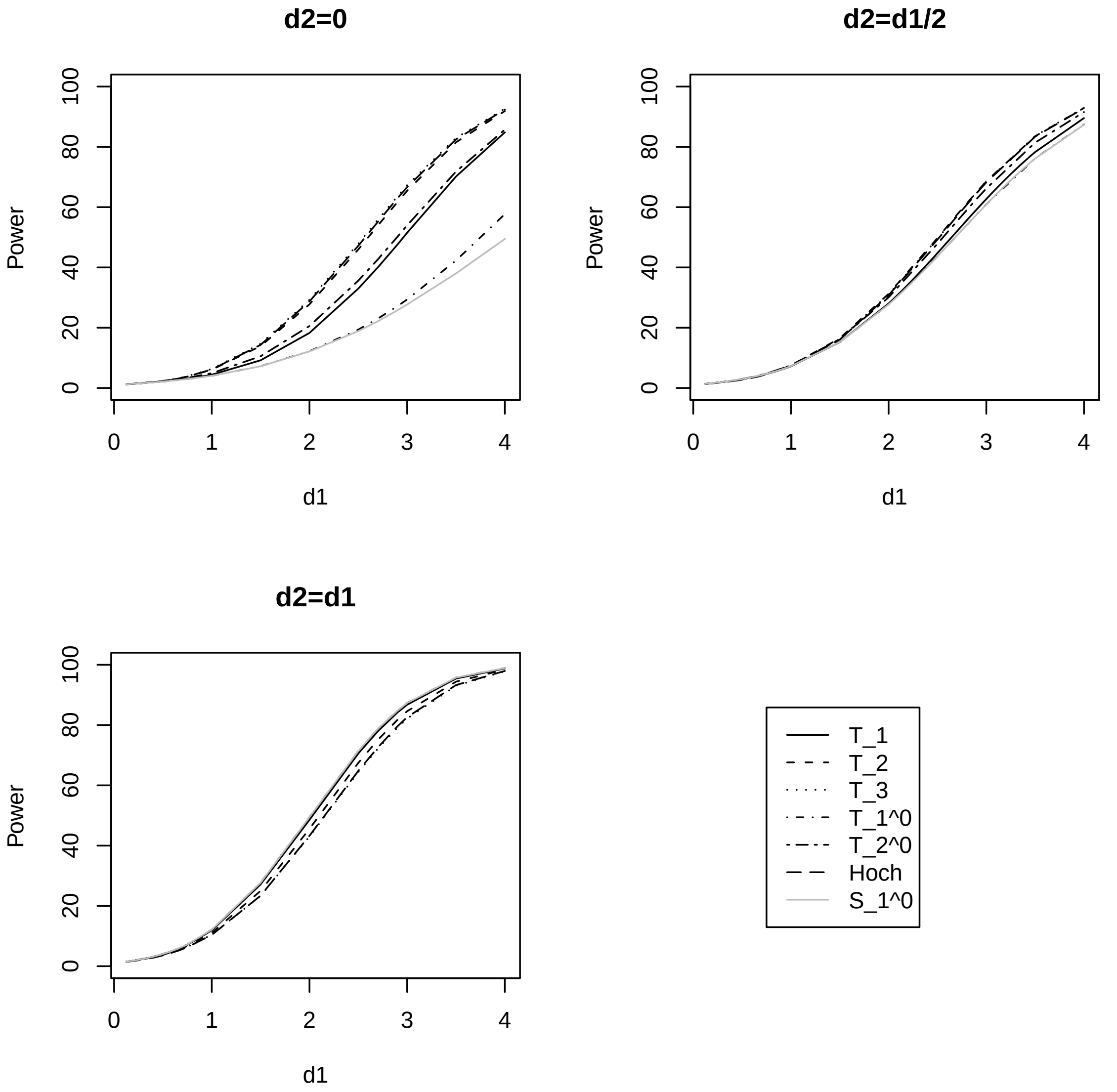

Figure 3.7: Power at 1%.

Table 3.9: Power of the tests using exact cutpoints for T_∞ and S_1^0 at 5, 2.5 and 1%.

		5%		2.5%		1%	
d_1	d_2	T_∞	S_1^0	T_∞	S_1^0	T_∞	S_1^0
2	0	53.820	31.307	41.977	21.043	29.263	12.121
2.5	0	72.243	42.091	61.564	30.361	47.887	18.876
3	0	86.330	53.620	78.624	41.063	67.311	27.696
3.5	0	94.398	64.819	90.294	52.587	82.957	38.010
2	1	56.897	53.620	44.450	41.063	31.045	27.696
2.5	1.25	74.350	70.014	63.573	58.281	49.509	43.689
3	1.5	87.490	83.133	79.870	74.044	68.490	60.912
3.5	1.75	94.909	91.881	90.916	85.865	83.641	76.124
2	2	70.028	74.914	58.009	63.888	43.016	49.446
2.5	2.5	85.938	89.465	77.255	82.461	64.404	71.405
3	3	94.877	96.564	90.338	93.564	82.174	87.340

our numerical results are not identical to those of Mukerjee et al. (1985), but the overall conclusions are consistent with theirs. Our results are also broadly consistent with many in the multiple comparisons literature, e.g. Horn and Dunnett (2004), who found that Dunnett's procedure, Hochberg's procedure and several other methods have similar any-pair powers. Again, exact comparisons of the results are not possible, due to the slightly different objectives of the tests.

3.3.2 Effect of estimating the cutpoints

We noted in Section 3.1.4 that, for those test statistics for which exact distributions are known, the cutpoints estimated from the simulations were very close to those calculated from the exact distributions. We now check to see what impact they have on the estimated powers. These are shown in Table 3.9. Comparing these with the results in Tables 3.3, 3.5 and 3.7, we find very little difference between them, in many cases no difference. Even in the worst case, S_1^0 at 1%, the powers are precise to two decimal places. This gives us additional confidence in the precision of our cutpoints. As noted above, even if the cutpoints were exact, the simulation variance in the powers would be greater than the differences we have found here, so that this is clearly a more important source of imprecision than the estimated cutpoints.

Table 3.10: Adjusted power of the tests at 5% when $\Delta_2 = 0$.

d_1	d_2	T_1	T_2	T_∞	T_1^0	T_2^0	S_1^0	Hoch
2.5	0	64.429	70.959	72.060	47.385	64.845	41.828	71.471
2.7	0	71.419	77.377	78.415	53.648	71.681	46.527	77.810
2.9	0	77.694	83.036	83.948	60.085	77.853	51.207	83.417
3	0	80.576	85.421	86.280	63.131	80.636	53.551	85.769
3.5	0	91.365	93.946	94.388	77.775	91.287	64.808	94.128

Table 3.11: Adjusted power of the tests at 5% when $\Delta_2 = \Delta_1/2$

d_1	d_2	T_1	T_2	T_∞	T_1^0	T_2^0	S_1^0	Hoch
2.5	1.25	65.730	68.057	68.162	62.995	66.864	62.668	68.099
2.7	1.35	72.232	74.394	74.534	69.546	73.211	69.146	74.388
2.9	1.45	78.081	80.016	80.194	75.436	78.986	75.003	80.043
3	1.5	80.642	82.490	82.611	78.083	81.523	77.633	82.458
3.5	1.75	90.476	91.451	91.544	88.622	90.919	88.243	91.449

3.3.3 Adjusted power results

Table 3.10 shows the adjusted powers at the 5% level for the same sets of values as the powers in Table 3.3 above with $\Delta_2 = 0$, when type-III errors could influence the result. The adjusted power is very similar to the power, since in this case type III errors are very rare. This is reassuring in the sense that, when H_0 is rejected, it is, with high probability, in favour of a treatment which is better than the control.

When $\Delta_2 = \Delta_1/2$, as shown in Table 3.11 the adjustment based on the probability of type-IV error makes some difference. In terms of this adjusted power function, T_∞ seems better, relative to T_2, than in terms of power. Again, though, T_2 and Hochberg are very close to T_∞.

Note that when $\Delta_2 = \Delta_1$ the adjusted power is identical to the power, since no type III or IV error is possible in this case, and so it is not shown.

Although it is straightforward to produce the adjusted powers for all simulations, we do not present any more of these, because we do not believe that the adjusted power (as defined here) is a useful concept. For example, if Δ_2 is very close, but not equal, to Δ_1, then the probability of type IV error approaches 1/2 and the adjusted power then drops to almost half of the power. There is then a discontinuity when $\Delta_2 = \Delta_1$, unless we arbitrarily define treatment 1 to be optimal and count the selection of treatment 2 as a type IV error in the latter case. However, in practice this is meaningless, since the choice of a treatment which is equally as good as the best (or even very close) cannot sensibly be defined as an error. Given the power and the expected loss, the adjusted power does not seem to add anything. Hence, we do not pursue the adjusted power further in this thesis.

3.3.4 Expected loss results

Tables 3.12 to 3.14 show the expected loss for each of the test statistics considered, for the same set of values of Δ_1 and Δ_2 as were given for the power. In these tables, we also show a reasonable upper bound on the expected loss, which is what we would get from randomly rejecting H_0 with probability α and then randomly selecting a treatment. This is calculated as

$$E(\text{loss}) = (1-\alpha)\Delta_1 + \frac{\alpha}{I}\sum_{i=2}^{I}(\Delta_1 - \Delta_i).$$

These and additional results over a wider set of values of d_1 and d_2 are presented graphically in Figures 3.8 to 3.10.

The first thing we note from the figures is that, across the values of d_1 and d_2, the most difficult cases for the tests, i.e. those that give the highest expected loss, seem to be when $\Delta_2 = 0$ and/or $d_1 \approx 1.5$, as can be seen in the figures. The former is consistent with the power being lowest for these configurations, while the latter is interesting, if not completely surprising. Smaller values of Δ_1 mean that type II and type III/IV errors have high probability, but the consequences of any wrong decision are less important, so the expected loss is small. Larger values of Δ_1 mean that type II errors are very unlikely and type III/IV errors are either very unlikely (if Δ_2 is small) or have unimportant consequences (if Δ_2 is close to Δ_1), so again the expected loss is small.

The ordering of the test procedures is almost exactly the same as for power. Note, however, that it is not identical, e.g. for tests at the 2.5% level, when $\Delta_1 = 2.9$ and $\Delta_2 = 1.45$, T_2 has higher power than Hochnerg's procedure, but higher expected loss. Overall, T_∞ is best when $\Delta_2 = 0$, being slightly better than the Hochberg procedure and T_2. T_2, T_∞ and Hochberg are about equally good when $\Delta_2 = \Delta_1/2$. Among these three tests, T_2 is clearly best when $\Delta_2 = \Delta_1$, but in this case S_1^0 is best overall.

Expected losses for negative values of Δ_2 are shown in Tables 3.15 to 3.17 for 5%, 2.5%

Table 3.12: Expected loss at 5%.

d_1	d_2	Bound	T_1	T_2	T_∞	T_1^0	T_2^0	S_1^0	Hoch
2.5	0	2.4375	0.8893	0.7260	0.6985	1.3154	0.8789	1.4543	0.7132
2.7	0	2.6325	0.7717	0.6108	0.5828	1.2515	0.7646	1.4438	0.5991
2.9	0	2.8275	0.6469	0.4920	0.4655	1.1575	0.6423	1.4150	0.4809
3	0	2.925	0.5827	0.4374	0.4116	1.1060	0.5809	1.3935	0.4269
3.5	0	3.4125	0.3022	0.2119	0.1964	0.7779	0.3050	1.2317	0.2055
2.5	1.25	2.4063	0.7666	0.7170	0.7186	0.8332	0.7410	0.8414	0.7170
2.7	1.35	2.5988	0.6612	0.6103	0.6104	0.7324	0.6367	0.7432	0.6113
2.9	1.45	2.7913	0.5515	0.5011	0.4996	0.6272	0.5268	0.6397	0.5011
3	1.5	2.8875	0.4991	0.4490	0.4485	0.5750	0.4742	0.5885	0.4506
3.5	1.75	3.3688	0.2700	0.2384	0.2371	0.3346	0.2553	0.3478	0.2389
2	2	1.9000	0.5237	0.5696	0.5994	0.5021	0.5235	0.5016	0.5804
2.5	2.5	2.3750	0.2828	0.3212	0.3516	0.2634	0.2804	0.2634	0.3322
2.7	2.7	2.5650	0.2014	0.2349	0.2630	0.1853	0.2012	0.1850	0.2443
2.9	2.9	2.7550	0.1377	0.1630	0.1852	0.1259	0.1357	0.1258	0.1709
3	3	2.8500	0.1126	0.1330	0.1537	0.1030	0.1119	0.1031	0.1404

Table 3.13: Expected loss at 2.5%.

d_1	d_2	Bound	T_1	T_2	T_∞	T_1^0	T_2^0	S_1^0	Hoch
2.7	0	2.6663	1.1432	0.8816	0.8392	1.6852	1.1123	1.7721	0.8534
2.9	0	2.8638	1.0102	0.7491	0.7075	1.6466	0.9820	1.7776	0.7210
3	0	2.9625	0.9387	0.6808	0.6423	1.6136	0.9131	1.7697	0.6550
3.5	0	3.4563	0.5664	0.3692	0.3400	1.3354	0.5545	1.6598	0.3490
2.5	1.25	2.4531	1.0565	0.9705	0.9706	1.1171	1.0081	1.1199	0.9711
2.7	1.35	2.6494	0.9528	0.8540	0.8526	1.0212	0.8990	1.0247	0.8547
2.9	1.45	2.8456	0.8303	0.7303	0.7279	0.9086	0.7760	0.9131	0.7295
3	1.5	2.9438	0.7660	0.6664	0.6652	0.8474	0.7122	0.8523	0.6663
3.5	1.75	3.4344	0.4553	0.3743	0.3706	0.5462	0.4130	0.5542	0.3725
2	2	1.9500	0.7435	0.7985	0.8398	0.7224	0.7410	0.7222	0.8233
2.5	2.5	2.4375	0.4588	0.5161	0.5686	0.4385	0.4574	0.4385	0.5458
2.7	2.7	2.6325	0.3481	0.4000	0.4459	0.3298	0.3461	0.3298	0.4269
2.9	2.9	2.8275	0.2507	0.2955	0.3374	0.2346	0.2488	0.2346	0.3200
3	3	2.9250	0.2077	0.2490	0.2898	0.1931	0.2077	0.1931	0.2727

Table 3.14: Expected loss at 1% .

d_1	d_2	Bound	T_1	T_2	T_∞	T_1^0	T_2^0	S_1^0	Hoch
2.9	0	2.8855	1.5224	1.1138	1.0565	2.1148	1.4469	2.1542	1.0652
3	0	2.985	1.4559	1.0374	0.9812	2.1198	1.3815	2.1701	0.9895
3.5	0	3.4825	1.0446	0.6470	0.5966	2.020	0.9871	2.1698	0.6046
2.9	1.45	2.8783	1.2463	1.0678	1.0635	1.2929	1.1405	1.2939	1.0629
3	1.5	2.9775	1.1808	0.9955	0.9911	1.2327	1.0732	1.2338	0.9901
3.5	1.75	3.4738	0.8133	0.6240	0.6157	0.8863	0.7025	0.8884	0.6170
2.5	2.5	2.4750	0.7356	0.8134	0.8898	0.7148	0.7333	0.7147	0.8700
2.7	2.7	2.6730	0.5953	0.6727	0.7513	0.5738	0.5908	0.5738	0.7308
2.9	2.9	2.8710	0.4589	0.5304	0.6065	0.4428	0.4561	0.4428	0.5840
3	3	2.9700	0.3975	0.4622	0.5347	0.3797	0.3935	0.3797	0.5136
3.5	3.5	3.4650	0.1597	0.1964	0.2440	0.1486	0.1562	0.1486	0.2296

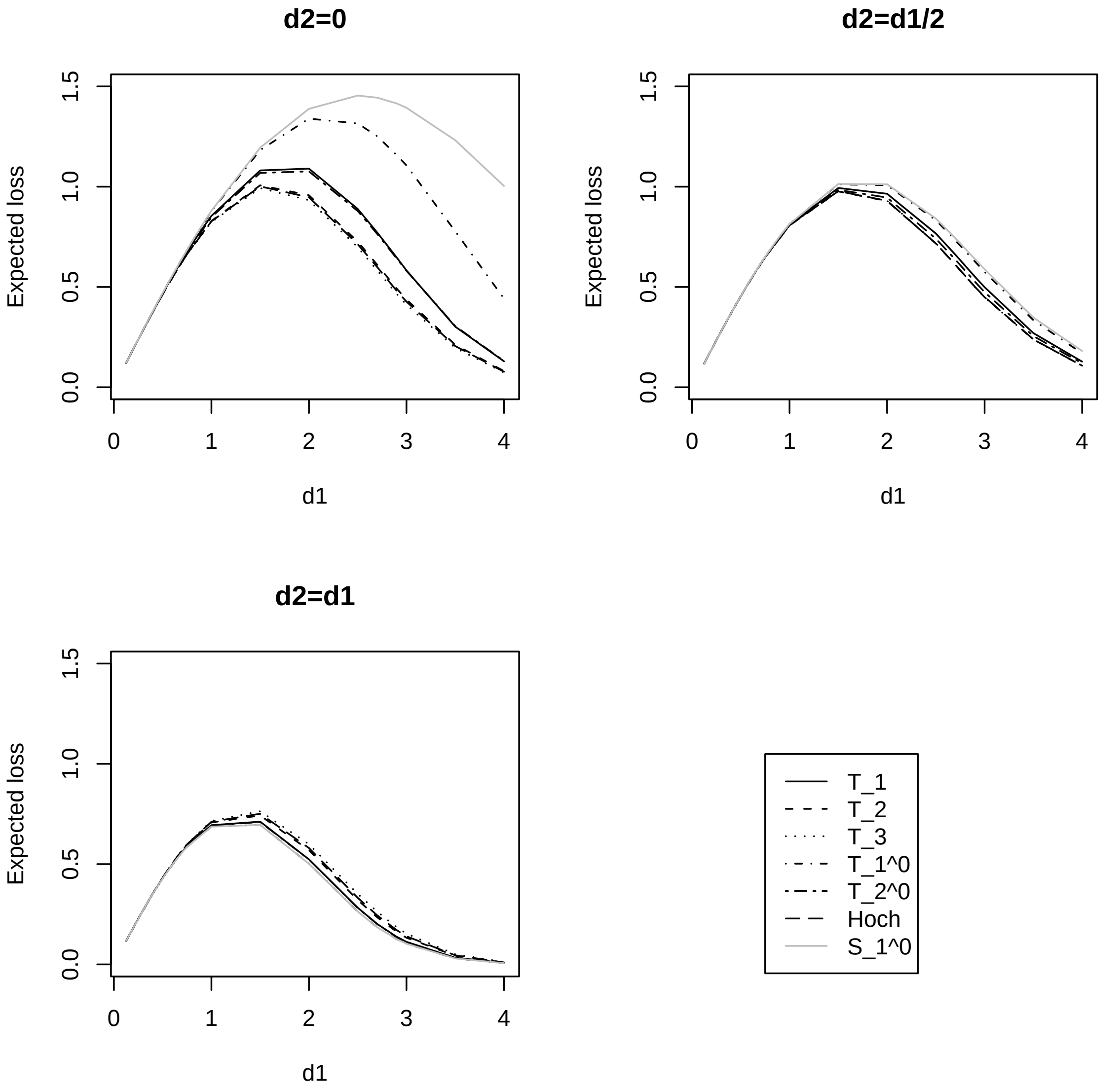

Figure 3.8: Expected loss at 5%.

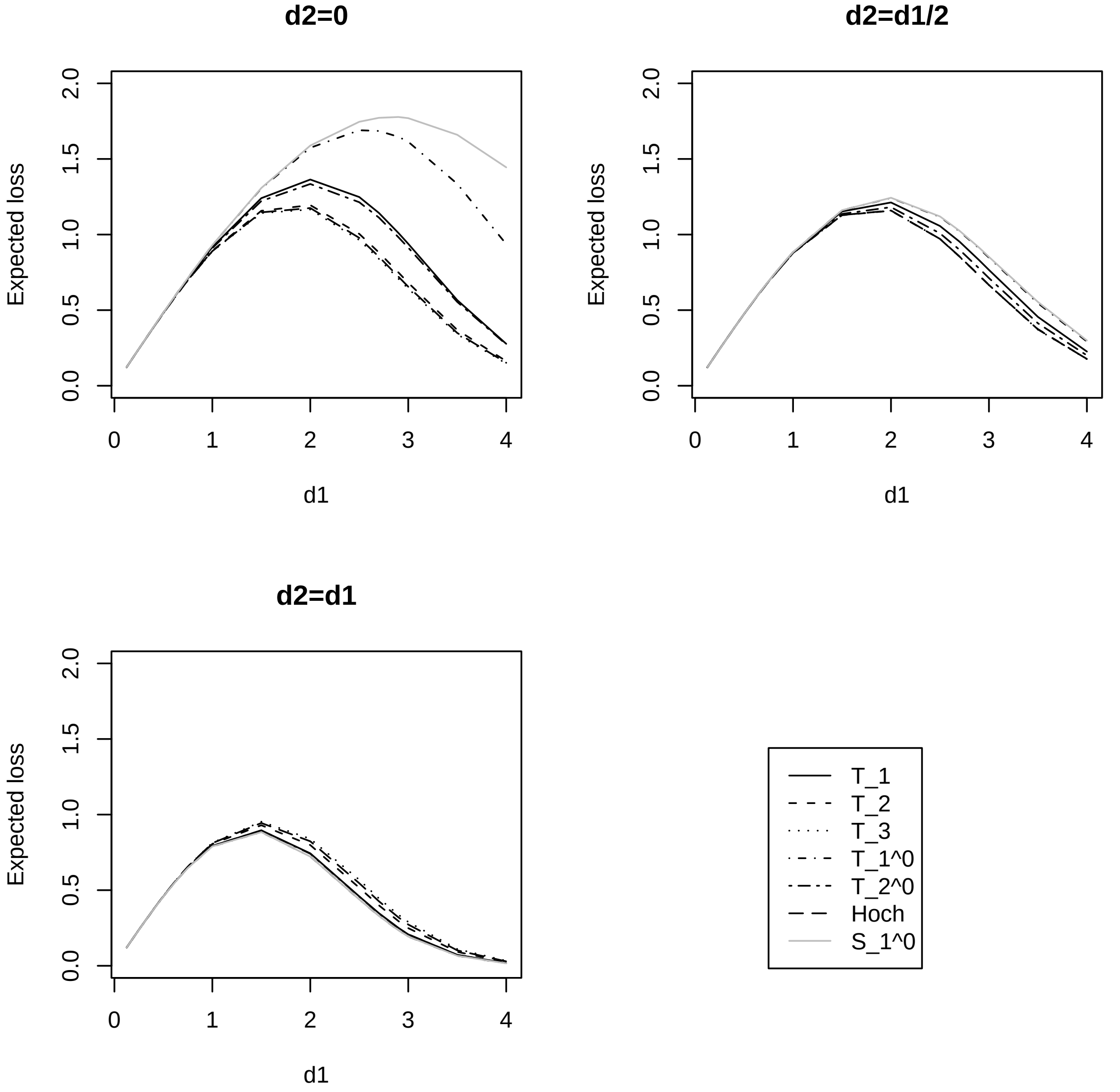

Figure 3.9: Expected loss at 2.5%.

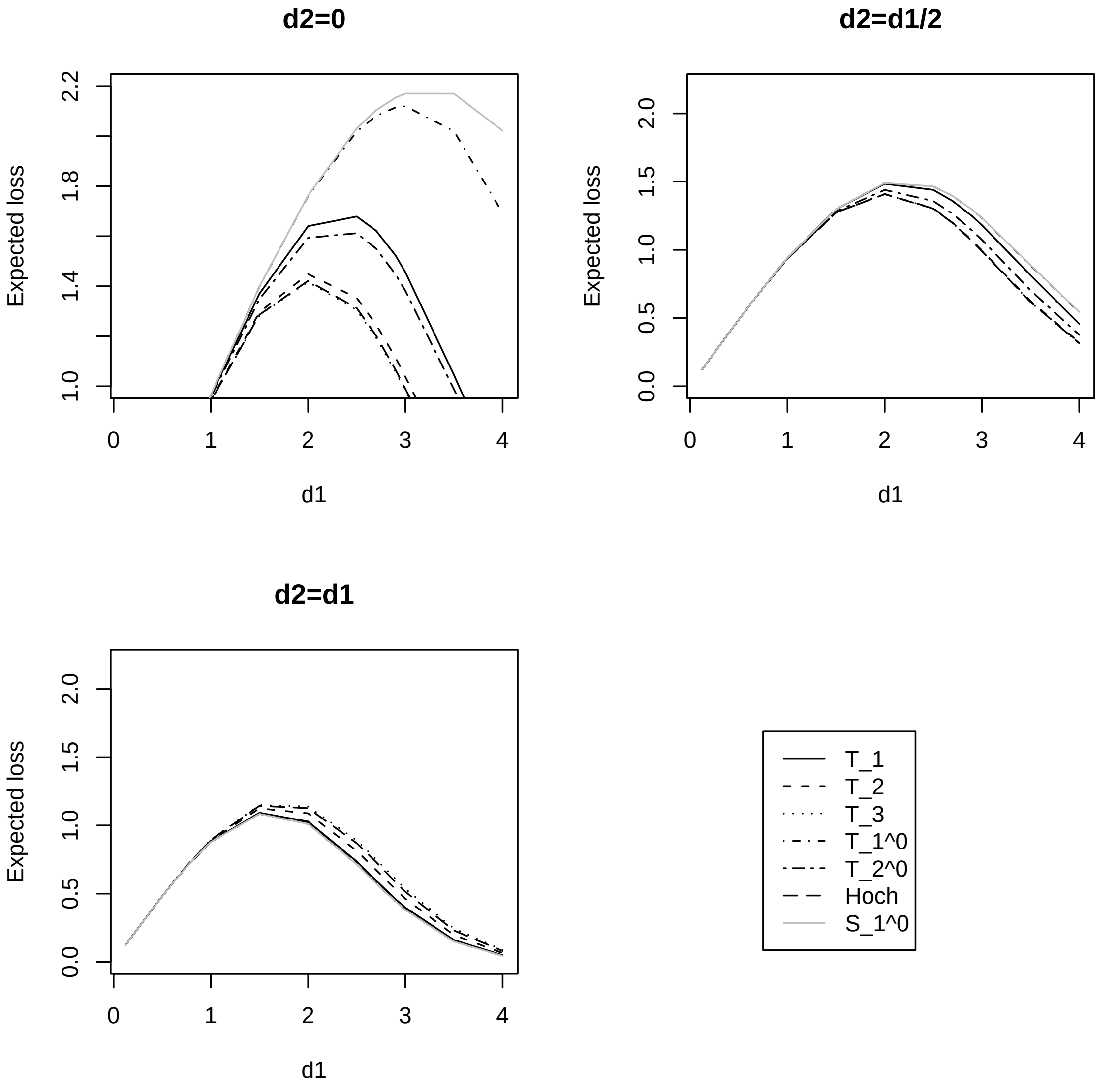

Figure 3.10: Expected loss at 1%.

Table 3.15: Expected loss at 5% when $d_2 < 0$.

d_1	d_2	Bound	T_1	T_2	T_∞	T_1^0	T_2^0	S_1^0	Hoch
2.5	-0.625	2.4531	0.9028	0.7262	0.6959	1.4798	0.9075	1.7804	0.7117
3	-0.75	2.9438	0.5893	0.4375	0.4107	1.2605	0.5986	1.9048	0.4264
3.5	-0.875	3.4344	0.3047	0.2120	0.1962	0.8779	0.3116	1.9259	0.2053
2.5	-1.25	2.4688	0.9061	0.7264	0.6957	1.5567	0.9165	2.0553	0.7117
3	-1.5	2.9625	0.5900	0.4374	0.4106	1.3070	0.6014	2.3447	0.4264
3.5	-1.75	3.4563	0.3047	0.2119	0.1961	0.8957	0.3126	2.5775	0.2053

Table 3.16: Expected loss at 2.5% when $d_2 < 0$.

d_1	d_2	Bound	T_1	T_2	T_∞	T_1^0	T_2^0	S_1^0	Hoch
3	-0.75	2.9719	0.9488	0.6816	0.6417	1.8502	0.9443	2.2366	0.6547
3.5	-0.875	3.4672	0.5707	0.3696	0.3398	1.5421	0.5704	2.3467	0.3490
3	-1.5	2.9813	0.9501	0.6818	0.6417	1.9320	0.9504	2.5885	0.6548
3.5	-1.75	3.4781	0.5709	0.3696	0.3398	1.5917	0.5721	2.9000	0.3490

and 1% respectively. The overall pattern is similar to that for power, in that T_1, T_2, T_∞ and Hochberg's show similar results as with $\Delta_2 = 0$, while the other tests perform poorly. However, an interesting feature is that both T_∞ and Hochberg's procedure actually show smaller expected loss with negative Δ_2. For T_2 the expected loss is larger when $\Delta_2 = -\Delta_1/4$ than when $\Delta_2 = 0$, but in some cases then decreases again when $\Delta_2 = -\Delta_1/2$, especially at the 5% level. Since the power is lower with negative Δ_2, this is clearly due to the reduced probability of type-III errors. This shows yet another advantage of these three test statistics over the others. It also shows the extra subtlety of the expected loss over the power as a measure for comparing tests.

Table 3.17: Expected loss at 1% when $d_2 < 0$.

d_1	d_2	Bound	T_1	T_2	T_∞	T_1^0	T_2^0	S_1^0	Hoch
3	-0.75	2.9888	1.4716	1.0395	0.9809	2.3991	1.4296	2.5443	0.9894
3.5	-0.875	3.4869	1.0518	0.6475	0.5965	2.3393	1.0170	2.7694	0.6046
3	-1.5	2.9925	1.4732	1.0396	0.9809	2.5107	1.4415	2.7834	0.9894
3.5	-1.75	3.4913	1.0522	0.6476	0.5965	2.4309	1.0216	3.1712	0.6046

Over all configurations studied, T_∞ has the lowest maximum loss ($= 0.9929$), when testing at 5%. The next lowest maximum loss is for Hochberg ($= 1.0015$) followed by T_2 ($= 1.0069$). These maxima all occur at $d_1 \approx 1.5$ and $d_2 = 0$, so are not shown in the tables above. The expected loss from T_2 is never more than 0.06 higher than for T_∞, while the expected loss for T_∞ can be more than 0.07, but never more than 0.08 higher than for T_2. Hochberg's procedure's expected loss is always between these two, but closer to T_∞'s, and is never more than 0.06 higher than for T_2.

3.4 Conclusions for three-arm trials

We do not find any test that is uniformly best. Each test has its own advantages and disadvantages, depending on the sizes of the Δ_is. As anticipated, when the Δ_i's are close to being equal, T_1, T_1^0 and S_1^0 perform well, but their loss of power when Δ_2 is smaller than Δ_1 is too great for them to be recommended, unless there is strong prior knowledge that the experimental treatments will be similar to each other. When one of the Δ_is is close to zero, T_∞ performs best, but T_2 is slightly better otherwise, as anticipated.

The conclusions from the comparisons of power and expected loss are broadly similar, though not identical. The power is the major component in the calculation of expected loss, $E(\text{loss}) = (\Delta_1 - \Delta_2) \times P(\text{type III/IV error}) + \Delta_1 \times (1 - \text{power})$, since type-III/IV errors have either small probability, relative to type-II errors, or a very small impact (when Δ_2 is close to Δ_1). The greatest difference is around $\Delta_2 = \Delta_1/2$, where the impact of a type-IV error is important and the probability of a type-IV error is non-negligible. In these cases the expected loss seems to favour T_∞ over T_2 slightly more than the power. The few cases we noted where different conclusions were reached about the ordering of tests are not enough to make a difference to our overall recommendations, but they do serve as a warning that the power should not be blindly accepted as the only measure of test performance.

Overall, T_2, T_∞ and Hochberg's procedure give a stable performance for a wide range of values of Δ_1 and Δ_2 and any of them is probably acceptable in practice. The performance of T_2 seems very slightly better than the others and we recommend it for practical applications, although up to now it does not seem to have been used. However, some might prefer T_∞ due to its simplicity and almost equally good performance. Although almost as good as these two tests, Hochberg's procedure is more complicated to use, especially for more than three arms, and does not seem to have any particular advantages over T_2 and

T_∞. If Hochberg's procedure is used without adjustment, it is conservative and is then not as good as T_2 and T_∞.

3.5 A technical problem regarding type-III errors

The results in Section 3.3 and further results (not shown) with $\Delta_2 = 0$ show that the probability of a type-III error is not large enough to be of any practical consequence. However, it is of interest to consider whether type-III error can ever be large enough to be important. In particular, when a test is being conducted at significance level α, it is desirable, perhaps essential, that the probability of a type-III error cannot be greater than α for any configuration of $\mathbf{\Delta}$. As noted in Chapter 2, for practical purposes, a type-III error can be considered as equivalent to a type-I error, so that our test procedure must control for type-III errors as well as type-I errors. In this section, we study this issue. First we show that there cannot be a problem with T_∞. Next we perform calculations which show that, for S_1^0 it is possible that $P(\text{type-III error}) > \alpha$, though only for very small α. Finally, we carry out simulations to show that the same can be true for T_1 and T_2.

First, though, we note that $P(\text{type-III error}) \leq \text{Power}/2$, since $P(\text{reject } H_0 \bigcap Z_2 > Z_1) \leq P(\text{reject } H_0)/2$ if $\Delta_2 < \Delta_1$. Therefore, no problem can arise for very small values of d_1, but only for those values which are large enough to give a power greater than 2α. However, large values of d_1 will make it very unlikely that a type-III error will occur. Informally, this might immediately suggest that the problem we are considering could not possibly arise. However, although this is true for practical values of α, we shall see that it is not true for very small values of α, since in this case a power of 2α occurs when d_1 is still very small.

3.5.1 Type-III errors for T_∞

Consider the case where $\Delta_2 = 0$, which maximises the probability of a type-III error. When using T_∞, a type-III error can only occur when Z_2 is greater than the cutpoint of the rejection region. Since this cutpoint C_α is the upper α point on the distribution of $\max(Z_1, Z_2)$, it is larger than $z_{1-\alpha}$. Hence, for T_∞,

$$P(\text{type-III error}) = P(T_\infty > C_\alpha \bigcap Z_2 > Z_1 | d_1 = d, d_2 = 0)$$

$$= P(\max(Z_1 + d, Z_2) > C_\alpha \bigcap Z_2 > Z_1 + d | d_1 = d_2 = 0),$$

where $d > 0$. Then, continuing to condition on $d_1 = d_2 = 0$, but dropping it from the notation (for simplicity),

$$\begin{aligned} P(\text{type-III error}) &= P(\max(Z_1 + d, Z_2) > C_\alpha | Z_2 > Z_1 + d) P(Z_2 > Z_1 + d) \\ &= P(Z_2 > C_\alpha | Z_2 > Z_1 + d) P(Z_2 > Z_1 + d) \\ &= P(Z_2 > C_\alpha \bigcap Z_2 > Z_1 + d) \\ &\leq P(Z_2 > C_\alpha \bigcap Z_2 > Z_1) \\ &< P(max(Z_1, Z_2) > C_\alpha \bigcap Z_2 > Z_1) \\ &= \frac{\alpha}{2} \end{aligned}$$

This extends to more than two experimental arms, since in the above result the upper α point on the distribution of $\max(Z_1, Z_2)$ is replaced by the upper α point on the distribution of $\max(Z_1, \ldots, Z_I)$ and $z_{1-\alpha}$ is replaced by the upper α point on the distribution of $\max(Z_J, \ldots, Z_I)$ for some $2 \leq J \leq I$. Clearly, the above inequalities would also hold in this case. Therefore, with T_∞, there is no need to be concerned about the probability of a type-III error becoming greater than the size of the test. This result and the proof are quite similar to the result of Bofinger (1985), who showed that for Tukey's multiple comparison test for all pairwise comparisons, with a two-sided alternative, the probability of a type-III error is always less than α.

3.5.2 Type-III errors for S_1^0

We consider the case of equal allocation, so that $Cov(Z_1, Z_2) = 1/2$. In this case the distribution of S_1^0 is known to be $N(d_1 + d_2, 3)$, it is straightforward to calculate $P(\text{type-III error})$ for any α and check whether or not it is greater than α. Let $D_1^0 = Z_1 - Z_2 \sim N(d_1 - d_2, 1)$ and therefore $Cov(S_1^0, D_1^0) = 0$. Consider $d_2 = 0$, so that $S_1^0 \sim N(d_1, 3)$ and $D_1^0 \sim N(d_1, 1)$. Then

$$\begin{aligned} P(\text{type-III error}) &= P\left(\text{Reject } H_0 \bigcap Z_2 > Z_1\right) \\ &= P\left(S_1^0 > C_\alpha \bigcap D_1^0 < 0\right), \end{aligned}$$

where C_α is the cutpoint of the rejection region for S_1^0. Since $P\left(S_1^0 > C_\alpha | d_1 = d_2 = 0\right) = \alpha$, $C_\alpha = \sqrt{3} z_{1-\alpha}$.

Now, since S_1^0 and D_1^0 are independent,

$$\begin{aligned} P\,(\text{type-III error}) &= P\left(S_1^0 > \sqrt{3}z_{1-\alpha}\right) P\left(D_1^0 < 0\right) \\ &= P\left(\frac{S_1^0 - d_1}{\sqrt{3}} > z_{1-\alpha} - \frac{d_1}{\sqrt{3}}\right) P\left(D_1^0 - d_1 < -d_1\right) \\ &= \left\{1 - \Phi\left(z_{1-\alpha} - \frac{d_1}{\sqrt{3}}\right)\right\} \Phi(-d_1) \\ &= \Phi\left(z_\alpha + \frac{d_1}{\sqrt{3}}\right) \Phi\left(-d_1\right). \end{aligned}$$

It is, therefore, simple to calculate the probability that S_1^0 gives a type-III error for any α and any d_1.

The results are shown in Figure 3.11 for $\alpha = 0.01\%$ and it is immediately clear that the probability of a type-III error can be greater than α for a range values of d_1 from less than 1 to greater than 2. In fact this happens for at least some values of d_1 for all $\alpha < 0.000753$ (to 6 decimal places). At this significance level, $P(\text{type-III error}) > \alpha$ for $d_1 = 1.15$ and values very close (equal to two decimal places) to this only.

3.5.3 Type-III errors for other test statistics

The results for S_1^0 are reassuring in the sense that they show that problems arise only for very small values of α which are unlikely to be used in practice to perform a significance test. However, they also show that there is a possibility of a problem arising. For T_1 and T_2, it is necessary to carry out simulations to check for possible problems. A set of two million errors were simulated and used to obtain rejection cutpoints for different levels of significance. Various values of d_1 were added to the errors with $d_2 = 0$ to estimate the probability of type-III error. For tests at the 5%, 2.5% and 1% levels, the probability of a type-III error never approaches α, as shown in Table 3.18. However, we also carried out simulations of tests at the 0.1%, 0.01% and 0.001% levels of significance and these results are shown in Table 3.19.

The results show that T_1 fails even at 0.1%, where S_1^0 is still satisfactory. On the other hand, T_2 fails at 0.001%, but is still clearly satisfactory at 0.01%. More extensive simulations are needed if a very small level of significance is to be used, although the overall pattern of results is clear. As proved above, T_∞ never has any problem and we can see that the probability of a type-III error remains less than $\alpha/2$ throughout. Thus, T_∞ appears to be the only test statistic for which the probability of a type-III error is never greater than the size of the test.

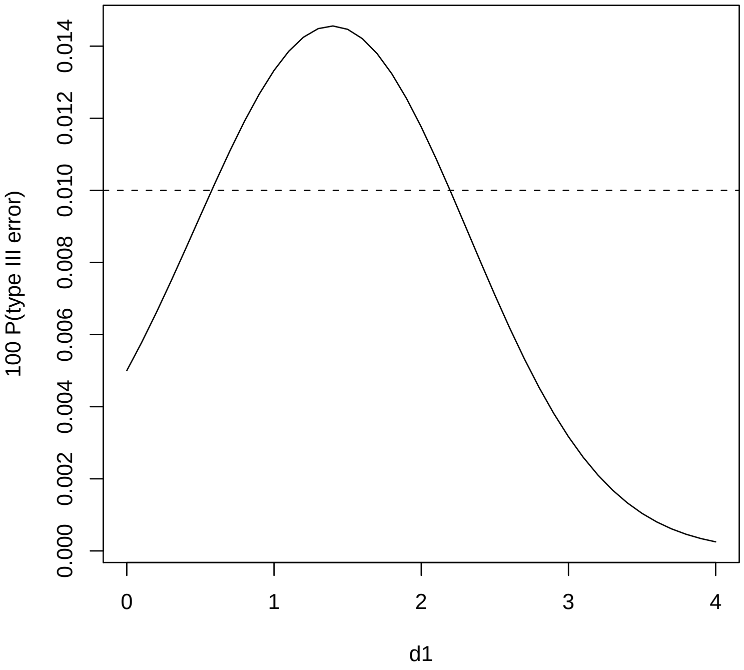

Figure 3.11: Probability of type-III error for S_1^0 at 0.01%.

Table 3.18: Percentage of type-III errors when $d_2 = 0$ at 5, 2.5 and 1%

	Significance level											
	5%				2.5%				1%			
d_1	T_1	T_2	T_∞	S_1^0	T_1	T_2	T_∞	S_1^0	T_1	T_2	T_∞	S_1^0
0.125	2.593	2.535	2.491	2.651	1.364	1.272	1.225	1.364	0.572	0.519	0.511	0.563
0.25	2.604	2.450	2.394	2.691	1.376	1.252	1.196	1.409	0.618	0.534	0.504	0.601
0.5	2.615	2.272	2.173	2.745	1.449	1.195	1.097	1.474	0.683	0.522	0.471	0.662
1	2.169	1.739	1.581	2.262	1.328	0.979	0.835	1.323	0.655	0.428	0.375	0.630
1.5	1.401	1.080	0.967	1.473	0.917	0.657	0.542	0.915	0.533	0.319	0.244	0.488

Table 3.19: Percentage of type III errors when $d_2 = 0$ with small sizes

		Test statistic			
100α	d_1	T_1	T_2	T_3	S_1^0
0.1	1	0.1098	0.0568	0.0412	0.0953
	1.1	0.1110	0.0570	0.0399	0.0952
	1.2	0.1108	0.0574	0.0384	0.0962
	1.3	0.1093	0.0565	0.0362	0.0935
0.01	1	0.0153	0.0075	0.0043	0.0133
	1.1	0.0166	0.0082	0.0042	0.0154
	1.2	0.0173	0.0090	0.0041	0.0160
	1.3	0.0188	0.0091	0.0039	0.0163
0.001	1	0.0020	0.0010	0.0004	0.0021
	1.1	0.0026	0.0011	0.0004	0.0023
	1.2	0.0027	0.0011	0.0004	0.0023
	1.3	0.0026	0.0012	0.0004	0.0024

3.5.4 Proposed solution

The results in this section show why it was necessary to redefine the size of the test, as in Section 2.2, to be $\alpha^\dagger = \max\left[\alpha, \max_{\boldsymbol{\Delta}}\{P(\text{type-III error})\}\right]$. Given that the best test statistics are T_∞, where $\alpha^\dagger = \alpha$ for all α, and T_2, where $\alpha^\dagger = \alpha$ for all except very small values of α, it is unlikely that this will ever make any difference in practice. In particular, all results reported in this thesis remain unchanged with this new definition of size. However, it is important to note that, should anyone ever desire to use T_2 to carry out a significance test at a level of significance less than 0.01%, further work will be needed to determine the true cutpoints of the null distribution. The problem is more serious for T_1 and could have a bearing on practical tests, but we would not recommend this test statistic anyway.

3.6 Trials with more than three arms

The methodology described above can be applied to trials with any number of experimental arms and we might expect similar results, although there are more different configurations with more arms. Rather than try to be comprehensive, we look at just four-arm and eight-arm trials with a few different configurations. For $I = 3$, we consider configurations of the types $(d_1, d_1, 0)$, $(d_1, 0, 0)$ and $(d_1, 0.5d_1, 0.25d_1)$. For $I = 7$ we consider configurations with one non-zero, two non-zero, etc. and some other mixed sizes of effects. We also restrict attention to tests in the T_k family.

The simulations were run exactly as before except that we give cutpoints based on only 1 million simulation runs. The standard errors of these cutpoints are of the order of 0.003, based on the previous results, but should be slightly smaller than this due to greater stability with more arms. Again, we can have confidence that they are precise to two decimal places.

3.6.1 Four-arm results

Cutpoints for $I = 3$ are shown in Table 3.20, those for 2.5% having been done later in a separate simulation. These are obviously at larger values than for $I = 2$, since there is more chance of at least one arm being greater than any particular amount.

We present the power and expected loss only at the 5% level. The power for various configurations is shown in Table 3.21. Some of these configurations, with more different

Table 3.20: Cutpoints of the rejection region for four arms.

Test	5%	2.5%*	1%
T_1	2.5102	2.9291	3.4426
T_2	2.1282	2.4321	2.7814
T_∞	2.0581	2.3496	2.6837

* 2.5% critical values from a separate simulation.

values of d_1, are illustrated in Figure 3.12. The overall pattern is what we expect. When all treatments have similar effects, T_1 is best, but this test is clearly inferior in other circumstances. Overall, there are only small differences in power between T_2 and T_∞, with T_∞ being slightly better when one treatment is considerably better than the others, but T_2 being slightly better when several treatments have large effects. We can see that with more arms there are perhaps more configurations which make T_2 better than T_∞. For example, configurations like $(2, 2, 0)$ cannot occur with $I = 2$.

The corresponding results for expected loss are shown in Table 3.22 and Figure 3.13. These are broadly in line with those for power, although perhaps very slightly more in favour of T_∞, presumably because of the impact of type III/IV errors.

3.6.2 Eight-arm results

The cutpoints for $I = 7$ are shown in Table 3.23.

Again we present the power and expected loss only for 5% for a few different configurations. The powers for several cases are shown in Table 3.24 and a few of these are illustrated over a wider range of d_1 values in Figure 3.14. The same general comments apply as before, but we note that with just three arms better than the control, T_2 is already better than T_∞. For the irregular configurations in the bottom section of the table, we see that T_2 can be considerably better than T_∞. The differences between tests can be quite large in this case, with T_∞ having almost 5% greater power than T_2 and vice versa.

The corresponding expected losses are shown in Table 3.25 and Figure 3.15. The overall pattern of results is very much similar to what we described regarding the power. The last section of the table shows that there are many different configurations for which T_2 gives a smaller expected loss than T_∞. In particular, if suboptimal treatments give results which are fairly close, but not very close, to the optimum, then T_2, by exploiting contributions

Table 3.21: Power for four-arm trials at 5%.

d_1	d_2	d_3	T_1	T_2	T_∞
2.5	0	0	51.769	64.948	67.109
3	0	0	69.592	80.869	82.679
2.5	1.25	0	61.655	68.502	69.176
3	1.5	0	77.413	83.285	83.916
2.5	1.25	0.625	63.540	69.104	69.448
3	1.5	0.75	78.481	83.585	84.033
2.5	1.25	1.25	68.280	71.292	70.903
3	1.5	1.5	82.131	85.034	84.936
2	2	0	63.624	65.157	64.316
2.5	2.5	0	81.929	82.764	81.903
3	3	0	93.092	93.429	92.972
2	2	1	67.353	66.869	65.399
2.5	2.5	1.25	84.052	83.704	82.440
3	3	1.5	93.964	93.792	93.183
2	2	2	77.483	75.078	72.852
2.5	2.5	2.5	91.354	89.693	88.040
3	3	3	97.579	96.884	96.099

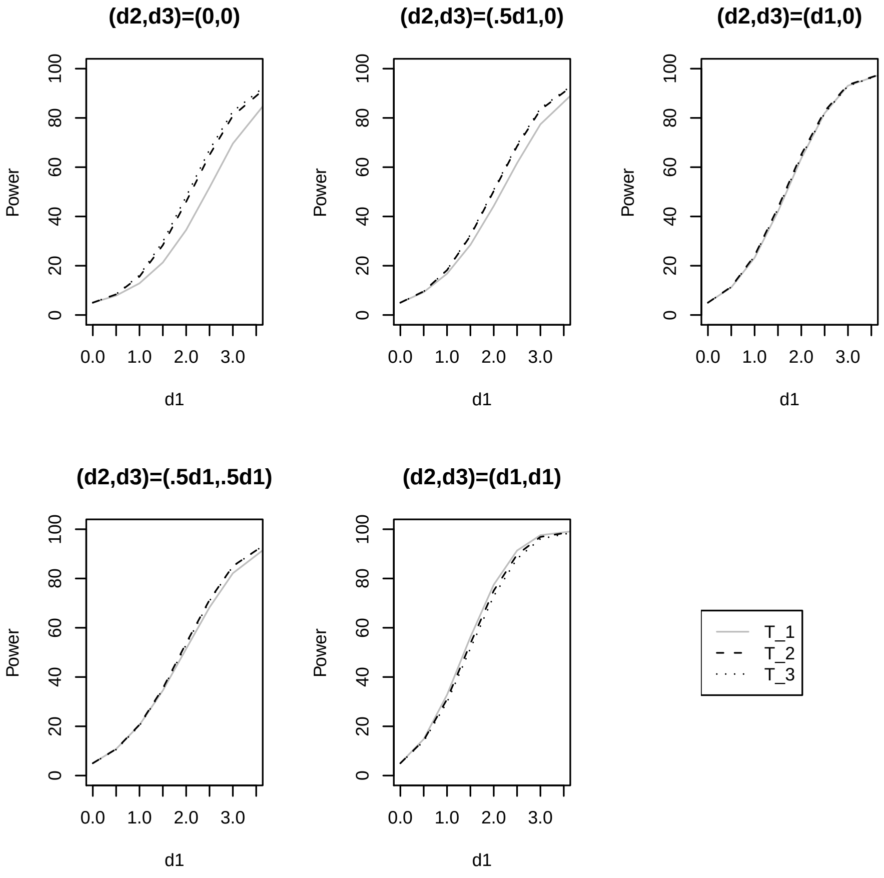

Figure 3.12: Power at 5% for four arms.

Table 3.22: Expected loss at 5% for four-arm trials.

d_1	d_2	d_3	Bound	T_1	T_2	T_∞
2.5	0	0	2.4583	1.2149	0.8843	0.8294
3	0	0	2.9500	0.9157	0.5771	0.5223
2.5	1.25	0	2.4375	1.0398	0.8629	0.8423
3	1.50	0	2.9250	0.7536	0.5735	0.5511
2.5	1.25	0.625	2.4271	1.0120	0.8633	0.8492
3	1.50	0.75	2.9125	0.7331	0.5745	0.5565
2.5	1.25	1.25	2.4167	0.9402	0.8517	0.8537
3	1.50	1.50	2.9000	0.6756	0.5796	0.5761
2	2	0	1.9333	0.7314	0.7001	0.7164
2.5	2.5	0	2.4167	0.4528	0.4318	0.4532
3	3	0	2.9000	0.2075	0.1973	0.2111
2	2	1	1.9167	0.6927	0.6976	0.7243
2.5	2.5	1.25	2.3958	0.4344	0.4397	0.4694
3	3	1.50	2.8750	0.2071	0.2109	0.2279
2	2	2	1.900	0.4503	0.4984	0.5430
2.5	2.5	2.5	2.3750	0.2162	0.2577	0.2990
3	3	3	2.8500	0.0726	0.0935	0.1170

Table 3.23: The critical values at 5%, 2.5% and 1% for 8-arm.

Test	5%	2.5%	1%
T_1	3.3764	3.8756	4.4848
T_2	2.4937	2.7895	3.1366
T_∞	2.3436	2.6163	2.9357

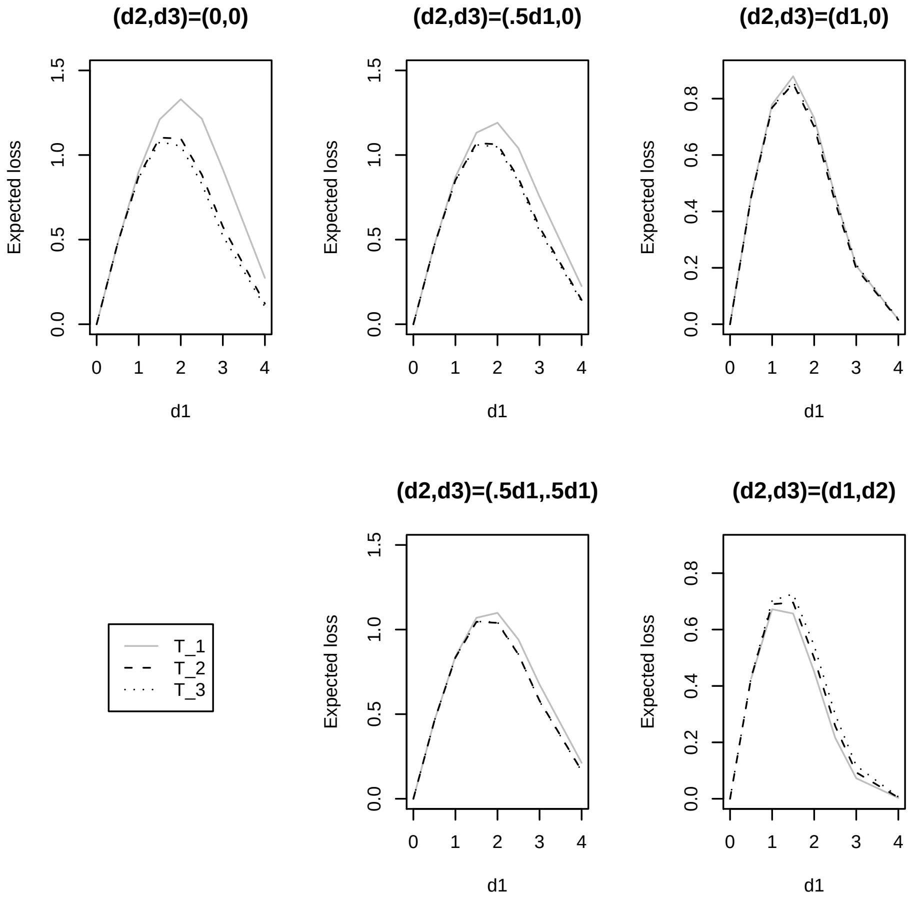

Figure 3.13: Expected loss at 5% for four arms.

Table 3.24: Power at 5% for eight-arm trials.

d_1	d_2	d_3	d_4	d_5	d_6	d_7	T_1	T_2	T_∞
3	0	0	0	0	0	0	38.656	69.786	74.523
3.5	0	0	0	0	0	0	56.545	84.405	87.616
2.5	2.5	0	0	0	0	0	60.031	71.855	72.796
3	3	0	0	0	0	0	79.204	87.218	87.533
3.5	3.5	0	0	0	0	0	91.579	95.504	95.548
2	2	2	0	0	0	0	55.904	61.866	61.399
2.5	2.5	2.5	0	0	0	0	76.939	81.303	80.533
3	3	3	0	0	0	0	90.859	92.957	92.374
2	2	2	2	0	0	0	66.653	68.853	67.254
2.5	2.5	2.5	2.5	0	0	0	85.198	86.402	84.888
3	3	3	3	0	0	0	95.240	95.588	94.735
2	2	2	2	2	0	0	73.772	73.756	71.387
2.5	2.5	2.5	2.5	2.5	0	0	89.715	89.525	87.753
2	2	2	2	2	2	0	78.770	77.351	74.539
2.5	2.5	2.5	2.5	2.5	2.5	0	92.474	91.657	89.753
2	2	2	2	2	2	2	82.402	80.100	76.981
2.5	2.5	2.5	2.5	2.5	2.5	2.5	94.273	93.069	91.212
2.5	2	1.5	1	0.5	0	-0.5	59.577	67.277	67.315
2.5	2.25	2	0.5	0.5	-0.5	-0.5	67.707	73.561	73.070
2.5	2	1	0	0	-2	-2	51.961	64.177	65.416
3.2	1.6	0.8	0.4	0.2	0.1	0.05	59.051	79.124	81.571
2	1.5	1.5	1.5	1.5	1.5	1.5	65.071	63.352	60.298
2.25	1.75	1.75	1.75	1.75	1.75	1.75	76.198	74.511	71.413
2.2	1.95	1.85	1.5	1.5	0	0	66.203	67.741	65.882
2.7	2.6	2.5	2	0	0	0	84.340	86.191	84.964
2.8	2.8	2.7	2.6	1.5	0	0	91.208	91.742	90.480

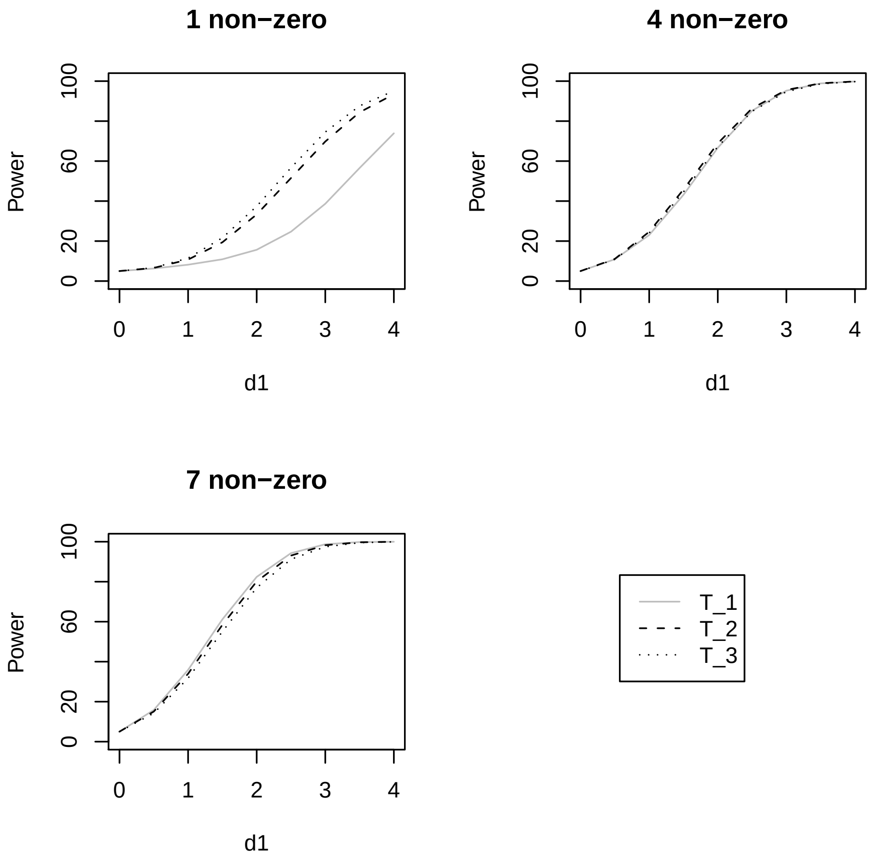

Figure 3.14: Power at 5% for eight arms.

from them to reject H_0 seems somewhat better than T_∞. In practice, we would not be surprised to see configurations of this sort, rather than the extreme case where only one or two treatments are better than the control, so this seems like an important practical result in favour of T_2.

3.6.3 Conclusions

The conclusions regarding three-arm trials seem to hold for trials with more arms. In particular, in the absence of strong prior knowledge that all experimental treatments will have similar effects, we should avoid using T_1. Overall, we would recommend T_2, but T_∞ is very nearly as good and simpler to calculate. However, the differences between test procedures is greater with more arms, so if there is prior knowledge that only a small number of treatments is likely to be better than the control, then we could use T_∞. Generally, either T_2 or T_∞ is acceptable in all cases.

Table 3.25: Expected loss at 5% for eight-arm trials.

d_1	d_2	d_3	d_4	d_5	d_6	d_7	Bound	T_1	T_2	T_∞
3	0	0	0	0	0	0	2.9786	1.8458	0.9115	0.7691
3.5	0	0	0	0	0	0	3.4750	1.5224	0.5473	0.4347
2.5	2.5	0	0	0	0	0	2.4643	1.0029	0.7067	0.6827
3	3	0	0	0	0	0	2.9571	0.6245	0.3839	0.3744
3.5	3.5	0	0	0	0	0	3.4500	0.2948	0.1574	0.1559
2	2	2	0	0	0	0	1.9571	0.8864	0.7664	0.7750
2.5	2.5	2.5	0	0	0	0	2.4464	0.5774	0.4681	0.4873
3	3	3	0	0	0	0	2.9357	0.2743	0.2114	0.2288
2	2	2	2	0	0	0	1.9429	0.6688	0.6245	0.6561
2.5	2.5	2.5	2.5	0	0	0	2.4286	0.3702	0.3401	0.3779
3	3	3	3	0	0	0	2.9143	0.1428	0.1324	0.1580
2	2	2	2	2	0	0	1.9286	0.5254	0.5256	0.5728
2.5	2.5	2.5	2.5	2.5	0	0	2.4107	0.2572	0.2619	0.3062
2	2	2	2	2	2	0	1.9143	0.4250	0.4533	0.5095
2.50	2.50	2.50	2.50	2.50	2.50	0	2.3929	0.1882	0.2086	0.2562
2	2	2	2	2	2	2	1.9000	0.3520	0.3980	0.4604
2.5	2.5	2.5	2.5	2.5	2.5	2.5	2.3750	0.1432	0.1733	0.2197
2.5	2	1.5	1	0.5	0	-0.5	2.4500	1.1734	0.9854	0.9786
2.5	2.25	2	0.5	0.5	-0.5	-0.5	2.4518	0.9339	0.7940	0.8035
2.5	2	1	0	0	-2	-2	2.4893	1.3050	1.011	0.9779
3.2	1.6	0.8	0.4	0.2	0.1	0.05	3.1546	1.3729	0.7332	0.6523
2	1.5	1.5	1.5	1.5	1.5	1.5	1.9214	0.9186	0.9423	0.9909
2.25	1.75	1.75	1.75	1.75	1.75	1.75	2.1589	0.7933	0.8216	0.8790
2.2	1.95	1.85	1.5	1.5	0	0	2.1357	0.9134	0.8776	0.9113
2.7	2.6	2.5	2	0	0	0	2.6300	0.5412	0.4917	0.5220
2.8	2.8	2.7	2.6	1.5	0	0	2.7114	0.3151	0.2998	0.3336

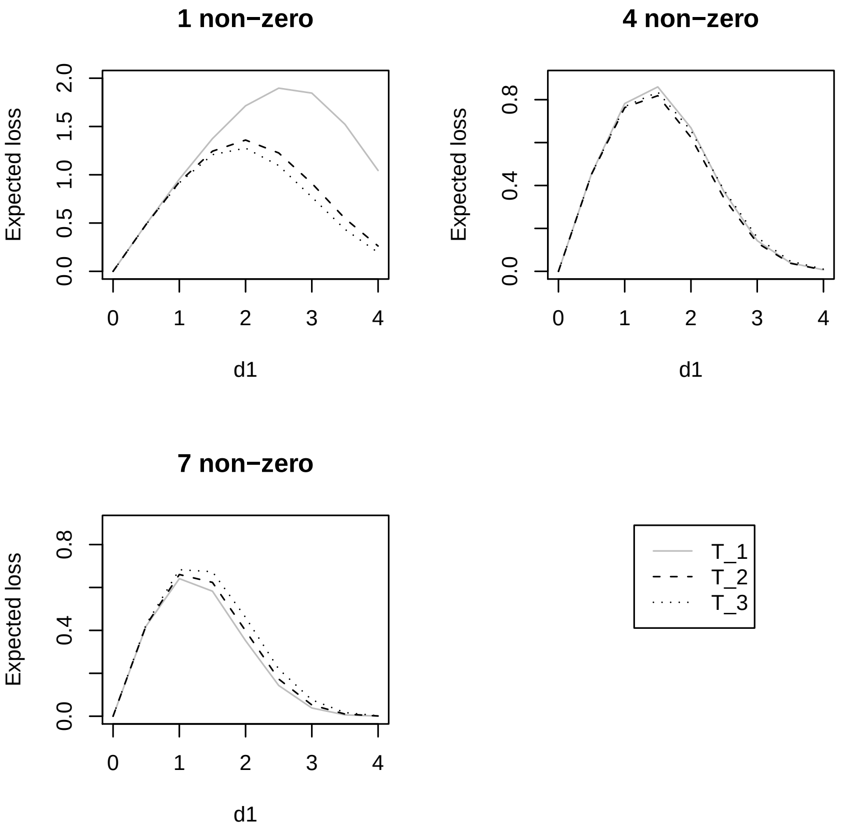

Figure 3.15: Expected loss at 5% for eight arms.

Chapter 4

Design Issues in Multi-Arm Trials

In clinical trials, as in any other experiment, it is important to decide how many subjects are required for a particular experiment and how they should be allocated to treatments. The bigger the trial, the more expensive it is to run and, in particular in clinical trials, recruiting subjects can be a difficult task. However, bigger trials are also more informative about the relative performances of the treatments. Therefore it is very important for an experimenter or drug company to try to produce a convincing result with limited resources by selecting an appropriate sample size. If the practicalities allow it, experimenters should also allocate proportions of subjects to the different treatments in such a way as to maximise the information on their comparison. In this chapter we address these problems.

The comparison of test statistics in Chapter 3 was based on each arm, including the control, having equal numbers of subjects allocated to it. However, given the emphasis on comparisons with the control, we can expect to get better results by allocating more subjects to the control than to each experimental arm. In this chapter we focus on how the proportion of subjects on the control arm influences the power and the expected loss for $I = 2$. For three arm trials, we consider a range of values between $1/3$ and $1/2$ for the proportion on control. Throughout, we will assume equal numbers of subjects on each experimental arm, since it would be practically and ethically more difficult to do anything else and this is optimal (Bechhofer and Nocturne, 1972) in the absence of prior knowledge about the relative performances of the treatments.

After reviewing the existing literature in Section 4.1, a general strategy is described in Section 4.2. A simulation study is described in Section 4.3 which compares different sample allocations in terms of power and expected loss at sizes 5% and 1% and estimates the optimal allocations to maximise power at size 5%. We also address the question of choosing the sample size to achieve a required power or expected loss in Section 4.4. Finally, some general comments are made in Section 4.5.

4.1 Background Literature

Several authors, including Dunnett (1955, 1964), Bechhofer and Tamhane (1983), Fleiss (1986), Hsu (1989), Tang and Lin (1997), Liu (1997), Horn and Vollandt (1998) and Dunnett et al. (2001), look at sample size calculations or allocation in multiple comparisons of I treatments with a control. Most of this literature is based on Dunnett type tests, i.e. testing the individual hypotheses H_{0i}. Tang and Lin (1997) obtain sample sizes for an approximate LR test for $H_0^* : \mathbf{\Delta} = \mathbf{0}$, as well as some other tests, but, as noted in Chapter 2, such tests are not sensibly applicable to our null hypothesis. Horn and Vollandt (1998) consider multiple t tests without adjustment for multiple comparisons, as well as Dunnett's test, but this is not of interest in this thesis. Hence, the results available in the literature are relevant only to our test statistic T_∞.

Many of these authors give tables of sample sizes needed to achieve a specified power for one- or two-sided alternatives or both. Others study the proportion of subjects to allocate to the control to achieve maximum power or minimum total sample size for fixed power. Several authors, including Dunnett (1955), Bechhofer and Tamhane (1983), Hsu (1989) and Liu (1997), consider optimal allocations to ensure the maximum coverage of one-sided (or two-sided) confidence intervals of fixed finite endpoint (or fixed width), or highest/lowest endpoint (or minimum width) for a fixed coverage. As noted in Chapter 2, rejecting $H_0 : \Delta_i \leq 0 \; \forall i$ if at least one one-sided confidence interval includes zero is equivalent to testing using T_∞ and maximising the endpoint of Dunnett's one-sided confidence intervals is equivalent to maximising the power of T_∞.

Dunnett (1955) gave numerical results to show that the optimal allocation gives slightly fewer subjects to the control than the allocation which minimises the average variance of the estimators $\hat{\Delta}_i$, the so-called *square root* allocation. This has $n_0/n_i = \sqrt{I}$ and, in terms of δ is given by $\delta = (1 - I^{-1/2})/(I - 1)$ for $I = 2, 3, \ldots$ (Finney, 1952). Dunnett also showed that the square root allocation gives very close to maximum power and can be used for all practical purposes if maximum power is required.

Better numerical approximations are used to improve the accuracy of Dunnett's results by Bechhofer (1969), later superseded by Bechhofer and Nocturne (1972), who also consider the two-sided case. Extensive tables of optimal allocations are given by Bechhofer and Tamhane (1983) and this can be considered the most complete set of recommendations to date. All of these results were obtained by numerical integration.

In many papers on sample size calculations, the power considered is the probability of rejecting $H_{0i} : \Delta_i \leq 0$ for all i such that $\Delta_i > \Delta^*$, where Δ^* is some specified practically important difference. Therefore their results are only relevant to T_∞ in the case that exactly one Δ_i is positive, the rest being zero.

Horn and Vollandt (1998), however, consider several different power characteristics, including what they call "any-pair" power which is the probability of rejecting at least one false H_{0i} and so is directly relevant to T_∞. Note, however, that their any-pair is not identical to the power of T_∞, since it is possible that a true H_{0i} is rejected, while all false H_{0i} are not rejected, in which case H_0 is false and is correctly rejected, even though no true H_{0i} is rejected. It is also different from the adjusted power calculated in Chapter 3, since in that calculation, if a false H_{0i} is rejected and corresponds to the largest Z_i, this is counted as a type-III error irrespective of whether one or more true H_{0i}s are also rejected. This can be considered as an extension of the consideration of directional errors in two-sided tests, as considered by Hsu (1989). Therefore, we do not expect to replicate their results. However, for large Δ_i, we would expect the difference between their any-pair power and the power of T_∞ to be negligible. Horn and Vollandt note that the different types of power require different sample sizes. Following the arguments from Chapter 3, in addition to the power of the hypothesis test, we prefer to consider the expected loss, rather than the adjusted power or Horn and Vollandt's any-pair power when considering the selection of the best treatment.

Horn and Vollandt, and others such as Liu (1997), argue that, if any difference of at least Δ^* is of interest, then we should choose the sample size to ensure a power of at least $1 - \beta$ for *any* configuration of Δ_is. Therefore, they recommend using the least favourable configuration (LFC), which, in the absence of prior knowledge, is $(\Delta^*, 0, \ldots, 0)$. A slightly different version of the LFC is used by Thall et al. (1988), who defined a marginal improvement Δ_1^* and called $(\Delta^*, \Delta_1^*, \ldots, \Delta_1^*)$ the LFC. In their examples they use $\Delta_1^* = \Delta^*/4$. They use the adjusted power and argue that it is not reasonable to try to detect very small differences between treatments and so do not consider values of Δ_2 close to Δ_1. Horn and Vollandt also consider the case of prior knowledge that at least g of the treatments will differ from the control by at least Δ^* and consider configurations of the form $(\Delta^*, \ldots, \Delta^*, 0 \ldots, 0)$, although they suggest that this is of little practical importance in clinical trials.

Horn and Vollandt give tables of results for equal allocation and for the square root allocation. They find that, although the latter allocation is better, the differences are not very great. Marschner (2007) follows the methodology of Horn and Vollandt (1998) very closely, but concentrates on the most favourable configuration (MFC), where $\Delta_i = \Delta^* \; \forall i \in \{1, \ldots, I\}$, noting that the choice of an appropriate configuration to choose the sample size is not an obvious one.

4.2 A Strategy for Choosing the Sample Size

The practical question of which configuration of Δ_is to choose the sample size for has received very little attention in the literature. Using the LFC is rather extreme and can lead to over-powered studies if the true configuration is different from the LFC.

Note that this question does not arise in two-arm trials. In these, choosing a sample size to achieve a power of $1 - \beta$ to detect a difference of Δ^* leads to a power which changes smoothly from $1 - \beta$ as the true difference changes from Δ^*. In multi-arm trials, however, choosing a sample size to achieve a power of $1 - \beta$ to detect at least one difference when the differences are $\boldsymbol{\Delta}^* = [\Delta_1^*, \ldots, \Delta_I^*]$ leads to a power function which changes in different ways as $\boldsymbol{\Delta}$ changes from $\boldsymbol{\Delta}^*$ in different directions. For example, in a two-arm trial, ensuring 80% power for a difference of Δ^* will ensure less than 80% power for any difference of less than Δ^*. In a three-arm trial, ensuring 80% power for $\Delta_1 = \Delta^*, \Delta_2 = 0$ will achieve considerably more than 80% power for Δ_1 and Δ_2 both slightly less than Δ^*. The LFC approach might, therefore, lead to excessively large sample sizes being chosen.

If we were to run I separate trials to compare each experimental treatment to the control, in practice we would choose a sample size for each to achieve a specified power to detect a difference of Δ^*. Considering the multi-arm trial as a way to save resources over individual trials would then suggest using the MFC to choose the sample size and allocation. This is rather extreme in the opposite direction from the LFC and could easily lead to studies which are too small to detect treatment differences in all but the most favourable configuration.

Instead of either of these two approaches, it might be sensible to ensure a power of $1 - \beta$ for $\Delta_1 = \Delta^*$ and a best guess as to how much worse the inferior experimental treatments will be. If there is reason to believe that the treatments will have similar effects, $\Delta_i = (I +$

$1-i)\Delta_1/I$ might be a reasonable guess, while otherwise $\Delta_i = \Delta_1/i$ might be a reasonable and moderately conservative choice. For $I = 2$, these both imply that $\Delta_2 = \Delta_1/2$, which we will study later, along with the extreme configurations.

In a review paper promoting the use of Bayesian thinking in frequentist statistics, Bayarri and Berger (2004) suggest a procedure which, in our problem, would involve specifying a prior distribution for $\boldsymbol{\Delta}$ and choosing a sample size to ensure at least 80% power averaged over that prior. In this chapter, we study the power at different values and configurations of $\boldsymbol{\Delta}$ and recommend a procedure to choose the sample size which depends on a prior point estimate of each Δ_i. It would be a simple modification of this to implement Bayarri and Berger's procedure but, in keeping with the likelihood-based approach taken here, we do not go as far as using Bayesian methods. This does, however, argue against using the LFC or MFC as the basis for choosing the design, as these would correspond to minimax and minimin procedures respectively, i.e. they represent the most extreme possible prior beliefs.

Note also that, if the argument for using the LFC is accepted, the results in Chapter 3 suggest that we should use the test statistic T_∞, whereas the argument for using the MFC would lead to us using S_1^0. There, however, we argued that T_2 was better over a broader range of configurations and should at least be considered as a strong competitor to T_∞. The attitude we take to choosing the allocation and sample size in this chapter is logically coherent with that conclusion.

In Section 4.3, we study the effect of the allocation for the test statistics T_1, T_2 and T_∞ for $I = 2$ at the 5% and 1% levels of significance. Although Hochberg's procedure also seems good, giving results in between those of T_2 and T_∞, it is more complicated to use in practice. It has no univariate test statistic and, for each allocation, requires us to find a new nominal significance level α^* to achieve the correct size. We will not study it further in this thesis. The other test statistics studied in Chapter 3 seemed clearly inferior or at least had some disadvantages and no clear advantages, so we do not study them further. For each of the statistics considered, we look for the choice of δ (the proportion of subjects on each experimental arm) which maximises the power. Some of the authors mentioned in Section 4.1 instead look for the allocation which minimises the sample size required to achieve a particular power, typically 80%. These are essentially different ways of looking at the same issue, but by exploiting a suitable version of the scaling used in Chapters 2 and 3, we can present the results in a form which allows us to illustrate the results across a reasonable range of powers, such as 65-95%, simultaneously.

Given prior information, we will see in Section 4.4.1 how these can be used to choose the sample size to achieve a specified power, conditional on the allocation chosen. By separating the choice of allocation from the sample size, we allow suboptimal allocations to be used if this is appropriate for ethical or practical reasons, i.e. a simple method for calculating the sample size can be used whatever allocation proportions have been chosen. In Section 4.3 we also consider the impact of patient allocation on the expected loss.

4.3 Effects of Allocation on Power and Expected Loss

Changing the proportion, δ, of subjects allocated to each experimental arm also controls the correlation, ρ. For $I = 2$, $\rho = \delta/(1-\delta)$. By increasing δ we increase ρ and by increasing the number of subjects on the control arm we decrease ρ. The range of δ and ρ are therefore $0 < \delta < 1/2$ and $0 < \rho < 1$, respectively. $\delta = 0$ would mean no subjects on the experimental arms, while $\delta = 1/2$ would mean no subjects on the control. Getting uncorrelated estimators of the Δ_i by having a different control group for each experimental arm would not help, because we lose power by increasing the variances of the treatment comparisons, i.e. each comparison is between two groups of $N/4$ subjects, rather than between two groups of, say, $N/3$ subjects. This was confirmed by Proschan and Follmann (1995), who showed that, even if we fail to adjust for multiple testing when performing separate experiments, higher power is still obtained in a single experiment with a common control.

The optimal allocation depends on the relative sizes of the unknown Δ_i, but exact results can be obtained in only a few special cases. At one extreme, if one assumes that $\Delta_i = \Delta\ \forall i$ and $S_1^0 = \sum_{i=1}^{I} Z_i$ is used, i.e. the test is based on a single contrast comparing the control with the average of the active treatments which are all considered identical, then the trial reduces to the two-arm case and $\delta = 1/(2I)$ is optimal, so that half the subjects would receive the control treatment. We would expect a similar proportion to be optimal for $T_1 = X_1^+ + \sum_{i=2}^{I} \frac{(X_i - E_i)^+}{\sqrt{V_i}}$ (which excludes arms which contribute less than expected taking into account the correlation ρ under H_0), which we saw in Chapter 3 was somewhat better than S_1^0 over a range of configurations.

At the other extreme, we noted in Chapter 3 that T_∞ is best for a set of values of the form $(\Delta, 0, \ldots, 0)$. Several authors have studied this situation, as described in Section 4.1, mainly for Dunnett's test, from which we can obtain the results for T_∞ as a special case. We note there that the square root allocation $\delta = 1/(I + \sqrt{I})$ has been found to be very close to optimal for maximising the power.

These extremes suggest that, for $I = 2$, a sensible range of values of δ is $\left[\frac{1}{4}, \frac{1}{3}\right]$, i.e. between 1/3 and 1/2 of subjects should be allocated to the control. General results for T_1 and T_2 cannot be obtained, so we proceed by simulation. For completeness, and to check for consistency with published results, we also simulate results for T_∞. Details of the simulation are described in Subsection 4.3.1 and the results are described in Subsections 4.3.2 and 4.3.4.

4.3.1 Simulations

The simulations described in the previous chapter had to be modified to allow for different patient allocations. In Chapter 3, the simulated values of Z_i, for $i = 1, 2$, were calculated by $w_i + (\Delta_i/\sigma)$, where w_i is a random value from the null distribution of Z_i and $E(Z_i) = \Delta_i/\sigma$. However, if we change the patient allocation δ, the value of $\sigma = \sqrt{\{1-(I-1)\delta\}/\{\delta(1-I\delta)\}}$ and hence $E(Z_i)$ changes. In other words, the standardisation used is different for different patient allocations. To compare different allocations, we need to use a single standardisation.

We choose a baseline value of σ, denoted σ^*, and calculate Z_i from $w_i + (\Delta_i/\sigma^*)$ so that data from different patient allocations are being simulated from the same population. The particular baseline we choose is the minimum value of σ for $I = 2$. By differentiating $\sigma^2 = (1-\delta)/\{\delta(1-2\delta)\}$ with respect to δ and equating to zero, we find $\delta^* = \arg\min \sigma^2 = 1 - \left(1/\sqrt{2}\right)$ and hence $\sigma^* = 1/\sqrt{3-2\sqrt{2}}$. We define $d_i^* = \Delta_i/\sigma^* = \sqrt{3-2\sqrt{2}}\Delta_i$. Apart from this, the simulations proceed as in Chapter 3. Note that for the allocation $\delta = \delta^*$, d_i and d_i^* are the same.

Various values of the proportion on the control, $1 - 2\delta$, between 1/3 and 1/2 were used, along with 0.2 and 0.6, to show how much lower the power is outside the sensible range of allocations. To save computing time, we did not simulate one billion sets of data to get mean cutpoints. Instead, a fresh set of 100,000 simulations was used for each value of δ and used to obtain the critical values of the test statistics and the same values used to estimate the powers. As in Chapter 3, each set of simulations was generated simultaneously to obtain results for all test statistics and for all values of d_1^* and d_2^*, i.e. they are all subject to the same random errors. As in Chapter 3, the approximate standard error for any estimated power is about 0.158, 0.126 and 0.069 for 50%, 80% and 95% power respectively. For comparing different allocations, therefore, the estimated standard error of a difference will typically be $\sqrt{2} \times 0.126 = 0.178$ and so any difference in power greater than about 0.35% can safely be interpreted as being real and not due to simulation variance. For comparing tests at the same allocation, the standard error of a difference will be smaller than this, due to the use of the same set of simulated errors.

Results were calculated for $d_1^* \in \{2, 2.5, 3, 3.5, 4\}$ and $d_2^* \in \{0, d_1^*/2, 3d_1^*/4, d_1^*\}$. However, in Subsection 4.3.2 we present the results only for cases which gave power between 65 and 95% for size 5% and 1% tests. In Subsection 4.3.4 we present the expected losses for the same cases.

4.3.2 Effect of Allocation Proportion on Power

Table 4.1 shows the simulated powers against the proportion on the control, $1 - 2\delta$, for tests at the 5% level of significance for various values of d_1^* and d_2^*. Figure 4.1 shows the same information graphically for the configuration $\Delta_2 = 0$. In this and other graphs, T_3 refers to T_∞. Note that T_1 does not appear in the top left panel, since its power is always lower than 65%. In this situation, T_∞ is expected to be the best statistic and it is known that the optimal allocation for T_∞ is when $1 - 2\delta$ is slightly less than $\sqrt{2} - 1 \approx 0.414$. This is confirmed by the results, which show the highest powers for T_∞ at $1 - 2\delta = 0.4$. Interestingly, the highest power for each of the other tests is also attained very close to 0.4. However, the tables and figure show that the power is stable over a wide range of allocations, so that choosing the optimal allocation is not crucial. In Chapter 3 we saw that T_2 is only slightly worse than T_∞ for equal allocation with this configuration. The results here show that this is true across all allocations.

The corresponding results for testing at the 1% level of significance are given in Table 4.2 and illustrated in Figure 4.2. The slight jumpiness in this plot is due to simulation variation and suggests that we have done only just enough simulations to see the correct pattern for testing at 1%. These results show a similar overall pattern to the results for 5%, but the differences between the tests are somewhat greater. However, as d_1^* increases, the difference between tests decreases again. The optimal allocation seems to be close to 0.4 for all the test statistics.

Table 4.1 and Figure 4.3 show the powers for testing at 5% for $\Delta_2 = \Delta_1/2$, with d_1^* increasing from 2.5 to 3.5. The corresponding results for the 1% significance level are shown in Table 4.2 and Figure 4.4. As well as confirming the result from the Chapter 3 that T_2 is slightly better than T_∞ in this case, these simulations show some interesting

Table 4.1: Power for different proportions on the control at the 5% level.

			Control					
d_1^*	d_2^*	Test	.5	.45	$\sqrt{2}-1$	.4	.35	1/3
2.5	0	T_1	61.634	63.160	63.475	63.514	63.352	63.117
2.5	0	T_2	68.907	70.147	70.317	70.496	69.940	69.570
2.5	0	T_∞	70.329	71.538	71.626	71.950	71.179	70.685
3	0	T_1	77.514	79.083	79.365	79.610	79.462	79.193
3	0	T_2	83.485	84.594	84.757	85.064	84.595	84.376
3	0	T_∞	84.674	85.662	85.799	86.107	85.487	85.101
3.5	0	T_1	89.099	90.191	90.559	90.637	90.548	90.299
3.5	0	T_2	92.851	93.525	93.709	93.751	93.528	93.326
3.5	0	T_∞	93.512	94.111	94.264	94.342	94.036	93.722
2.5	1.25	T_1	74.704	74.647	74.243	73.894	72.371	71.745
2.5	1.25	T_2	75.052	75.438	75.115	75.061	73.835	73.236
2.5	1.25	T_∞	73.826	74.625	74.442	74.673	73.564	72.926
3	1.5	T_1	87.251	87.397	86.959	86.718	85.734	85.056
3	1.5	T_2	87.876	88.047	87.913	87.949	87.044	86.531
3	1.5	T_∞	86.972	87.537	87.567	87.720	86.856	86.370
3.5	1.75	T_1	94.679	94.772	94.591	94.483	93.812	93.387
3.5	1.75	T_2	95.095	95.334	95.244	95.197	94.669	94.434
3.5	1.75	T_∞	94.703	95.059	95.082	95.091	94.625	94.317
2.5	1.875	T_1	83.519	83.406	82.773	82.226	80.640	79.735
2.5	1.875	T_2	81.609	81.623	81.108	80.870	79.441	78.730
2.5	1.875	T_∞	79.385	79.775	79.556	79.566	78.256	77.523
3	2.25	T_1	93.382	93.250	92.773	92.661	91.459	90.911
3	2.25	T_2	92.282	92.343	92.017	91.874	90.860	90.342
3	2.25	T_∞	90.858	91.102	90.957	91.028	90.048	89.518
2	2	T_1	76.844	76.510	75.843	75.184	73.448	72.535
2	2	T_2	73.909	73.595	73.119	72.865	71.201	70.286
2	2	T_∞	71.034	71.129	71.082	71.061	69.551	68.646
2.5	2.5	T_1	90.891	90.735	90.222	89.833	88.490	87.760
2.5	2.5	T_2	88.849	88.947	88.437	88.120	86.826	86.100
2.5	2.5	T_∞	86.760	87.165	86.838	86.672	85.587	84.833

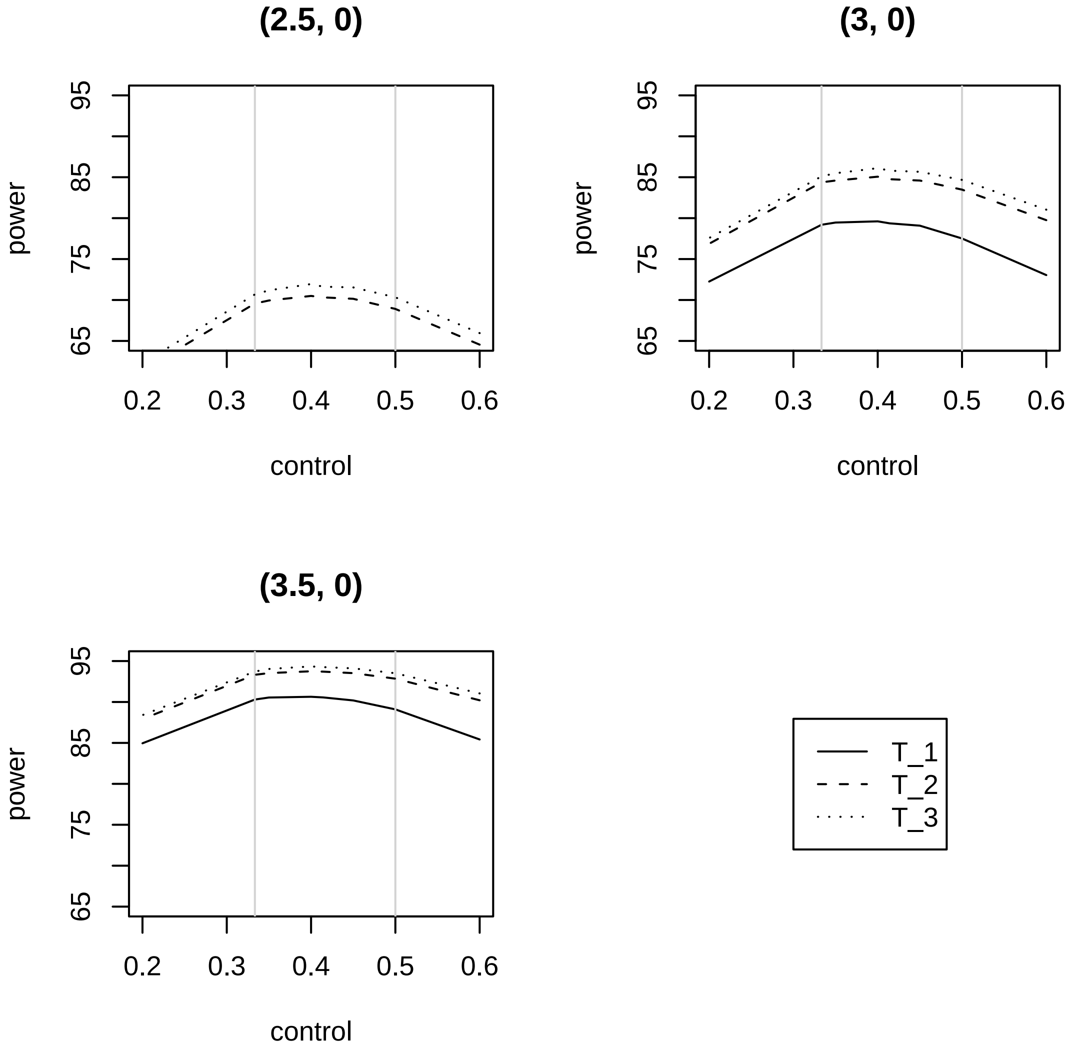

Figure 4.1: Power for different patient allocations with $d_1^* = 2.5, 3, 3.5$ and $\Delta_2 = 0$ (at 5%).

Table 4.2: Power for different proportions on the control at the 1% level.

			Control					
d_1^*	d_2^*	Test	.5	.45	$\sqrt{2}-1$	.4	.35	1/3
3.5	0	T_1	64.957	67.447	67.733	69.189	68.298	68.341
3.5	0	T_2	79.058	80.698	80.515	81.133	80.022	80.008
3.5	0	T_∞	80.797	82.159	82.270	82.736	81.511	81.357
4	0	T_1	80.608	82.550	82.939	83.996	83.281	83.302
4	0	T_2	90.277	91.467	91.179	91.589	90.988	90.776
4	0	T_∞	91.345	92.281	92.191	92.479	91.856	91.570
3	1.5	T_1	65.738	65.675	64.587	65.041	61.919	60.857
3	1.5	T_2	68.900	69.931	68.920	69.309	67.176	66.751
3	1.5	T_∞	66.949	68.440	68.206	68.700	67.068	66.586
3.5	1.75	T_1	80.774	80.628	79.766	80.165	77.529	76.782
3.5	1.75	T_2	83.732	84.440	83.871	84.093	82.430	82.157
3.5	1.75	T_∞	82.368	83.440	83.294	83.733	82.328	82.083
4	2	T_1	90.991	91.065	90.349	90.583	88.887	88.324
4	2	T_2	93.010	93.533	92.989	93.125	92.215	91.868
4	2	T_∞	92.236	92.923	92.740	92.983	92.232	91.955
3	2.25	T_1	79.619	79.080	77.885	78.026	74.890	73.719
3	2.255	T_2	77.321	77.668	76.417	76.436	73.955	73.274
3	2.25	T_∞	72.765	73.854	73.480	73.664	71.661	70.966
3.5	2.625	T_1	91.169	91.019	89.990	90.048	87.892	86.968
3.5	2.625	T_2	89.874	90.163	89.250	89.202	87.470	86.878
3.5	2.625	T_∞	86.800	87.451	87.158	87.331	85.744	85.365
2.5	2.5	T_1	75.027	74.560	73.307	73.410	70.036	68.958
2.5	2.5	T_2	70.658	70.938	69.581	69.361	66.510	65.699
2.5	2.5	T_∞	64.860	65.847	65.437	65.551	63.215	62.481
3	3	T_1	89.936	89.735	88.660	88.663	86.384	85.406
3	3	T_2	87.096	87.392	86.191	86.128	83.884	83.053
3	3	T_∞	82.740	83.412	83.061	83.144	81.205	80.449

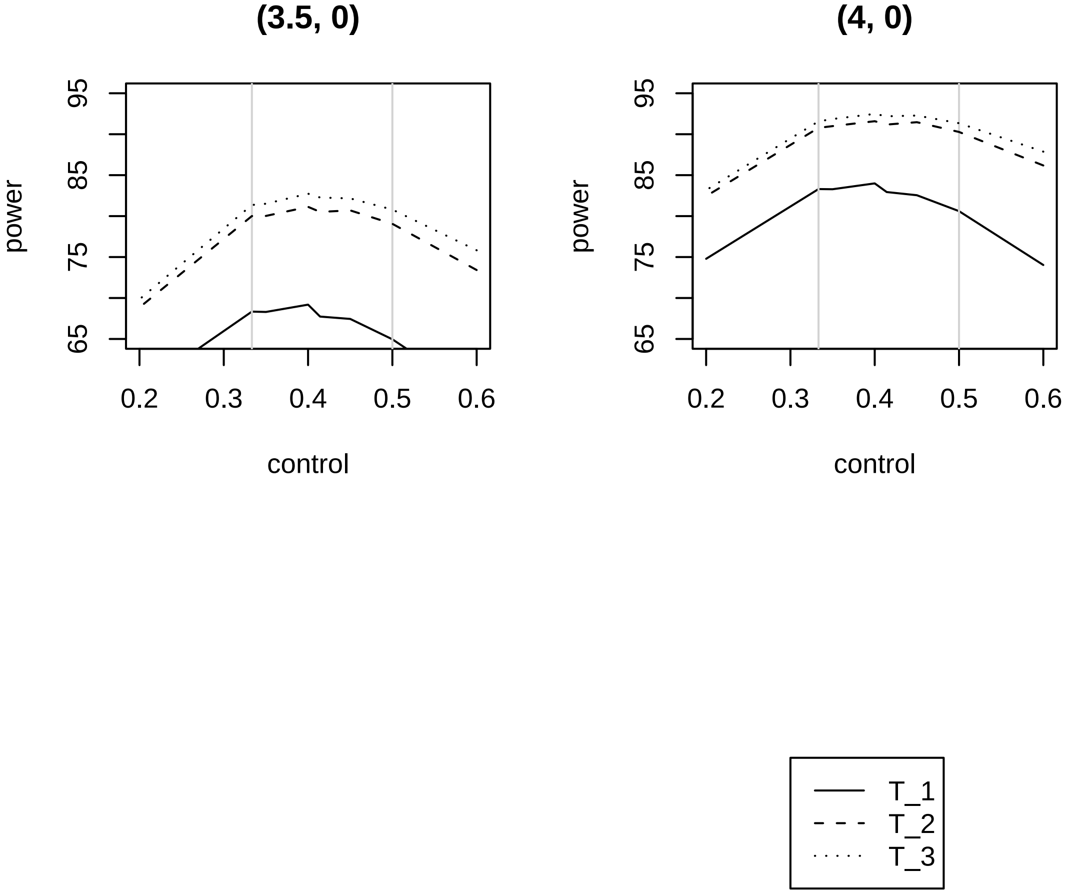

Figure 4.2: Power for different patient allocations with $d_1^* = 3.5, 4$ and $\Delta_2 = 0$ (at 1%).

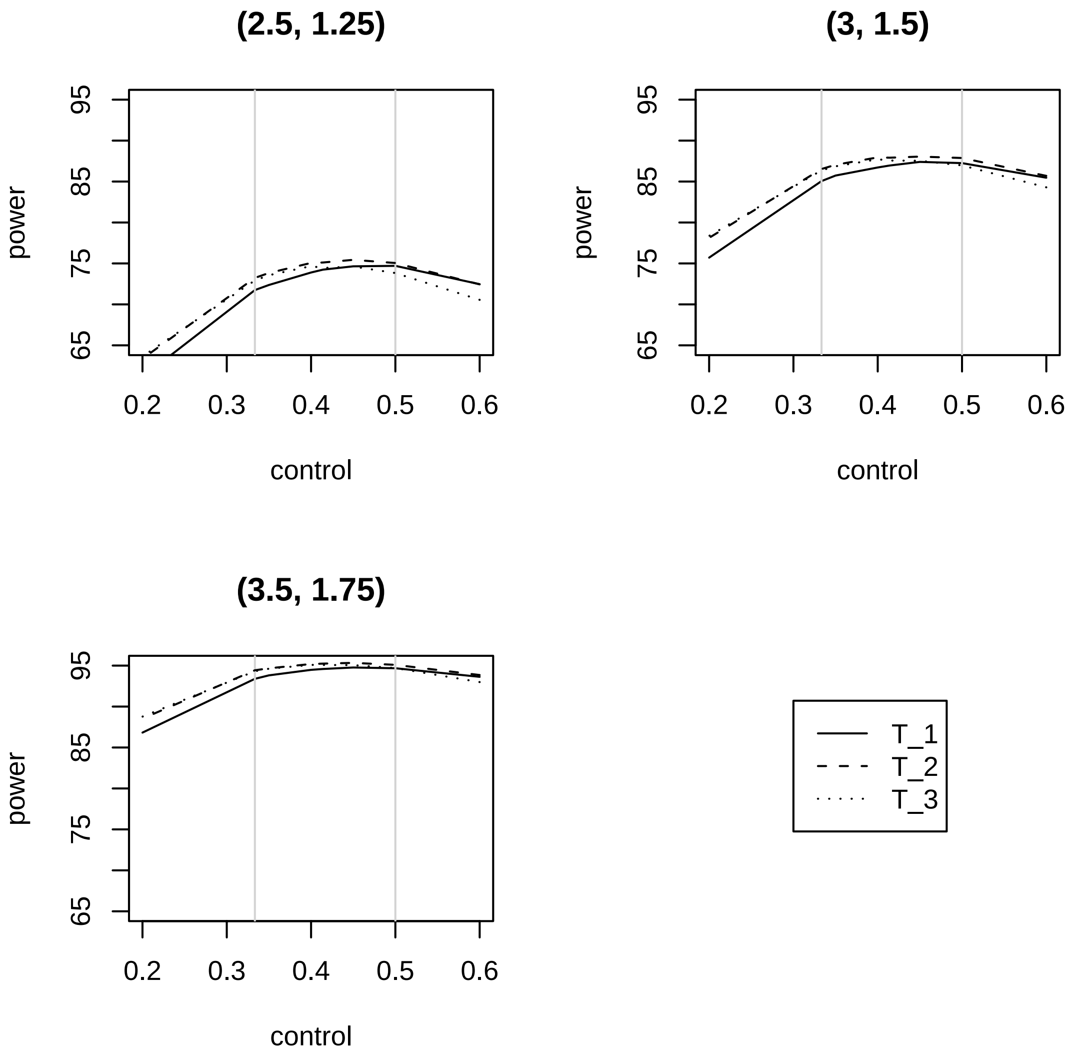

Figure 4.3: Power for different patient allocations with $d_1^* = 2.5, 3, 3.5$ and $\Delta_2 = \frac{1}{2}\Delta_1$ (at 5%).

points. First, for T_∞, the optimal allocation for testing at 5% remains close to $1-2\delta = 0.4$, but at 1% moves towards 0.45. However, for T_2, it is better to increase the proportion of subjects on the control, the optimum being close to 0.45. For T_1, it is even higher, being somewhere between 0.45 and 0.5. Figures 4.3 and 4.4 show that the power curves even cross and, at its optimal allocation, T_1 is more powerful than T_∞ for testing at 5%, even though it is considerably less powerful when equal allocation is used. Similarly, at its optimal allocation, T_2's advantage over T_∞ is somewhat greater than at equal allocation, especially when testing at the 1% level.

Results with $\Delta_2 = 3\Delta_1/4$ at the 5% level are shown in Table 4.1 and Figure 4.5. The results for the 1% level are given in Table 4.2 and Figure 4.6. In this case, the optimal

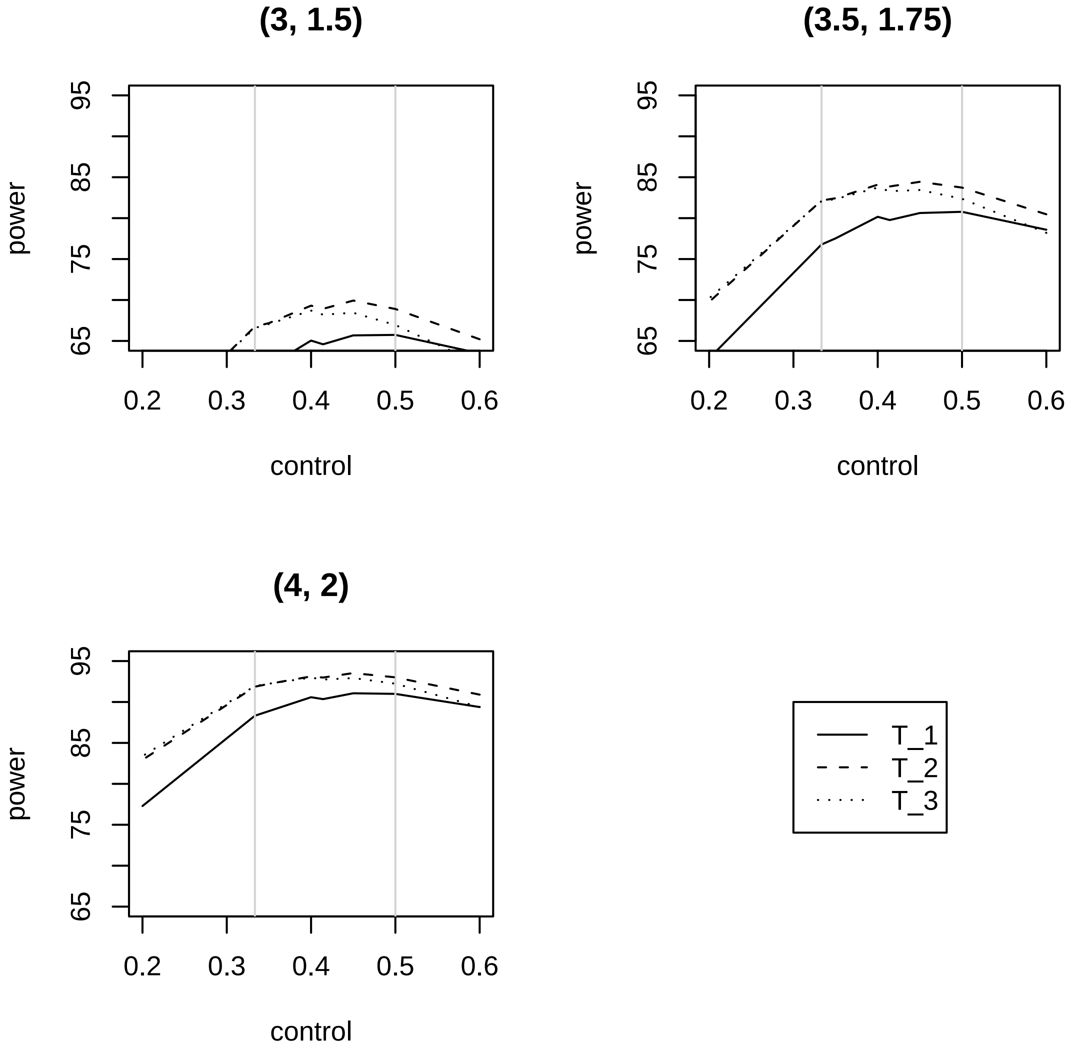

Figure 4.4: Power for different patient allocations with $d_1^* = 3, 3.5, 4$ and $\Delta_2 = \frac{1}{2}\Delta_1$ (at 1%).

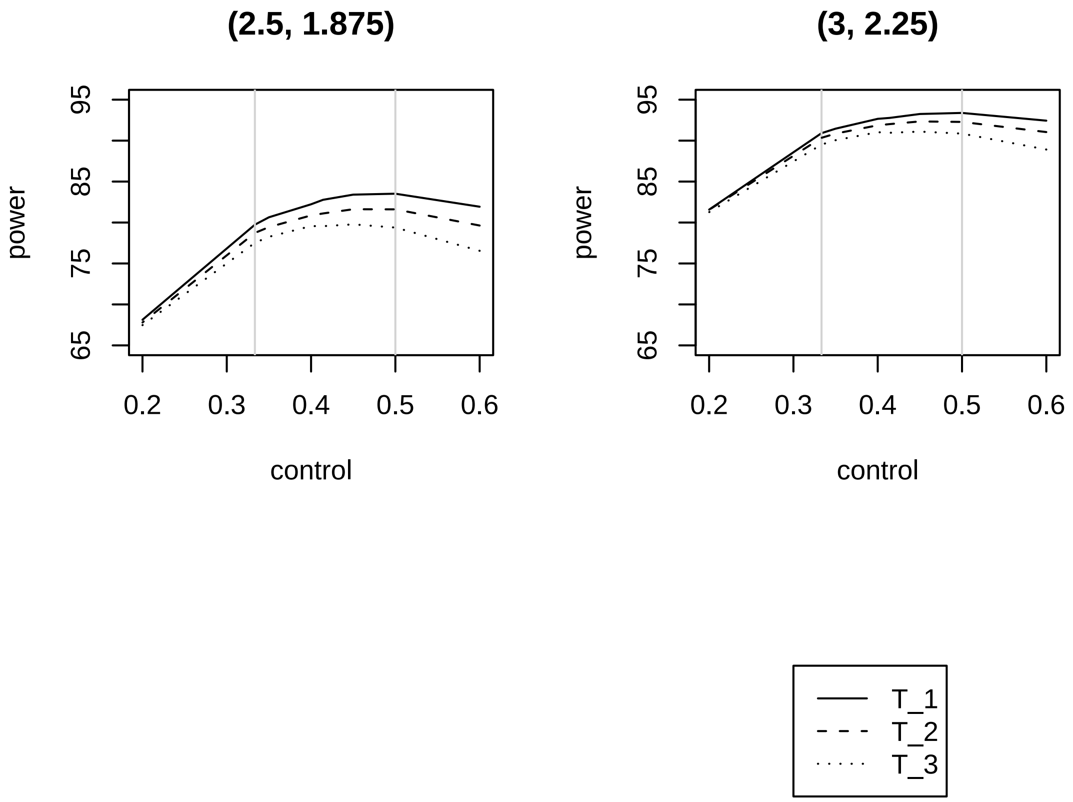

Figure 4.5: Power for different patient allocations with $d_1^* = 2.5, 3$ and $\Delta_2 = \frac{3}{4}\Delta_1$ (at 5%).

allocation on control is close to 0.45 for all tests, or between 0.45 and 0.5 for T_1. Again the advantages of T_1 and T_2 over T_∞ are greater at their optimal allocation than at equal allocation and are greater at 1% than at 5%. In Table 4.2, for example, we see that T_2 can have almost 4% higher power than T_∞, much greater than any advantage of T_∞ over T_2 that we have observed.

Finally, results for $\Delta_2 = \Delta_1$ are shown in Table 4.1 and Figure 4.7 for the 5% level of significance and in Table 4.2 and Figure 4.8 for the 1% level of significance. As expected, the optimal allocation for T_1 is close to 0.5 and for T_2 it gets closer to 0.5 than 0.45 for testing at 5%, though not for 1% tests. For T_∞, the optimal proportion on the control is around 0.45. Once again, we see that the differences in power are bigger when testing at the 1% level, with T_2 giving more than 5% higher power than T_∞ in some cases in Table

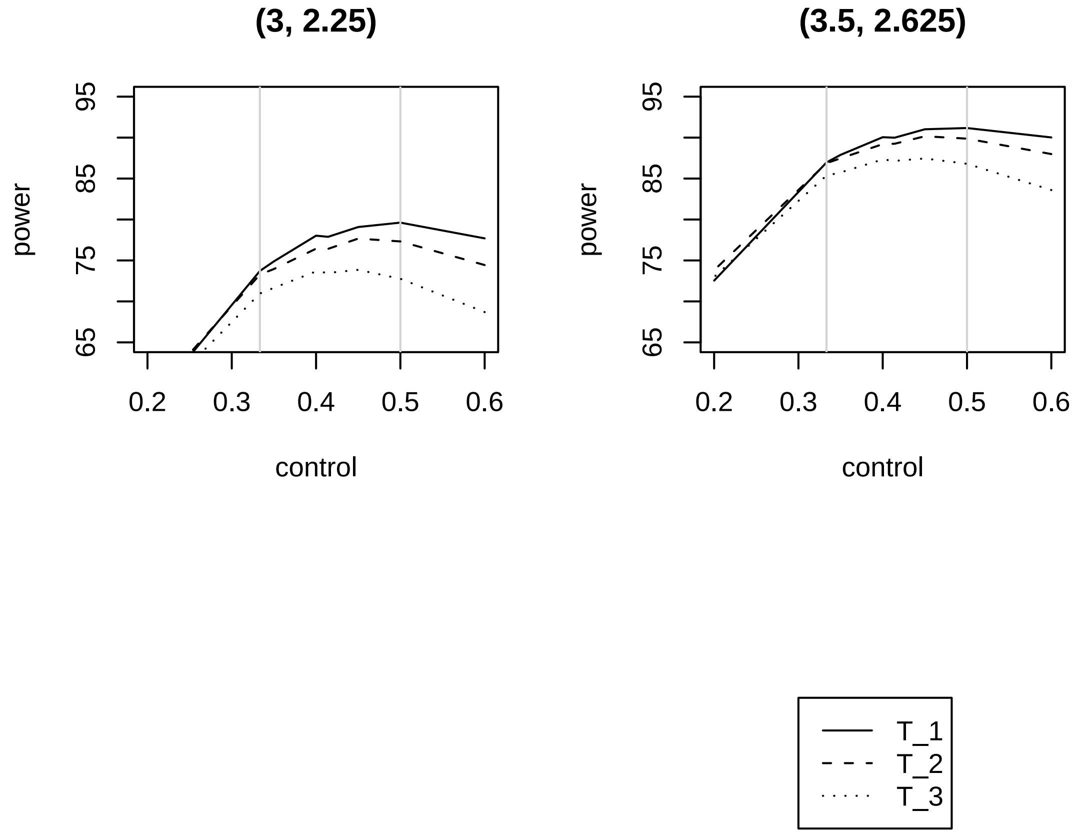

Figure 4.6: Power for different patient allocations with $d_1^* = 3, 3.5$ and $\Delta_2 = \frac{3}{4}\Delta_1$ (at 1%).

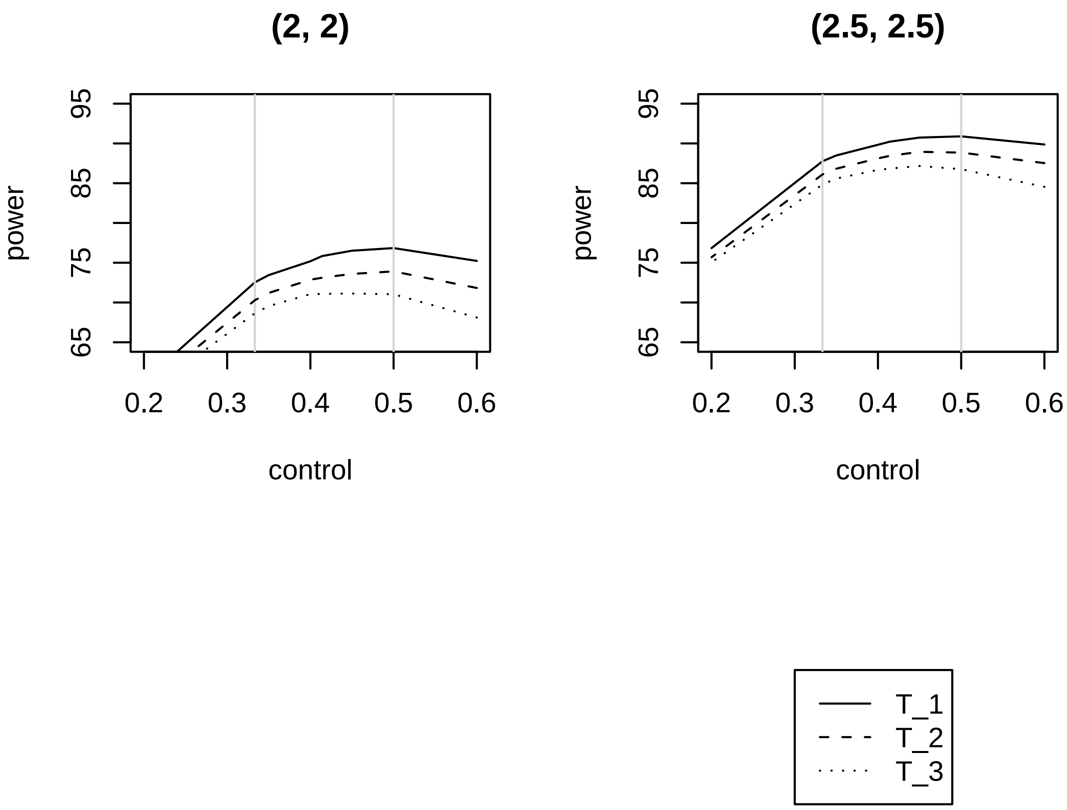

Figure 4.7: Power for different patient allocations with $d_1^* = 2, 2.5$ and $\Delta_2 = \Delta_1$ (at 5%).

4.2.

It can be seen that, along with the values of Δ_1 and Δ_2, the optimality of allocation depends on which test statistic is being used and on the significance level. With lower values of Δ_2 the optimal allocation is around $1 - 2\delta = 0.4$. As seen before, T_∞ works best for this configuration. As Δ_2 gets bigger relative to Δ_1 the optimal allocation changes towards $1 - 2\delta = 0.5$ for T_1 and, more slowly, in the same direction for the other tests. For T_∞ this occurs later than for the other two tests. For these configurations, T_2 is better than T_∞ and, for the larger values of Δ_2, T_1 is even better.

The fact that the optimal allocation for T_k gets closer to $1 - 2\delta = 0.5$ for smaller k and for larger Δ_2 is not surprising. It is known that if $\Delta_2 = \Delta_1$, S_1^0 with $1 - 2\delta = 0.5$ is optimal and T_1 and T_2 can be seen as compromises between S_1^0 and T_∞. It was less predictable

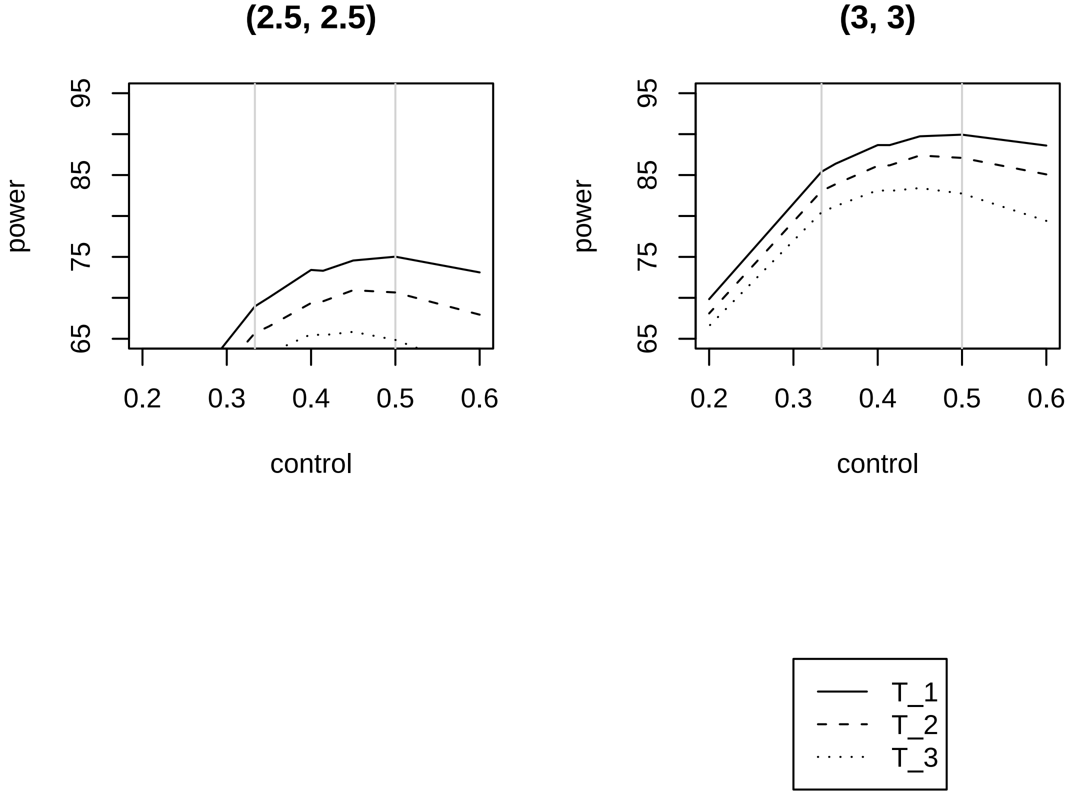

Figure 4.8: Power for different patient allocations with $d_1^* = 2.5, 3$ and $\Delta_2 = \Delta_1$ (at 1%).

that the differences would be so much greater at 1% than at 5%, but this is probably due to the impact of the differences being greater in the tails of the distributions.

In all cases the balanced design ($\delta = 1 - 2\delta = 1/3$) is sub-optimal. Within the range of values of δ suggested, differences in the power are reasonably small but, especially for testing at the 1% level, not negligible. Of course, more extreme values of δ would lead to much lower powers.

The results obtained here seem to confirm that both T_2 and T_∞ are reasonable testing procedures, but suggest that, if unequal allocations can be used, T_2 has worthwhile advantages unless there is a strong prior belief that only one treatment will be better than the control.

4.3.3 Approximating the Allocation which Maximises Power

The simulated values presented in Subsection 4.3.2 give a good idea of the optimal value of δ. To get a smoother approximation of the dependence of the power on δ, we fit a quadratic regression of power at 5% on $1 - 2\delta$ and check for lack of fit by adding a cubic term. If the quadratic model is a good fit, then we estimate the maximum by differentiating the quadratic function and equating to zero. The R^2 values were all at least 90%, the quadratic terms were all highly significant and in only one case was a cubic term (marginally) significant. Hence, the quadratic curve seems to give a very good approximation to the dependence of the power on δ.

The estimated optimal proportions on the control treatment are shown in Table 4.3. The first section of the table confirms the result of Dunnett (1955) that, for T_∞ with the least favourable configuration, the optimal proportion on the control is slightly less than $\sqrt{2} - 1 \approx 0.4142$. Bechhofer and Tamhane (1983), using an approximation to the probabilities, found the optimal proportion on the control to be 0.399. Our simulations give systematically slightly higher optimal proportions to the control. Note also that our cutpoint for the test is obtained internally from the simulations and so tends to the exact value as the number of simulations increases, while the approximate cutpoint of Dunnett, used by Bechhofer and Tamhane, is known to give a test of size slightly greater than α. The results are close enough to give essentially the same practical conclusions. It can be seen that for larger Δ_2, the optimal proportion on the control treatment for T_∞ increases steadily to 0.452 when $\Delta_2 = \Delta_1$.

The optimal allocation for the other test statistics is similar, but for the LFC they require a

Table 4.3: Estimate of optimum allocation in the control arm for different configurations of d_1^*, d_2^* for maximising power of tests at 5%.

Configuration	T_1	T_2	T_∞
(2.5,0)	0.3900	0.4049	0.4101
(3,0)	0.3852	0.3977	0.4071
(3.5,0)	0.3865	0.4005	0.4082
(2.5,1.25)	0.4767	0.4517	0.4306
(3,1.5)	0.4702	0.4502	0.4277
(3.5,1.75)	0.4617	0.4441	0.4287
(2.5,1.875)	0.4865	0.4759	0.4474
(3,2.25)	0.4873	0.4695	0.4471
(2,2)	0.4982	0.4851	0.4521
(2.5,2.5)	0.4886	0.4739	0.4524

slightly smaller proportion of subjects on the control arm. However, the optimal allocation changes more quickly as Δ_2 increases. For T_2 the optimum is about 0.40 for the LFC, but reaches around 0.48 at the MFC. For T_1 the change is even greater, the optimum going from about 0.39 at the LFC to almost 0.5 at the MFC.

Although these results are interesting, they do not make a large difference in practice, as was seen in Subsection 4.3.2. Nothing here contradicts Dunnett's advice that allocating a proportion $\sqrt{2}-1$ to the control gives powers which are nearly optimal for T_∞ across a range of configurations. We could add that the same is also true for the other test statistics studied, unless there is a strong prior belief that the Δ_is will be nearly equal.

4.3.4 Effect of Allocation Proportion on Expected Loss

As noted in Chapter 3, the power does not tell the full story about the properties of the test statistics. In particular, if we assume that a selection will be made after performing the test, the probability of incorrect selection is also of interest. However, as was argued in Chapter 3, the probability itself is insufficient, as it does not take account of the consequences of an incorrect decision. There we also calculated the expected loss, defined as $E(\text{Loss})$, where

$$\text{Loss} = \begin{cases} 0 & \text{if we select treatment 1;} \\ \Delta_1 - \Delta_2 & \text{if we select treatment 2;} \\ \Delta_1 & \text{if we fail to reject } H_0. \end{cases}$$

It is perfectly possible to choose a patient allocation to minimise the expected loss, rather than to maximise the power. Similarly, the sample size could be chosen to achieve a specific expected loss and this is discussed in Subsection 4.4.2.

When running the simulations described in Subsection 4.3.1, we calculated the expected loss, as well as the power. The results for testing at the 5% level of significance, when $\Delta_2 = 0$ are shown in Table 4.4 and Figure 4.9. The conclusions from these are that the optimal allocation is to have a proportion around 0.4 on the control and that T_∞ is consistently slightly better than T_2 and considerably better than T_1. These are similar to the conclusions drawn from comparing the powers in Subsection 4.3.2.

The corresponding results for the 1% significance level are given in Table 4.5 and Figure 4.10. These show that the optimal allocation is close to 0.4 for all the test statistics.

The results at 5% when $\Delta_2 = \Delta_1/2$ are shown in Table 4.4 and Figure 4.11. For this configuration, the optimal proportion on control is close to 0.4 for T_∞, slightly greater than 0.4 for T_2 and 0.45 for T_1. T_2 consistently gives the smallest expected loss, but T_∞ is not much worse. This is slightly different from the results found for power in that, for T_2, slightly fewer subjects should be allocated to the control to minimise the expected loss than to maximise the power.

The corresponding results for testing at the 1% level are shown in Table 4.5 and Figure 4.12. These show a more complex pattern, with the optimal allocation for T_∞ being around 0.4 and for T_1 and T_2 decreasing from around 0.45 to around 0.4 as Δ_1 increases. Note that this is rather different from the conclusions in Subsection 4.3.2 for finding the allocation which maximises the power, even though the same simulated data sets are being used. Hence the decision about which allocation to use should depend on whether we are aiming to maximise the power or make the best treatment selection. We also note that, in terms of expected loss, the advantage of T_2 over T_∞ is greater at 1% than at 5%, but perhaps not so dramatically as when we considered power. Perhaps this is because T_∞ avoids type-III/IV errors better than T_2.

The expected losses when $\Delta_2 = 3\Delta_1/4$ when testing at 5% are shown in Table 4.4 and Figure 4.13. The corresponding results for 1% are shown in Table 4.5 and Figure 4.14.

Table 4.4: Expected loss for different proportions on the control at the 5% level.

			Control					
d_1^*	d_2^*	Test	.5	.45	$\sqrt{2}-1$	.4	.35	1/3
2.5	0	T_1	0.9771	0.9348	0.9229	0.9227	0.9241	0.9298
2.5	0	T_2	0.7914	0.7574	0.7496	0.7457	0.7575	0.7668
2.5	0	T_∞	0.7540	0.7209	0.7158	0.7082	0.7257	0.7382
3	0	T_1	0.6843	0.6343	0.6234	0.6156	0.6191	0.6271
3	0	T_2	0.5035	0.4678	0.4606	0.4514	0.4647	0.4711
3	0	T_∞	0.4664	0.4348	0.4288	0.4197	0.4375	0.4491
3.5	0	T_1	0.3850	0.3457	0.3318	0.3290	0.3314	0.3404
3.5	0	T_2	0.2533	0.2288	0.2214	0.2198	0.2270	0.2343
3.5	0	T_∞	0.2299	0.2079	0.2017	0.1990	0.2092	0.2203
2.5	1.25	T_1	0.7586	0.7514	0.7553	0.7592	0.7875	0.8008
2.5	1.25	T_2	0.7370	0.7203	0.7222	0.7196	0.7418	0.7549
2.5	1.25	T_∞	0.7584	0.7322	0.7320	0.7231	0.7436	0.7573
3	1.5	T_1	0.5084	0.4909	0.4969	0.4988	0.5183	0.5352
3	1.5	T_2	0.4816	0.4638	0.4620	0.4553	0.4725	0.4848
3	1.5	T_∞	0.5008	0.4733	0.4672	0.4577	0.4742	0.4863
3.5	1.75	T_1	0.2959	0.2776	0.2772	0.2743	0.2883	0.3000
3.5	1.75	T_2	0.2769	0.2541	0.2506	0.2459	0.2549	0.2605
3.5	1.75	T_∞	0.2866	0.2603	0.2536	0.2472	0.2542	0.2625
2.5	1.875	T_1	0.5654	0.5634	0.5755	0.5859	0.6200	0.6406
2.5	1.875	T_2	0.6050	0.6002	0.6095	0.6130	0.6434	0.6592
2.5	1.875	T_∞	0.6536	0.6407	0.6433	0.6415	0.6692	0.6852
3	2.25	T_1	0.3797	0.3763	0.3869	0.3858	0.4161	0.4301
3	2.25	T_2	0.4076	0.3989	0.4049	0.4049	0.4294	0.4424
3	2.25	T_∞	0.4451	0.4316	0.4324	0.4269	0.4505	0.4640
2	2	T_1	0.4631	0.4698	0.48314	0.4963	0.5310	0.5493
2	2	T_2	0.5218	0.5281	0.5376	0.5427	0.5760	0.5943
2	2	T_∞	0.5793	0.5774	0.5784	0.5788	0.6090	0.6271
2.5	2.5	T_1	0.2277	0.2316	0.2445	0.2542	0.2878	0.3060
2.5	2.5	T_2	0.2788	0.2763	0.2891	0.2970	0.3294	0.3475
2.5	2.5	T_∞	0.3310	0.3209	0.3291	0.3332	0.3603	0.3792

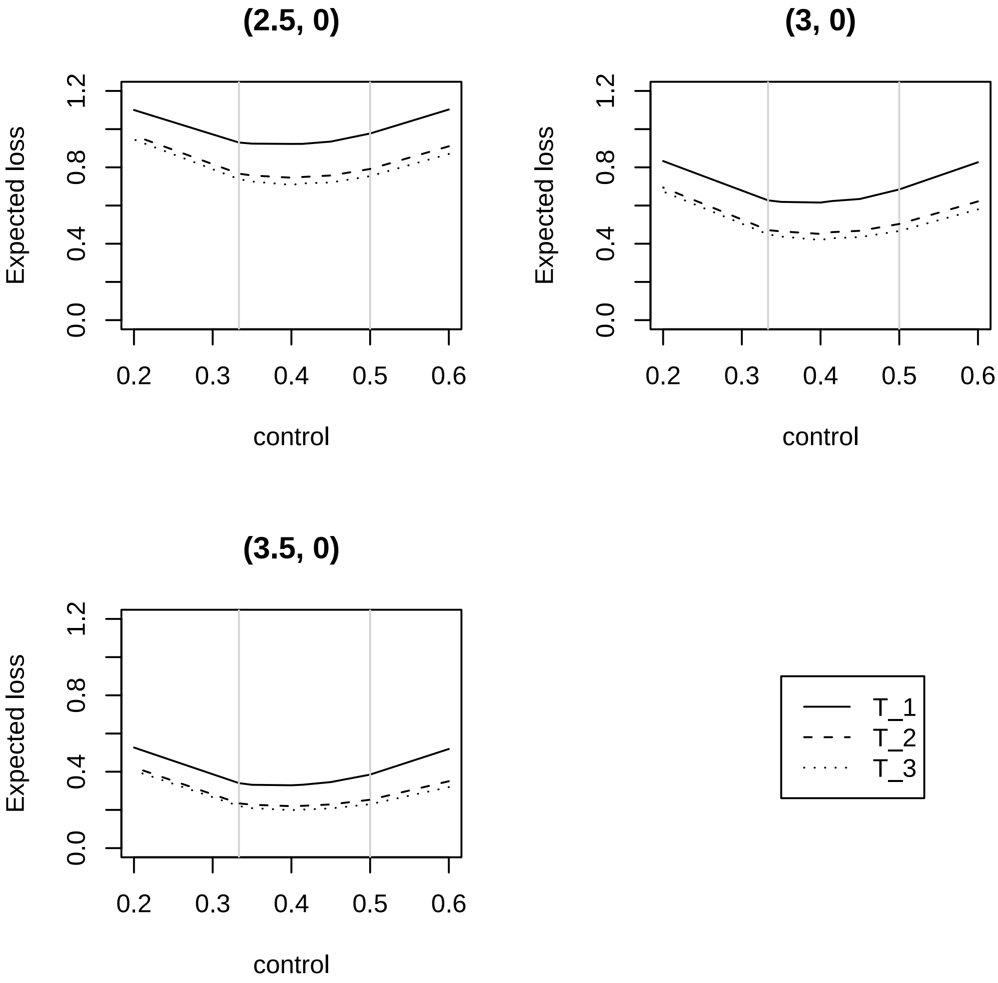

Figure 4.9: Expected loss for different patient allocations (5%) with $d_1^* = 2.5, 3, 3.5$ and $\Delta_2 = 0$.

Table 4.5: Expected loss for different proportions on the control at 1%.

			Control					
d_1^*	d_2^*	Test	.5	.45	$\sqrt{2}-1$	.4	.35	1/3
3.5	0	T_1	1.2289	1.1410	1.1304	1.0791	1.1101	1.1085
3.5	0	T_2	0.7347	0.6767	0.6827	0.6608	0.6995	0.7000
3.5	0	T_∞	0.6732	0.6252	0.6210	0.6046	0.6473	0.6525
4	0	T_1	0.7766	0.6984	0.6827	0.6404	0.6688	0.6679
4	0	T_2	0.3895	0.3416	0.3531	0.3366	0.3605	0.3690
4	0	T_∞	0.3466	0.3090	0.3126	0.3009	0.3258	0.3372
3	1.5	T_1	1.12589	1.1169	1.1412	1.1261	1.2089	1.2359
3	1.5	T_2	1.0164	0.9761	0.9983	0.9855	1.0392	1.0488
3	1.5	T_∞	1.0604	1.0089	1.0100	0.9950	1.0358	1.0476
3.5	1.75	T_1	0.7685	0.7602	0.7817	0.7625	0.8451	0.8658
3.5	1.75	T_2	0.6547	0.6184	0.6296	0.6176	0.6667	0.6720
3.5	1.75	T_∞	0.6913	0.6444	0.6420	0.6237	0.6648	0.6694
4	2	T_1	0.4414	0.4234	0.4440	0.4302	0.4886	0.5078
4	2	T_2	0.3554	0.3204	0.3342	0.3247	0.3521	0.3627
4	2	T_∞	0.3790	0.3392	0.3398	0.3266	0.3485	0.3569
3	2.25	T_1	0.7673	0.7770	0.8070	0.8011	0.8849	0.9150
3	2.255	T_2	0.8233	0.8084	0.8398	0.8372	0.9018	0.9178
3	2.25	T_∞	0.9447	0.9099	0.9172	0.9101	0.9619	0.9785
3.5	2.625	T_1	0.4896	0.4870	0.5161	0.5113	0.5754	0.6027
3.5	2.625	T_2	0.5265	0.5100	0.5343	0.5327	0.5828	0.5985
3.5	2.625	T_∞	0.6227	0.5951	0.5989	0.5907	0.6362	0.6449
2.5	2.5	T_1	0.6243	0.6360	0.6673	0.6648	0.7491	0.7761
2.5	2.5	T_2	0.7336	0.7266	0.7605	0.7660	0.8373	0.8575
2.5	2.5	T_∞	0.8785	0.8538	0.8641	0.8612	0.9196	0.9380
3	3	T_1	0.3019	0.3080	0.3402	0.3401	0.4085	0.4378
3	3	T_2	0.3871	0.3782	0.4143	0.4162	0.4835	0.5084
3	3	T_∞	0.5178	0.4976	0.5082	0.5057	0.5639	0.5865

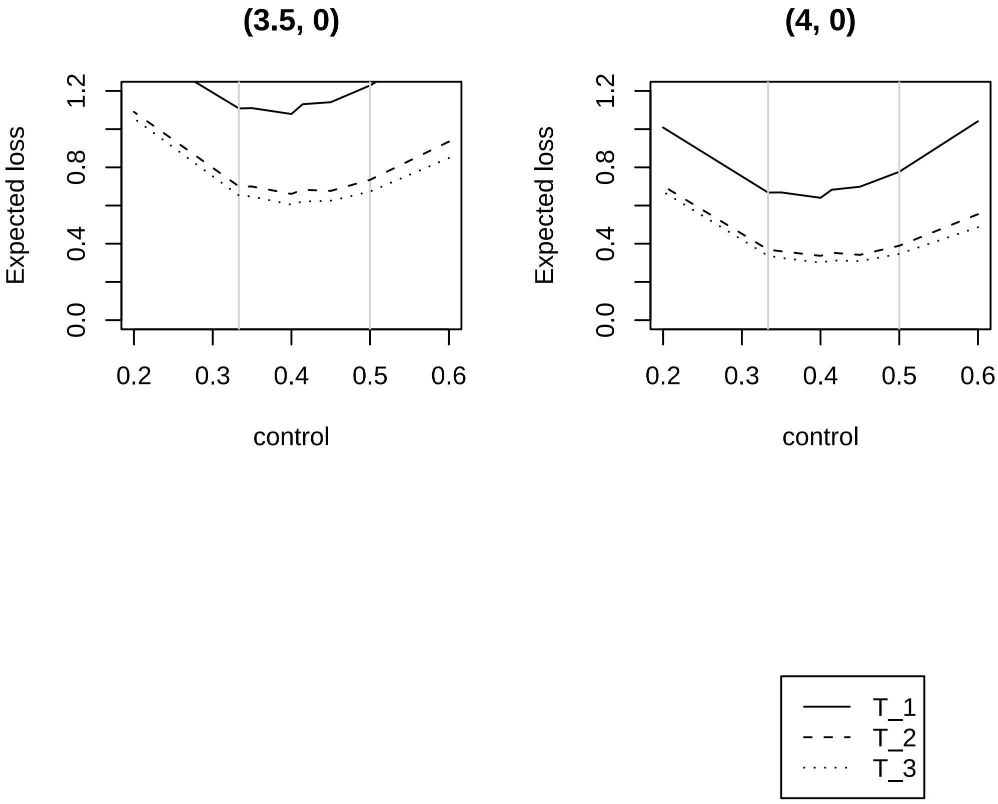

Figure 4.10: Expected loss for different patient allocations (1%) with $d_1^* = 3.5, 4$ and $\Delta_2 = 0$.

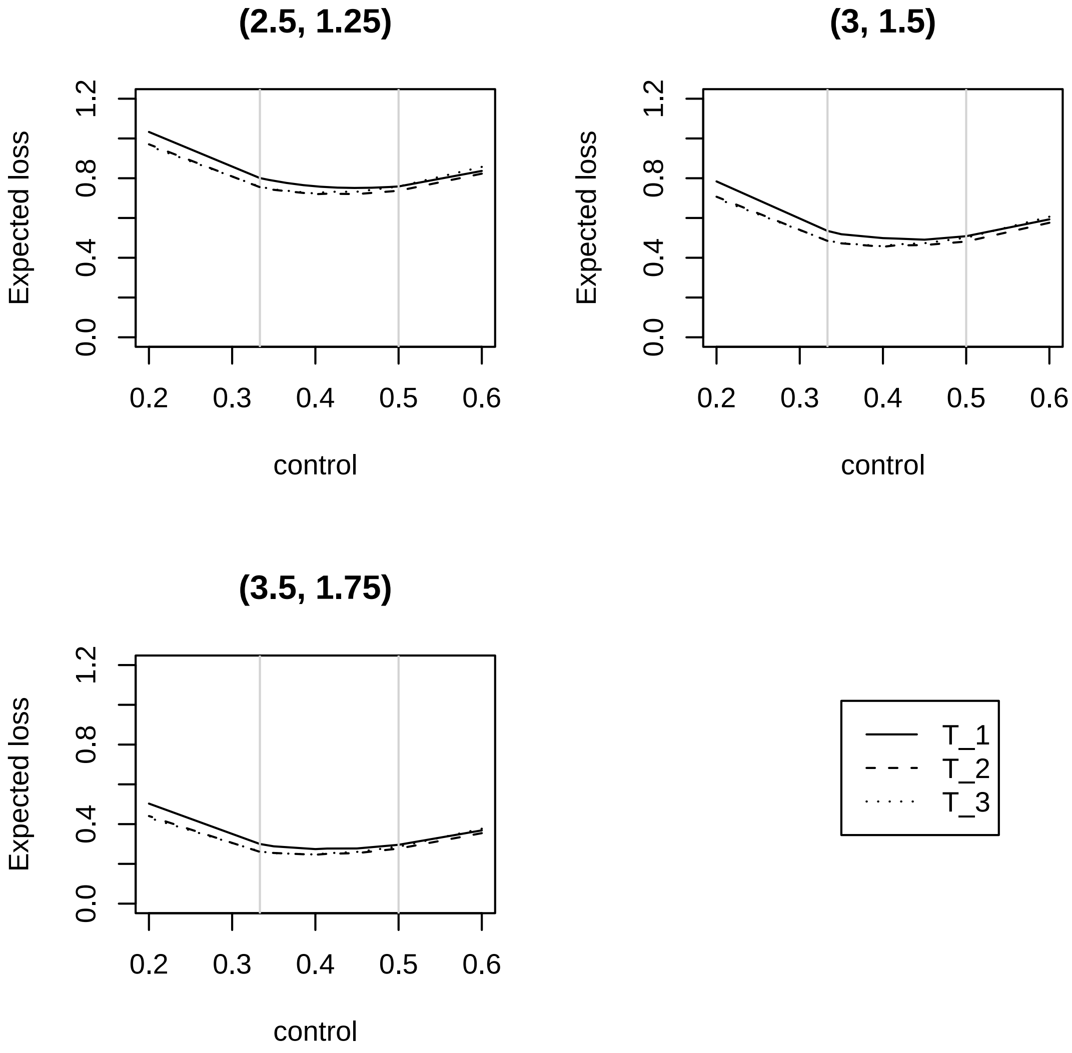

Figure 4.11: Expected loss for different patient allocations (5%) with $d_1^* = 2.5, 3, 3.5$ and $\Delta_2 = \frac{1}{2}\Delta_1$.

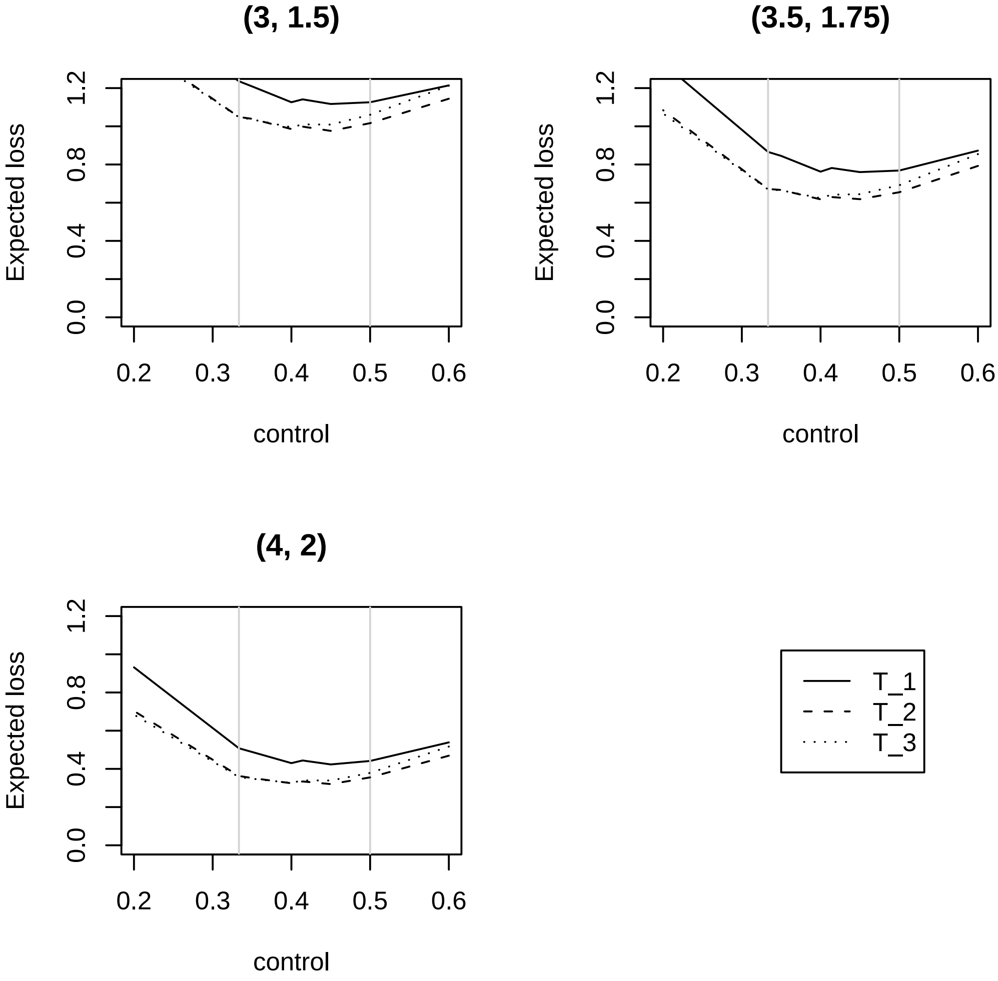

Figure 4.12: Expected loss for different patient allocations (1%) with $d_1^* = 3, 3.5, 4$ and $\Delta_2 = \frac{1}{2}\Delta$.

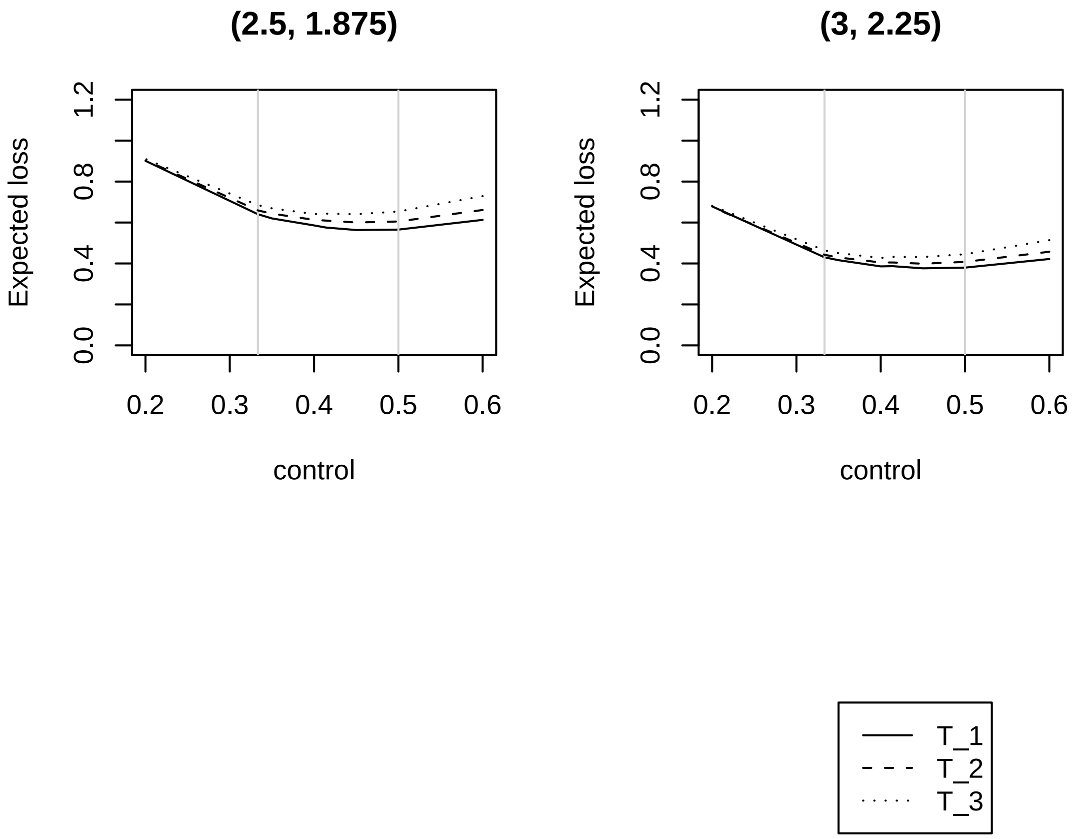

Figure 4.13: Expected loss for different patient allocations (5%) with $d_1^* = 2.5, 3$ and $\Delta_2 = \frac{3}{4}\Delta_1$.

This shows again that, as Δ_1 increases, the optimum allocation changes from close to 0.45 to closer to 0.4 for all tests. It also shows again that the advantage of T_2 over T_∞ is greater for testing at 1%.

Finally, the expected losses for $\Delta_2 = \Delta_1$ are shown in Table 4.4 and Figure 4.15 for testing at 5% and Table 4.5 and Figure 4.16 for testing at 1%. The optimal allocation seems to be close to 0.5 for T_1, between 0.45 and 0.5 for T_2 and close to 0.45 for T_∞. Again, the benefit of using T_2 over T_∞ seems greater at 1% than at 5% and certainly is not negligible.

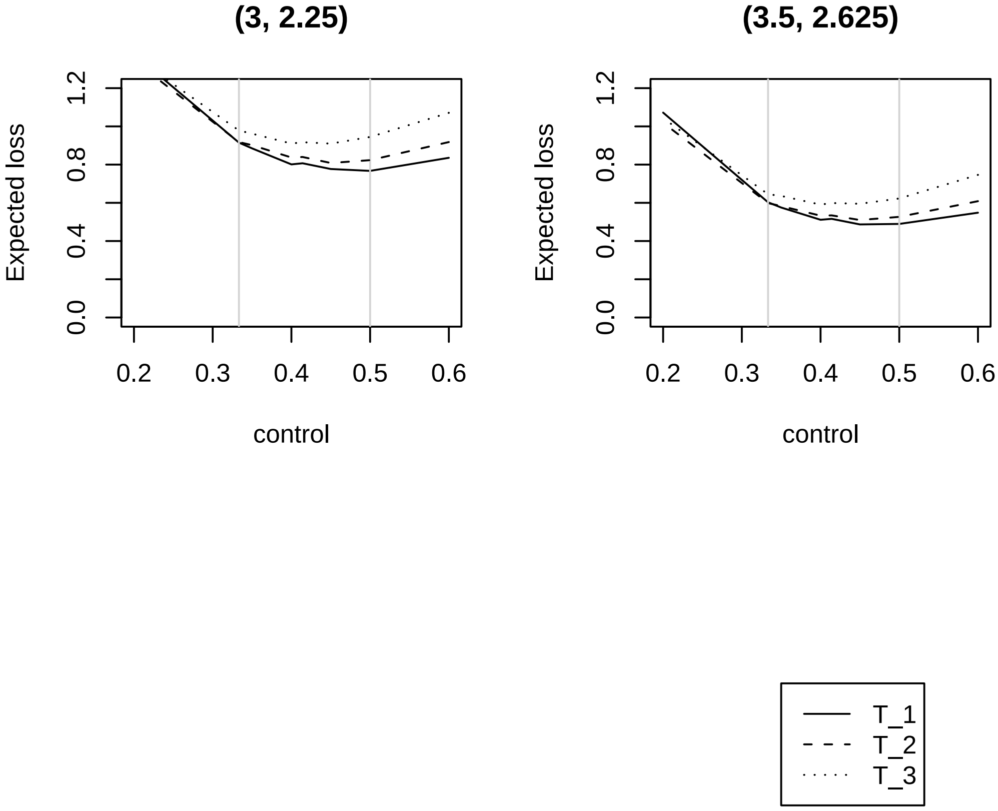

Figure 4.14: Expected loss for different patient allocations (1%) with $d_1^* = 3, 3.5$ and $\Delta_2 = \frac{3}{4}\Delta_1$.

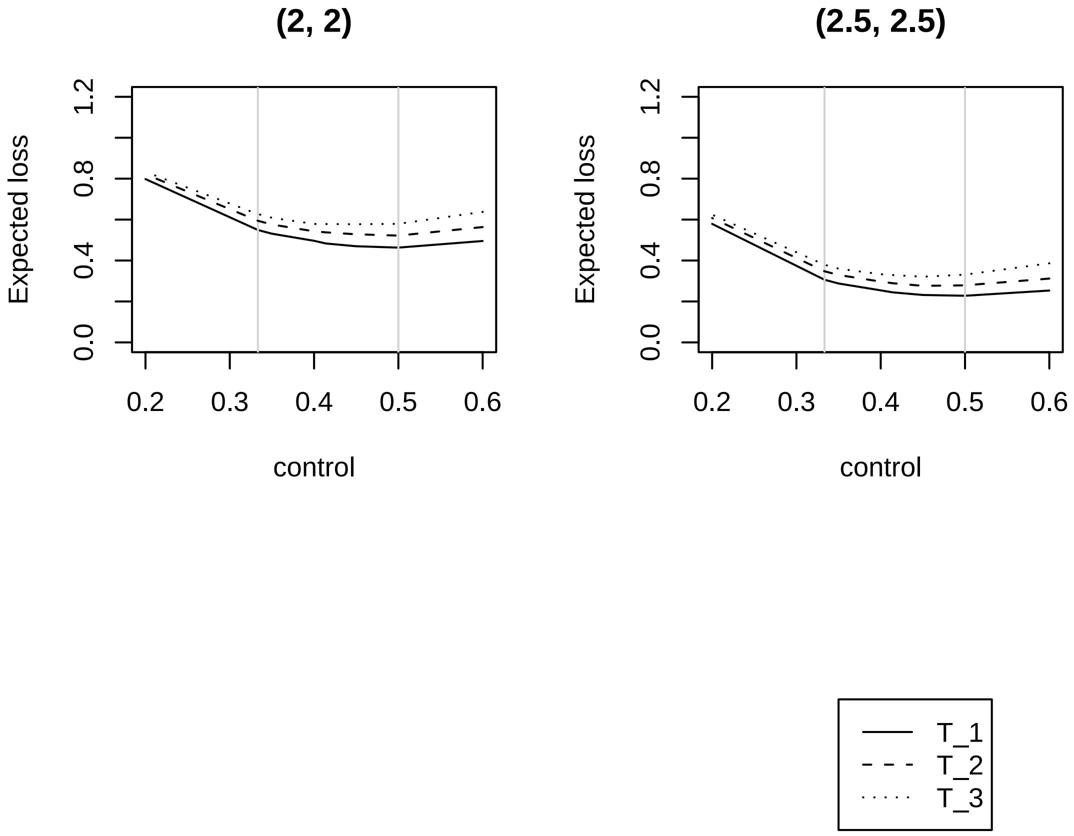

Figure 4.15: Expected loss for different patient allocations (5%) with $d_1^* = 2, 2.5$ and $\Delta_2 = \Delta_1$.

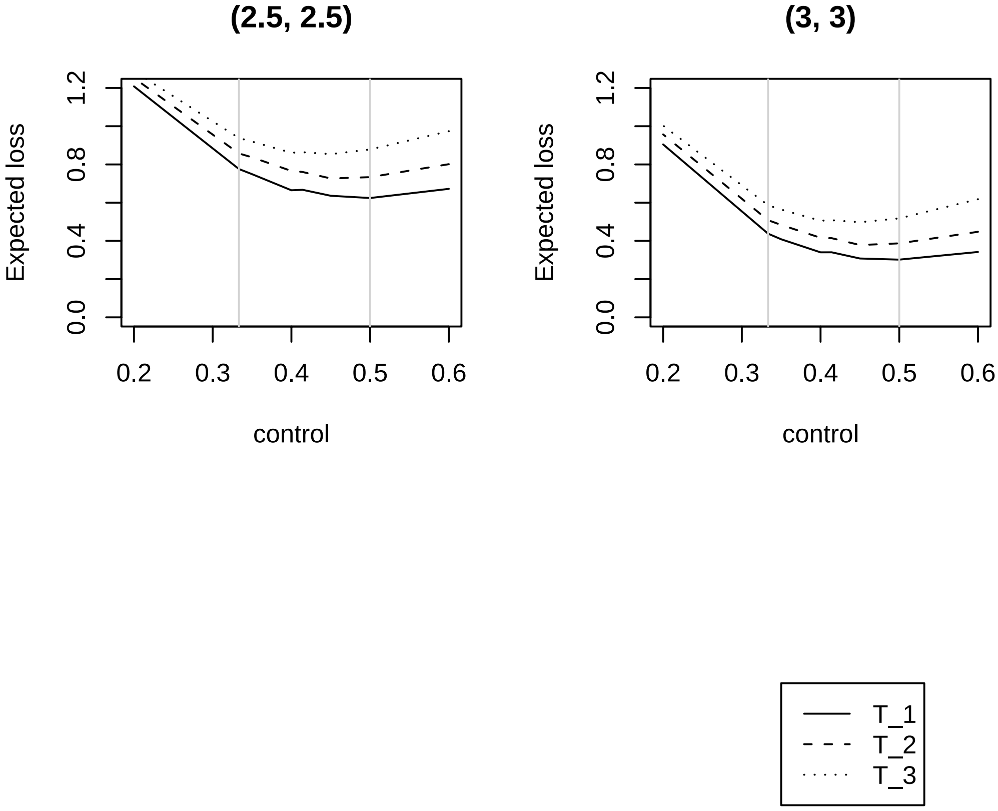

Figure 4.16: Expected loss for different patient allocations (1%) with $d_1^* = 2.5, 3$ and $\Delta_2 = \Delta_1$.

4.4 Sample Size Calculations

Whether the allocation proportion is chosen to maximise the power, minimise the expected loss or for other practical reasons, its choice can precede the calculation of a sample size for the trial. In this section we show how to find a sample size to achieve a pre-specified level of power or expected loss, given an allocation proportion, a level of significance and prior values of the treatment differences.

Up to now, we have worked mainly with versions of the treatment effects which are scaled by the number of observations. To choose a sample size, we need to relate them to the original scale of the response variable, so that realistic prior differences between the treatments can be used. Let

$$E(Z_i) = E\left(\sqrt{N}\frac{\theta_i}{\sigma}\right),$$

where σ is such that $Var(Z_i) = 1$. Then $\Delta_i = \sqrt{N}\theta_i$.

As above, we have $d_i^* = (\sqrt{2}-1)\Delta_i$ and so

$$d_i^* = \theta_i\sqrt{N}(\sqrt{2}-1).$$

Thus for any prior values of θ_1 and θ_2, it is straightforward to find the required d_1^* and d_2^* as functions of N. Note that in the case when $Z_i = \sqrt{N}(\hat{Y}_i - \hat{Y}_0)/\sigma$, θ_1 and θ_2 are on the same scale as Y_{ij}, so that the variance of the original responses affects the sample size through the scaling, i.e. for the same unscaled difference, θ_i is smaller if the variance is larger. We then need to find the d_1^* and d_2^* which achieve the required power or expected loss, using the same simulations as in Section 4.3, so that we can find the required sample size by equating the expressions for either d_1^* or d_2^* to give

$$N = \frac{d_i^{*2}}{\theta_i^2(\sqrt{2}-1)^2}. \tag{4.1}$$

Since the simulations were set up to find the power or expected loss for specific values of d_1^* and d_2^*, we use repeated interpolation to find the d_1^* and d_2^* which achieve a specific power or expected loss. Note that, since the same set of simulations is used for each $\boldsymbol{\Delta}$ for a specific allocation ratio, no new simulation runs are required.

4.4.1 Sample size to achieve specified power

As an example, the top left portion of Table 4.6 shows the values of d_1^* required to achieve 80% power when $\Delta_2 = \Delta_1/2$ for three different allocations, equal numbers on each arm, the

Table 4.6: Effect sizes yielding 80% power and sample size ratio to achieve 80% power for different δ for the tests at the 5% level with $d_2^* = d_1^*/2$.

	d_1^*			Ratio of required sample sizes	
δ	T_1	T_2	T_∞	T_1 v. T_2	T_∞ v. T_2
1/3 (equal)	2.7878	2.7336	2.7418	1.0400	1.0060
$\sqrt{2}-1$ (root)	2.7109	2.6769	2.6979	1.0256	1.0158
1/4 (half)	2.7046	2.6913	2.7273	1.0098	1.0269
Ratio of required sample sizes					
equal v. root	1.0575	1.0428	1.0328		
half v. root	0.9953	1.0108	1.0219		

square root allocation and half of the patients on the control. These are not immediately interpretable, but they do allow us to compare the sample sizes required by the different test statistics or for different allocation proportions. For a fixed value of θ_1, we can calculate the required sample size from d_1^*. So the ratio of the required sample sizes for two different test statistics, or for two different allocations, is the ratio of d_1^{*2} for these two cases.

The bottom portion of the table compares equal allocation and half on the control with the square root allocation. For T_2 just over 4% more observations are required for equal allocation than for the square-root allocation to achieve 80% power, whereas it only needs 1% more subjects when half of the subjects are allocated to the control. For T_∞ about 3% and 2% more subjects are required for equal allocation and for half on the control respectively compared with the square root allocation. About 6% more subjects are required for equal allocation compared with the square root allocation for T_1, but about the same number is required for half on the control and the square root allocation. Although equal allocation requires larger sample sizes for all three test statistics, these results show that if it is considered advantageous for practical or ethical reasons, the increase in the number of patients is not enormous.

For each allocation proportion, we also compare the test statistics T_1 and T_∞ relative to T_2 in the right hand portion of Table 4.6. For equal allocation T_∞ seems to require almost the same sample size as T_2, but T_1 requires 4% more than T_2. When the square root allocation is used, then T_∞ and T_1 require about 1.6% and 2.6%, respectively, more subjects than T_2. With half of the subjects on the control arm, T_∞ needs about 2.7% more subjects than T_2, whereas T_1 requires about the same number.

The configuration we have chosen here is in the middle where T_2 is best and it is not surprising that T_2 requires smaller sample sizes. Similarly we can calculate the ratios for any other configurations of interest, but we do not pursue this further here since it is a straightforward calculation.

Example

We use an example from Horn and Vollandt (1998) who calculate the sample size required to achieve at least 80% power for T_∞ with expected differences of 1 and 0 and variance of observations equal to 5. We standardise these to our scale to get $\theta_1 = 1/\sqrt{5}$ and $\theta_2 = 0$. For equal allocation, by interpolation from the simulation results, we find the required $d_1^* = 2.818$. Plugging this into equation (4.1), we get

$$N \geq \frac{2.818^2 \times 5}{(\sqrt{2}-1)^2} = 231.5.$$

Our required numbers of subjects on each arm are then the smallest integers which satisfy this inequality, which are ($n_0 = n_i = 78$), very close to Horn and Vollandt's ($n_0 = n_i = 77$). The slight difference could be due either to simulation error in our results, or to the fact that Horn and Vollandt's any-pair power does not correspond exactly to the power of T_∞. For the square root allocation, we find that T_∞ requires exactly the same number of n_0 and n_i ($n_0 = 93$ and $n_i = 66$) as Horn and Vollandt.

Using the same prior values we also calculated sample sizes for T_2. For the balanced design $n_0 = n_i = 80$ and, with the square root allocation, $n_0 = 95$ and $n_i = 68$. Thus the total sample sizes required for T_∞ are 234 and 225 for equal and square root allocation respectively, while for T_2 they are 240 and 231 respectively. It can be seen again that the differences between the allocations and between the test statistics are all small. However, this example is for the least favourable configuration, i.e. only one experimental arm is better than the control. We already know that this is the situation which is most favourable to T_∞.

4.4.2 Sample Size to Achieve Specified Expected Loss

The method described in Section 4.4.1 can be used in exactly the same way, but with a target minimum power replaced by a target maximum expected loss. We illustrate with

Table 4.7: Effect sizes yielding 0.5 expected loss and sample size ratio to achieve 0.5 expected loss for different δ for the tests at the 5% level with $d_2^* = d_1^*/2$.

		d_1^*		Ratio of required sample sizes	
δ	T_1	T_2	T_∞	T_1 v. T_2	T_∞ v. T_2
1/3 (equal)	3.0715	2.9721	2.9739	1.0334	1.0006
$\sqrt{2}-1$ (root)	2.9947	2.9185	2.9370	1.0261	1.0063
1/4 (half)	3.0205	2.9727	3.0067	1.0161	1.0114
Ratio of required sample sizes					
equal v. root	1.0184	1.0125	1.0257		
half v. root	1.0186	1.0237	1.0086		

an expected loss of 0.5. The value of 0.5 is chosen so that the sample sizes will be similar to those required for $80-90\%$ power. Table 4.7 corresponds to Table 4.6, but for the expected loss of 0.5. We see that, in terms of expected loss, the allocation is less crucial. For example, for T_2, equal allocation requires an increase in sample size of only about 1.25% compared with the square root allocation. We also see that the differences between the test statistics is very small with, in this case, T_∞ requiring an increase of less than 1% in sample size relative to T_2 at the square root allocation and almost no increase at equal allocation.

The practical consequence of these results is that the decision as to whether to use equal allocation or close to optimal allocation depends on whether one places more importance on the power of the test, or on the quality of the treatment finally selected. In the latter case, there is little to be gained by using anything other than equal allocation and, if this has practical benefits, we should stick to it. The fact that the differences between test statistics is so small might be reassuring, in that whichever we use will not give terrible results, but it does not tell us any more about which one we should use. If $\Delta_2 = \Delta_1/2$ seems like a plausible prior guess, then these results do not give any reason to choose anything other than T_2. The choice of sample size proceeds exactly as for power, but with the value of d_1^* being obtained from the expected loss, rather than the power.

4.5 Conclusions

We have shown that the optimal allocation depends on the unknown Δ_i and on the test statistic to be used. The optimum proportion to be allocated to the control treatment is between 0.4 and 0.5, so that equal allocation is never optimal. However, in most practical cases, including the ATAC, DASH and MORE trials described in Chapters 1 and 2, equal allocation is used. Other considerations can influence patient allocation, such as secondary comparisons between the experimental treatments and information on side effects. Also, the practicality of obtaining informed consent from subjects is easier with equal allocation. Also, if the control is suspected of being the worst treatment, we do not want to maximise the numbers on control. Finally, our results have shown that, although equal allocation is clearly inferior, the difference in the required sample size is not enormous.

Hence, the equal allocation case described in Chapter 3, is still of great practical importance and the conclusions given there are important in practice. However, if appropriate, careful consideration should be given to choosing an appropriate allocation and in this chapter we have shown how this can be done optimally.

Chapter 5

Binary Response Data

In this chapter we consider comparing several experimental treatments with a control using binary outcomes, paying particular attention to three arm trials. Binary responses (including dichotomized continuous variables) are very common in many experimental settings.

In previous chapters we assumed normality, which provides asymptotic approximations for other types of data for large samples. Now we consider the observed proportion of successes in experimental and control arms and use normal approximations to both the raw binomial random variables for each arm and univariate test statistics Z_i based on log odds ratios. Using Monte Carlo simulations for some particular sample sizes and probabilities of success, we approximate the sizes of the tests obtained by using the critical values calculated for normal distributions.

Although it is reasonable to apply the normal approximation to binary data for large samples, it might not be accurate enough for small samples. In this chapter we also derive the likelihood ratio test statistic λ, using the violator algorithm, for binary data. Since, it is not possible to obtain a standardized form of the null distribution in this case, we calculate the exact distribution of λ conditional on the observed total number of successes. This is somewhat in the spirit of Fisher's exact test, which is used for testing the null hypothesis of equality against the two-sided alternative.

In Section 5.1, we define the model and notation to be used in this chapter. Related literature is reviewed in Section 5.2 and the general analysis strategy we recommend is outlined in Section 5.3. Approximate methods suitable for large samples are described and assessed in Section 5.4, for equal allocation and, more briefly, for 40% allocated to the control. An exact conditional test is developed in Section 5.5. Examples are given in

Section 5.6 and some final points are discussed in Section 5.7.

5.1 Model and Notation

Throughout this chapter, we assume that there are n_i subjects on the ith arm, $i = 0, \ldots, I$, whose responses, $Y_{ij},\ j = 1, \ldots, n_i$, are independent with probability of success π_i. As is standard practice, we will refer to the outcomes as "success" ($Y_{ij} = 1$) or "failure" ($Y_{ij} = 0$). Let $Y_i = \sum_{j=1}^{n_i} Y_{ij}$. Then $Y_i \sim Bin(n_i, \pi_i)$. Our null and alternative hypotheses are $H_0 : \pi_i \leq \pi_0\ \forall\ i = 1, ..., I$ and $H_1 : \pi_i > \pi_0$, for at least one i, respectively. As in the previous chapters, we will refer to the null hypothesis of equality as $H_0^* : \pi_0 = \pi_1 = \cdots = \pi_I$ and the individual null hypotheses as $H_{0i} : \pi_i \leq \pi_0$ and $H_{0i}^* : \pi_i = \pi_0$.

5.2 Relevant literature

Several authors who have developed methods for comparing several experimental arms with a control for continuous data have mentioned in passing that they also provide an asymptotic approximation to the case of binary response data. Some of these authors have evaluated the large sample methods for use with binary data. Rather fewer have dealt with the binary case directly. Here we review what relevant literature there is. Many authors compare different methods, often including methods which are similar, but not identical. It is easiest to present the literature in chronological order.

The early paper by Paulson (1952), mentioned in Chapter 2, although mainly concentrating on the normal distribution, gives a brief description of the binomial case. He does an angular transformation and then uses, in effect, T_∞ to test H_0. Modifications of this approach are developed by Dunnett (1984) and Chen and Sarkar (2004).

For comparing several treatments with a control, to test H_0^* against the alternative of a simple order, Williams (1988) considers several test statistics, including the standard chi-squared (X^2) and likelihood ratio (G^2) test statistics for the two-sided alternative and the Cochran-Armitage statistic for testing a pre-specified trend. He develops the isotonic regression estimates for a simple order and suggests using a modified chi-squared test statistic ($\bar{X}^2$), in which the isotonic regression estimators replace the observed proportions. He also obtains the likelihood ratio test (LRT) statistic ($\bar{G}^2$) for the simple ordered alternative. He briefly mentions the simple tree ordered alternative and says $\bar{X}^2$ and $\bar{G}^2$ can be used with these estimates but he does not pursue this. He also looked at $\bar{T}$ which is the LRT statistic for testing equality in an order restricted model with normal data, mentioned briefly in Chapter 2. Finally he considers Dunnett's test.

Williams (1988) uses published approximations of the null distributions of all of these test statistics (except $\bar{G}^2$) and assesses their accuracy by finding the exact null distributions, conditional on $Y = \sum_{i=0}^{I} y_i$ (total number of successes), for $Y = 1, ..., 30$, and plots the achieved sizes against Y. He finds that X^2 is better than G^2, but is somewhat conservative. G^2 is conservative for small Y but liberal otherwise; it's liberality can be reduced, but not removed, by using a Bartlett adjustment. Williams finds that the approximate size of the Cochran-Armitage statistic is surprisingly accurate, $\bar{X}^2$ is no worse than X^2, $\bar{T}$ is not as good as $\bar{X}^2$ and Dunnett's test is even worse. He concludes that we should not rely on continuous approximate distributions, but should instead use the exact conditional distributions, in a manner similar to Fisher's exact test. He goes on to show how this can be done for a general test statistic under the null hypothesis. He notes it takes quite a lot of computing time for $\bar{X}^2$ and $\bar{G}^2$, but recommends this as being worthwhile.

In a paper on more general multiple comparisons with binary responses, Piegorsch (1990) briefly considers multiple comparisons with a control, using the one-sided alternative against H_0^*. He considers the properties of procedures similar to T_∞ when applied to statistics calculated from binary data. The statistics considered are those commonly used in generalized linear models, namely the logit, probit and complementary log-log link functions. The logit link is of most interest to us since it corresponds to using the log odds ratios of estimated proportions as the Z_i in any of the test statistics defined in Chapter 2. The tests considered are Dunnett's test, Simes' procedure, Hommell's modification (which is very similar to Hochberg's procedure) and the likelihood ratio test LRT for H_0^*.

Piegorsch carries out a simulation study and reports that Dunnett's procedure is too conservative for moderate sized samples and the LRT is slightly better, but sometimes liberal. Piegorsch recommended that the asymptotic procedures could be used for total sample sizes above about 100. In a follow-up paper concentrating on confidence intervals, Piegorsch (1991) found that the coverage of Dunnett-type intervals was unacceptable for sample sizes smaller than about 300, but could be improved by applying the so-called Jeffreys-Perks method of Beal (1987). This involves adjusting the standard error of $\hat{\pi}_i - \hat{\pi}_0$ using a Bayesian procedure.

The most comprehensive study comparing methods was carried out by Chuang-Stein and Tong (1995). They consider three methods for comparing several treatments with a control for binary outcomes, assuming the hypotheses $H_{0i} : \pi_i = \pi_0$ are compared against the two-sided alternatives $H_{1i} : \pi_i \neq \pi_0$ for $i \in \{1, \ldots, I\}$. The methods are therefore directly analogous to the two-sided test based on Dunnett's procedure. The three methods considered all make use of the difference between the largest observed proportion of successes and the proportion of successes on the control.

The following methods were considered by Chuang-Stein and Tong. The methods differ only in the rejection (or equivalently acceptance) region used.

1. The asymptotic Freeman-Tukey acceptance region uses a modified angular (arc sin) transformation of $\hat{\pi}_i$ and then uses a normal approximation to find the critical value based on Dunnett's method.

2. The binomial acceptance region uses acceptance boundaries assuming the true π_0 and π_is are equal to the pooled $\hat{\pi}$, i.e. the combined probability of success. They then calculate the full discrete null distribution of the test statistic, i.e. they calculate the test statistic for all possible values of $y_0, y_1, \ldots, y_I$ and the corresponding probabilities of these values under H_0^*, but with the additional assumption that $\pi_i = \hat{\pi}\ \forall i \in \{0, \ldots, I\}$. They then use a test statistic which is like T_∞ except that they use this pooled $\hat{\pi}$ to get the standard error in each comparison. They state that using a fixed standard error gives more stable null probabilities than using a different one for each arm.

3. Dunnett's method is used directly on the binary responses, relying on the normal approximation to the binomial distribution to give reasonable answers for large samples.

Chuang-Stein and Tong (1995) report a simulation study to check the sizes of the tests and find that they are all rather similar. For small sample sizes they can be very conservative whereas for moderate sample sizes they can be a bit liberal and for large sample sizes they give close to 5% type I errors. They conclude that direct use of Dunnett's method is as good as any of the other procedures in most cases. However, if the pooled $\hat{\pi}$ is between 1/3 and 2/3 then they recommend using the Freeman-Tukey transformation procedure.

Agresti and Coull (1996) consider testing H_0^* against the simple ordered alternative, with the additional complication of having other covariates. They obtain the LRT statistic for this case and approximate the conditional null distribution by simulation after fixing the marginal totals. This is similar to work we will present in Section 5.5, but in a different context.

Koch and Hothorn (1999) consider four different versions of the T_∞ test statistic, all using large sample normal approximations, but based on different Z_i's, namely:

1. $Z_i = \hat{\pi}_i - \hat{\pi}_0$ (unscaled);
2. scaled by unpooled variance estimates;
3. scaled by pairwise pooled variance estimates, i.e. π_i and π_0 are pooled.
4. scaled by the total pooled variance estimate, i.e. all are π_j pooled, $j = 0, ..., I$.

They describe how to get the exact conditional distributions for binary data. They note that options 2 and 3 above exhaust the α level better than 1 and 4, i.e. the null distribution has more distinct discrete values so that it is possible to get an actual size closer to the desired α.

Koch and Hothorn then go on to consider the exact unconditional null distributions of test statistics 2 and 3 above, which depend on the unknown π. They take the critical value to be the smallest such that the size of the test is less than or equal to α for all values of π, in a manner similar to Barnard's exact test. They then compare the exact tests for statistics 2 and 3 with statistics 2, 3 and 4 (by simulation) using Dunnett's critical values. The exact distribution for the unpooled statistic gives a test which is very conservative for most π, while the pairwise pooled statistic is slightly conservative for most π, but rather more conservative for large values of π. The tests based on the large sample approximations can be very liberal or very conservative depending on π. The main results presented for illustration are from a very unbalanced design with $n_0 = 40$, $n_1 = 10$ and $n_2 = 10$ and they state that this is worse than for the balanced design.

Peddada et al. (2001) extend the work of Hwang and Peddada (1994), mentioned in Chapter 2, to binary data. Considering the null hypothesis of equality, they explore the simple ordered and simple tree ordered alternatives. They show how to obtain the maximum likelihood estimators under a simple tree order from the pool adjacent violators algorithm (PAVA). They consider a test statistic similar to T_∞, but with the estimator of Hwang and Peddada (1994), i.e. the estimator of each experimental arm is replaced with its isotonic regression estimator from an arbitrary order. They obtain p-values by bootstrapping, but also use Bonferroni corrections assuming asymptotic normality.

It can be seen from this review that most work on binary response data uses tests which are similar to T_∞, which was described in Chapter 2, or simple modifications of it. On the other hand, Williams (1988), Piegorsch (1990) and Agresti and Coull (1996) consider likelihood ratio test statistics for testing the null hypothesis of equality assuming an order restricted model. As explained in Chapter 2, this restriction to the model is not appropriate in the cases we are considering and test statistics based on such an assumption have undesirable properties, such as decreasing as the response from some active arm increases. In this chapter, we develop more appropriate likelihood ratio tests, either based on T_2 with large sample approximations, or developed directly from the binomial probability function.

5.3 A strategy for analysing binary data

Even more than is the case for the normal distribution, most literature on comparing several experimental arms with a control using binary data deals with the null hypothesis of equality. As noted by Chuang-Stein and Tong (1995), compared with the normal distribution, there is rather little research on binary data in this context.

We first consider how to use the methods of earlier chapters with binary data, using large-sample approximations. The simplest approach is to apply the methods of earlier chapters to the total number of successes on each treatment, assuming that the normal approximation to the binomial distribution will give acceptable results. This is an extension of the simple test for comparing two proportions which is frequently presented in introductory textbooks and courses, e.g. Armitage et al. (2002), p.125. This approach has been studied, mainly as applied to Dunnett's test, by for example Williams (1988), Chuang-Stein and Tong (1995) and Koch and Hothorn (1999), as described in Section 5.2.

Another possibility we consider is to use univariate LRT statistics based on the observed log odds ratios and assume that these have approximately normal distributions. This is a direct extension of the asymptotic z-tests which are often used in two-arm trials and in other contexts, e.g. through the use of generalized linear models with a logistic link function - see, for example Armitage et al. (2002), p.127-128. This approach has been studied, again mainly through the use of Dunnett's test, by Williams (1988) and Piegorsch (1990, 1991).

For small samples we will find that neither of these approaches gives acceptable results and instead we develop an exact conditional test in the spirit of Fisher's exact test, for which see, for example, Armitage et al. (2002), p.134-137, and its extension to $(I+1) \times 2$ contingency tables (Mehta and Patel, 1983; Mehta, 1994). Just as the test statistic used in Fisher's test can be derived as the LRT statistic for testing equality against a two-sided alternative for binomial data, so we derive the LRT statistic for our hypotheses.

Then, just as in Fisher's exact test, we condition on the marginal totals, the total number of successes across all arms, and consider the distribution of the test statistic under all possible contingency tables satisfying this condition. Other authors who have studied exact conditional tests in similar contexts include Williams (1988) and Agresti and Coull (1996).

There is some controversy in the literature about whether exact conditional tests or exact unconditional tests should be used. The former condition on the total number of successes across all arms, while the latter either condition on a fixed value of π, such as $\hat{\pi}$, or take a minimax approach to removing the dependence on π, as in Barnard's exact test for the two-sided alternative to the null hypothesis of equality. Barnard's test has been shown to be more powerful for 2 × 2 tables, where Fisher's test is very conservative. However, for more arms the conservativeness of Fisher's test reduces, due to the discrete distribution having more points, while the conservativeness of Barnard's test, which is due to the minimax approach, as well as the discreteness, remains. Ultimately, there is no single correct answer, but most authors have preferred the conditional tests and we also find this to be the most natural approach. For some discussion of this topic, see Suissa and Shuster (1985) and the references contained therein.

5.4 Large sample approximations

We consider two ways of using normal approximations. Since the binomial variance depends on the mean, there are different ways of doing this. First, we use the original proportions, with the normal approximation to the binomial distribution. Then, following the suggestion in Chapter 2, we define the Z_i using the log odds ratios and assume that they are approximately normally distributed. We then use simulations to check the sizes of the asymptotic tests for some particular sample sizes and values of π_i.

5.4.1 Normal approximation to the binomial distribution

We apply the simple normal approximation to the binomial distribution and use the scaled difference of two proportions from independent binomial samples as a univariate test statistic. The normal approximation to the binomial distribution for our model gives

$$Z_i = \frac{\hat{\pi}_i - \hat{\pi}_0}{\sqrt{\hat{\pi}(1-\hat{\pi})(\frac{1}{n_0}+\frac{1}{n_i})}}, \tag{5.1}$$

for $i \in \{1, \ldots, I\}$, where $\hat{\pi}_i = y_i/n_i$ and

$$\hat{\pi} = \frac{\sum_{i=0}^{I} y_i}{\sum_{i=0}^{I} n_i},$$

since

$$Var(\hat{\pi}_i - \hat{\pi}_0) = \frac{\pi_0(1-\pi_0)}{n_0} + \frac{\pi_i(1-\pi_i)}{n_i},$$

which, for the purposes of hypothesis testing, we approximate by

$$\widehat{Var}(\hat{\pi}_i - \hat{\pi}_0) = \hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_0} + \frac{1}{n_i}\right).$$

Note that we are using the total pooled variance, following the recommendation of Koch and Hothorn (1999). Although it might seem more efficient to use

$$\check{Var}(\hat{\pi}_i - \hat{\pi}_0) = \frac{\hat{\pi}_0(1-\hat{\pi}_0)}{n_0} + \frac{\hat{\pi}_i(1-\hat{\pi}_i)}{n_i},$$

Koch and Hothorn (1999) found that this was too conservative for a test statistic corresponding to T_∞. Note that $\widehat{Var}(\hat{\pi}_i - \hat{\pi}_0) \geq \check{Var}(\hat{\pi}_i - \hat{\pi}_0)$ by Jensen's inequality.

It is immediately clear that $Var(Z_i) \simeq 1$ and, assuming equal numbers of subjects on each experimental arm,

$$\begin{aligned}
\rho = Cov(Z_i, Z_j) &= \frac{Cov(\hat{\pi}_i - \hat{\pi}_0, \hat{\pi}_j - \hat{\pi}_0)}{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_0} + \frac{1}{n_1}\right)} \\
&= \frac{Var(\hat{\pi}_0)}{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_0} + \frac{1}{n_1}\right)} \\
&\simeq \frac{n_1}{n_0 + n_1} \\
&= \frac{\delta N}{\delta N + (1 - I\delta)N},
\end{aligned}$$

where, as in earlier chapters, δ is the proportion of subjects on each experimental arm. After simplification, we get

$$\rho \simeq \frac{\delta}{1 - (I-1)\delta}.$$

Hence when $\pi_0 = \pi_1 = \cdots = \pi I$ $\mathbf{Z} = [Z_1, \ldots, Z_I]$ is approximately multivariate normal, with the same distribution as in Chapter 2.

5.4.2 Univariate likelihood ratio test statistics

Here we use the log odds ratios of the estimated proportions of successes in each arm relative to the control. This is also a large sample approximation, but based on the general asymptotic normality of LRT statistics. Let

$$\psi_i = \frac{\pi_i/(1-\pi_i)}{\pi_0/(1-\pi_0)},$$

$i = 1, \ldots, I$, be the odds ratio of success on treatment i and control, where π_i is the probability of success on treatment i, $i = 0, 1, \ldots, I$. For large N,

$$\log \hat{\psi}_i \sim N(\log \psi_i, Var(\log \hat{\psi}_i))$$

(Armitage et al., 2002, p.127), where

$$\hat{\psi}_i = \frac{\hat{\pi}_i/(1-\hat{\pi}_i)}{\hat{\pi}_0/(1-\hat{\pi}_0)},$$

and

$$Var(\log \hat{\psi}_i) = \frac{1}{n_0\pi_0(1-\pi_0)} + \frac{1}{n_i\pi_i(1-\pi_i)}.$$

In practice we use

$$\widehat{Var}(\log \hat{\psi}_i) = \frac{1}{n_0\hat{\pi}_0(1-\hat{\pi}_0)} + \frac{1}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}.$$

We scale the log odds ratios to get Z_i with unit variance, so that

$$\begin{aligned} Z_i &= \frac{\log \hat{\psi}_i}{\sqrt{\widehat{Var}(\log \hat{\psi}_i)}} \\ &= \log\left(\frac{\hat{\pi}_i/(1-\hat{\pi}_i)}{\hat{\pi}_0/(1-\hat{\pi}_0)}\right)\sqrt{\frac{n_0 n_i \hat{\pi}_0\hat{\pi}_i(1-\hat{\pi}_0)(1-\hat{\pi}_i)}{n_i\hat{\pi}_i(1-\hat{\pi}_i)+n_0\hat{\pi}_0(1-\hat{\pi}_0)}}. \qquad (5.2) \end{aligned}$$

From the general properties of LRT statistics, $Z_i \sim N(0,1)$.

To obtain the correlation, we note that, as $n_i \to \infty$, $\sqrt{n_i}(Z_i - \zeta_i) \to_p 0$, where

$$\zeta_i = \log \hat{\psi}_i \,/\, \sqrt{Var(\log \hat{\psi}_i)}.$$

Hence, the asymptotic correlation of Z_i and Z_j is

$$\begin{aligned} \rho = Cov(\zeta_i, \zeta_j) &= Cov\left(\frac{\log \hat{\psi}_i}{se(\log \hat{\psi}_i)}, \frac{\log \hat{\psi}_j}{se(\log \hat{\psi}_j)}\right) \\ &= \frac{Cov(\log \hat{\psi}_i, \log \hat{\psi}_j)}{Var(\log \hat{\psi}_i)}, \end{aligned}$$

since the standard errors are constant for equal numbers of subjects on each experimental arm. Hence,

$$\begin{aligned} \rho &= \frac{n_0 n_1 \pi(1-\pi)}{n_0+n_1}\left[Cov\left\{\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right), \log\left(\frac{\hat{\pi}_j}{1-\hat{\pi}_j}\right)\right\}\right. \\ &\quad -Cov\left\{\log\left(\frac{\hat{\pi}_0}{1-\hat{\pi}_0}\right), \log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right)\right\} - Cov\left\{\log\left(\frac{\hat{\pi}_0}{1-\hat{\pi}_0}\right), \log\left(\frac{\hat{\pi}_j}{1-\hat{\pi}_j}\right)\right\} \\ &\quad \left.+Cov\left\{\log\left(\frac{\hat{\pi}_0}{1-\hat{\pi}_0}\right), \log\left(\frac{\hat{\pi}_0}{1-\hat{\pi}_0}\right)\right\}\right] \\ &= \frac{n_0 n_1 \pi(1-\pi)}{n_0+n_1} Var\left\{\log\left(\frac{\hat{\pi}_0}{1-\hat{\pi}_0}\right)\right\} \\ &= \frac{n_0 n_1 \pi(1-\pi)}{n_0+n_1}\frac{1}{n_0\pi(1-\pi)} \\ &= \frac{\delta}{1-(I-1)\delta}, \end{aligned}$$

as before. Hence, the asymptotic distribution of $\mathbf{Z}$ is the same as in Chapter 2 and Section 5.4.1.

5.4.3 Simulations

Simulations of the null distribution were performed for the case $I = 2$. We generated 2,000,000 trials for three independent binomial samples for a given probability of success, assumed to be the same in each arm, then for each simulation run calculated Z_1 and Z_2, using each of the methods in Subsections 5.4.1 and 5.4.2. Then the values of T_2 and T_∞ were calculated for each approximation. Note that the simulations allow the extreme cases $y_i = 0$ and $y_i = 1$. In these cases the likelihoods were calculated directly to avoid computational problems.

We present results for $\pi = 0.5$ and $\pi = 0.1$ to represent the case with smallest variance and a case with reasonably large variance. Note that, since we are doing one-sided tests, the results for $\pi = 0.1$ cannot be assumed to apply for $\pi = 0.9$. We present results with small, moderate and fairly large sample sizes. In the case of equal allocation, we use the best available critical values from Chapter 3, shown in Table 3.2, to check if each simulation would reject H_0 or not. This allows us to estimate the true size of each test.

5.4.4 Results for equal allocation

We first present the results for the simple approximation on the original scale. The estimated sizes of the tests in the case of equal allocation are shown in Table 5.1 and plotted in Figures 5.1 and 5.2. Generally, with moderate to large sample sizes, T_2 seems liberal when testing at 5% or 2.5%, but slightly conservative when testing at 1%. For a total sample size of 600 or greater the sizes are very close to the nominal value. With a small sample size of 30, the picture is much more confused, the test being severely conservative or slightly liberal in different cases.

Note that the highly discrete nature of the true null distribution for such small sample sizes mean that a slight change in the critical value could lead to a large change in the size of the test. We could improve these results by using a randomized test, in which we replace the large sample critical value, which is impossible to achieve with discrete data, with the possible value immediately below it. If this value is obtained from our data, we randomly reject or do not reject H_0 in order to achieve a true size of exactly α. Such tests are rarely used in practice and we will not pursue them here.

Figure 5.1: Sizes of tests for T_2 using normal approximation to binomial. The solid line represents $\pi = 0.5$ and the dashed line represents $\pi = 0.1$. Black, red and blue represent 5%, 2.5% and 1% significance levels respectively.

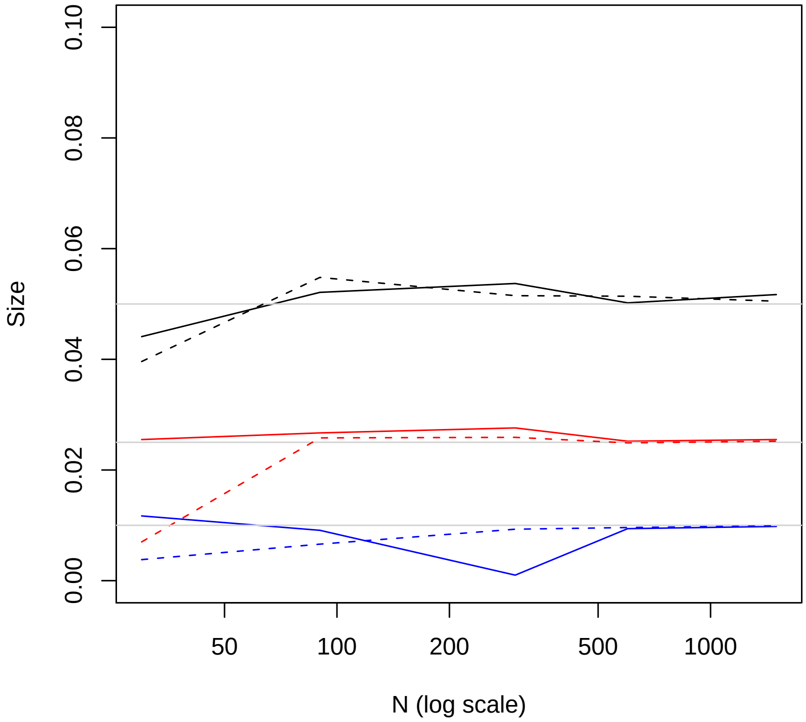

Figure 5.2: Sizes of tests for T_∞ using normal approximation to binomial. The solid line represents $\pi = 0.5$ and the dashed line represents $\pi = 0.1$. Black, red and blue represent 5%, 2.5% and 1% significance levels respectively.

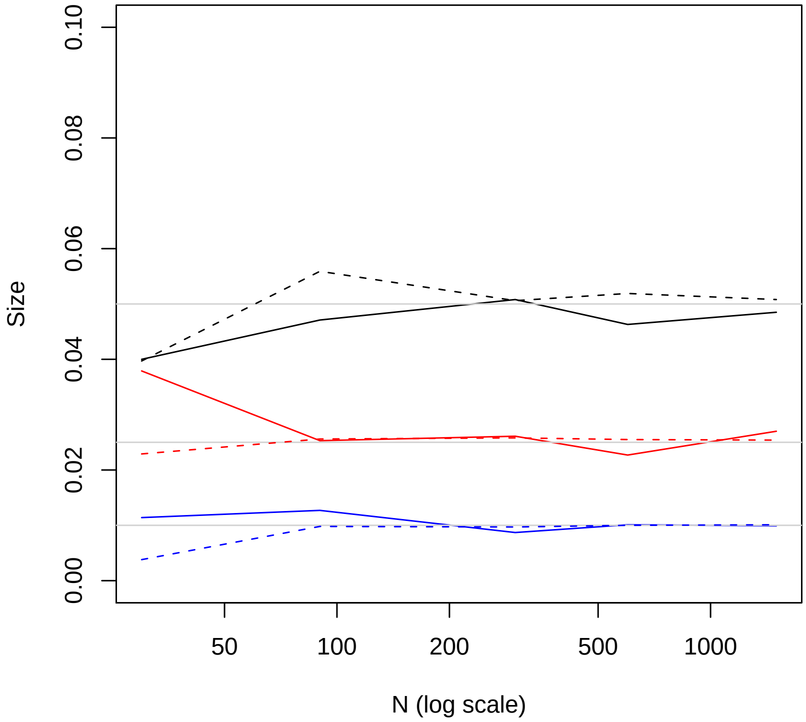

Table 5.1: Test sizes (%) using mean critical values for large sample approximation on the original scale with $\delta = 1/3$ for T_2 and T_∞

		Estimated size of the test					
N	π	T_2			T_∞		
		5%	2.5%	1%	5%	2.5%	1%
30	.5	4.41	2.55	1.17	4.00	3.79	1.14
	.1	3.96	0.70	0.38	3.97	2.29	0.38
90	.5	5.21	2.67	0.91	4.71	2.53	1.27
	.1	5.48	2.58	0.66	5.59	2.56	0.98
300	.5	5.37	2.76	0.10	5.08	2.61	0.87
	.1	5.15	2.59	0.93	5.06	2.58	0.97
600	.5	5.02	2.51	0.94	4.63	2.27	1.01
	.1	5.14	2.49	0.96	5.19	2.55	1.00
1500	.5	5.17	2.55	0.98	4.85	2.70	0.99
	.1	5.05	2.52	0.99	5.08	2.54	1.01

The results for T_∞ are rather different, showing more variation between π of 0.5 and 0.1. For $\pi = 0.1$, with moderate sample sizes it is generally liberal, whereas for $\pi = 0.5$ it can be conservative. For a sample size of 30, the results are similar to those for T_2. For T_∞, the sizes are reasonably close to the nominal values for sample sizes of 300 or greater. These results broadly agree with those presented by Chuang-Stein and Tong (1995) for a two-sided alternative hypothesis.

The corresponding results for the large sample approximation based on log-odds ratios with equal allocation are given in Table 5.2 and plotted in Figures 5.3 and 5.4. We find that the sizes of the tests are not close to the nominal values unless the total sample size is at least 600 and 1500 is better. For moderate sample sizes, T_2 is generally slightly liberal when $\pi = 0.5$, except at the 1% level and somewhat conservative when $\pi = 0.1$. For small sample sizes, it can be quite extreme in either direction, the results for $\pi = 0.1$ being particularly far from the nominal sizes. Note that for $N = 30$ and $\pi = 0.1$, H_0 is rejected if there are no successes on the control but at least one success on either of the other arms. The probability of this is

$$P\left(\{\hat{\pi}_0 = 0\} \bigcap \{\hat{\pi}_1 + \hat{\pi}_2 > 0\}\right) = 0.9^{10}\left(1 - 0.9^{20}\right) = 0.3063,$$

which is very close to the tabulated size. This suggests that having no successes on the

Table 5.2: Estimated test sizes (%) for the large sample approximation using log odds ratios with $\delta = 1/3$ for T_2 and T_∞

		Size of the tests					
N	π	T_2			T_∞		
		5%	2.5%	1%	5%	2.5%	1%
30	.5	4.33	1.85	0.61	3.89	1.29	0.49
	.1	30.65	30.64	30.64	30.65	30.64	30.64
90	.5	4.95	2.65	0.80	4.72	2.54	0.71
	.1	5.13	4.34	4.24	5.32	4.32	4.24
300	.5	5.32	2.42	0.98	5.03	2.60	0.84
	.1	4.22	1.87	0.54	4.10	1.83	0.52
600	.5	5.03	2.50	0.94	4.64	2.27	1.02
	.1	4.71	2.20	0.80	4.60	2.15	0.76
1500	.5	5.18	2.53	0.96	4.86	2.52	0.98
	.1	4.88	2.41	0.93	4.90	2.35	0.90

control accounts for almost all the probability of rejecting H_0. The results for T_∞ are broadly similar.

The inaccuracy of the sizes of the tests achieved might alone be enough to make us hesitant about using these asymptotic methods in small samples. However, in Section 5.6 we will apply them to some examples, for one of which we carry out a small power study.

5.4.5 Results for unequal allocation

In Table 5.3, we show some results for a different treatment allocation, namely having 40% of subjects on the control, which we found in Chapter 4 was close to optimal. In this case, since we do not have mean cutpoints for the unequal allocation, the critical values are estimated from two million simulations. The results for equal allocation are also shown, which are almost identical to those in Table 5.1, confirming that the critical values are adequately estimated. The pattern of results is broadly similar for the different allocations, but we can see that for unequal allocation some of the sizes are considerably further from their nominal values than for equal allocation. The results for T_∞ at 5% can be compared with those of Koch and Hothorn (1999), although they used an even more unequal allocation and different sample sizes. Our results are broadly similar to theirs, as

Figure 5.3: Sizes of tests for T_2 using normal approximation to log odds ratio. The solid line represents $\pi = 0.5$ and the dashed line represents $\pi = 0.1$. Black, red and blue represent 5%, 2.5% and 1% significance levels respectively.

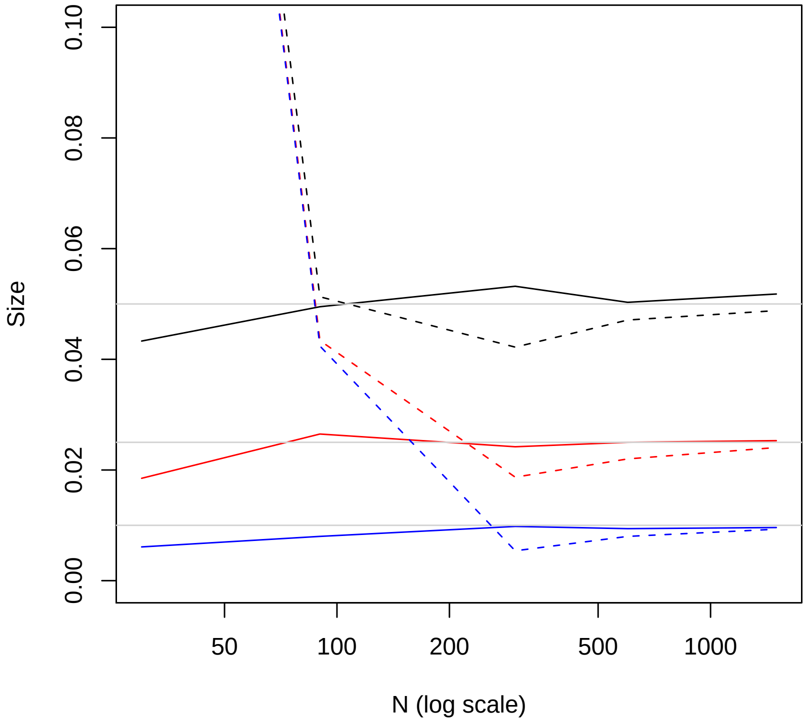

Figure 5.4: Sizes of tests for T_∞ using normal approximation to log odds ratio. The solid line represents $\pi = 0.5$ and the dashed line represents $\pi = 0.1$. Black, red and blue represent 5%, 2.5% and 1% significance levels respectively.

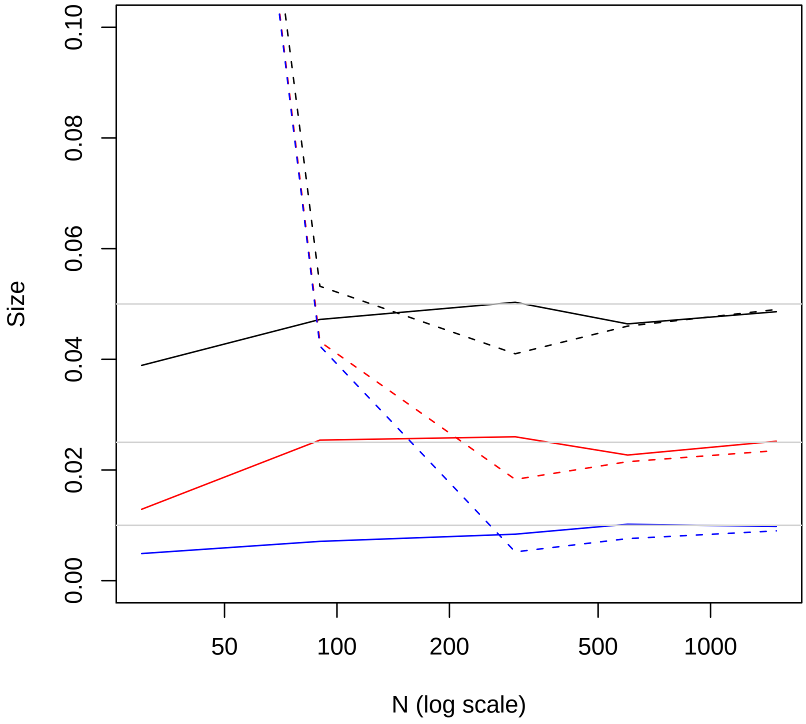

Table 5.3: Estimated test sizes (%) for T_2 and T_∞ for large sample approximation on original scale with different treatment allocations.

δ	N	π	Size of the tests					
			T_2			T_∞		
			5%	2.5%	1%	5%	2.5%	1%
1/3	30	.5	4.41	2.55	1.17	4.00	3.79	1.14
		.1	3.96	0.70	0.38	3.97	2.29	0.38
1/3	90	.5	5.19	2.66	0.91	4.68	2.52	1.27
		.1	5.47	2.56	0.64	5.59	2.55	0.97
3/10	30	.5	5.74	2.63	0.89	5.35	2.69	0.88
		.1	7.22	1.52	0.41	7.26	2.50	0.41
3/10	90	.5	4.96	2.50	1.02	5.03	2.51	0.96
		.1	5.78	3.03	1.00	5.64	2.82	1.00

they also found that with small π the test was liberal, although their results for $\pi = 0.5$ are slightly on the other side of the nominal value from ours. Overall, tests based directly on the normal approximation to the binomial distribution cannot be recommended for fewer than 30 subjects on each arm.

For the approximation based on log-odds ratios with unequal allocation, the results are given in Table 5.4. They show that in this case also, the sizes achieved are quite far from the nominal sizes for small sample sizes. If π is close to 0.5 the test is reasonably accurate with 30 patients in each arm, but otherwise requires at least 100 in each arm.

5.5 An exact likelihood ratio test

For large samples, we can use the normal approximations. Perhaps a reasonable requirement would be that $\min(n_i\pi_0, n_i(1-\pi_0))$ is at least 30, 45 or 60 for testing at 5%, 2.5% and 1% respectively. However, in the previous section we saw that for smaller samples the approximation is unreliable and for very small samples, such as 10 in each arm, it is essentially useless. Instead, we develop a test based directly on the binomial distribution. We derive the likelihood ratio test statistic for the simple tree order for binomial data. The null distribution of this is not known and, unlike in the normal case, a standardized form of the binomial distribution cannot be obtained. Instead we develop a method to

Table 5.4: Test sizes (%) for the large sample approximation using log odds ratios with different allocations for T_1 and T_∞.

δ	N	π	Size of the tests					
			T_2			T_∞		
			5%	2.5%	1%	5%	2.5%	1%
1/3	30	.5	4.33	1.85	0.61	3.89	1.29	0.49
		.1	30.65	30.64	30.64	30.65	30.64	30.64
1/3	90	.5	4.95	2.65	0.80	4.72	2.54	0.71
		.1	5.13	4.34	4.24	5.32	4.32	4.24
3/10	30	.5	3.94	1.74	0.73	4.95	2.14	0.74
		.1	24.09	24.02	24.01	24.08	24.02	24.01
3/10	90	.5	4.80	2.36	0.90	4.96	2.46	0.92
		.1	4.20	2.66	2.26	3.79	2.68	2.30

calculate the exact conditional p-value for a given marginal total number of successes. Because of the lack of a standardised form, the general properties of this cannot easily be studied, but we give some examples to show how they can be considered in specific cases.

5.5.1 Likelihood ratio test statistic

As in Chapter 2, we derive the likelihood ratio test statistic in the usual way by obtaining the restricted and unrestricted maximum likelihood estimators (MLEs) and plugging them into the log-likelihood. Let $Y_i \sim Bin(n_i, \pi_i)$, where $i = 0, ..., I$ and assume that Y_i $(i = 0, ..., I)$ are independent. Then the unrestricted MLEs are, as usual, $\hat{\pi}_i = y_i/n_i$, $i = 0, 1, \ldots, I$.

The restricted MLEs are those from a model using the simple tree order $\pi_i \leq \pi_0$, $i = 1, \ldots, I$. Barlow et al. (1972), p.93, showed that the MLEs for a simple tree order restriction for any one-parameter exponential family, which includes the binomial, distribution are given by the isotonic regression estimators. Robertson et al. (1988), p.8-11, showed that the pool adjacent violators algorithm (PAVA) gives the isotonic regression estimators - see also Barlow et al. (1972), p.102. We now apply this result to obtain our restricted MLEs. Note that these are essentially the same as the estimators mentioned by Williams (1988), although he suggested using them to test different hypotheses and did not give the form of the LRT statistic. Silvapulle and Sen (2005) also give several tests which use these and similar estimators, although they also did not derive the form of the LRT statistic.

The joint likelihood is given by

$$L\left(\pi_0, \ldots, \pi_I \mid \mathbf{y}\right) = \prod_{i=0}^{I} \binom{n_i}{y_i} \pi_i^{y_i} \left(1-\pi_i\right)^{n_i - y_i}.$$

Assuming that none of the π_is are 0 or 1, the log-likelihood is

$$\log L\left(\pi_0, \ldots, \pi_I \mid \mathbf{y}\right) = \sum_{i=0}^{I} \left\{ \log \binom{n_i}{y_i} + y_i \log \pi_i + \left(n_i - y_i\right) \log\left(1-\pi_i\right) \right\}.$$

However, in what follows, we will need to consider the boundary cases, so we define

$$\log_+ \frac{A}{B} = \begin{cases} 1 & \text{if } A = 0 \text{ or } B = 0; \\ \log \frac{A}{B} & \text{otherwise.} \end{cases}$$

and we write

$$\log L\left(\pi_0, \ldots, \pi_I \mid \mathbf{y}\right) = \sum_{i=0}^{I} \left[\log \binom{n_i}{y_i} + y_i \log_+\left(\pi_i\right) + \left(n_i - y_i\right) \log_+\left(1-\pi_i\right) \right]. \tag{5.3}$$

Let $\tilde{\boldsymbol{\pi}}$ denote the restricted MLEs under H_0. We calculate $\tilde{\boldsymbol{\pi}}$ using the PAVA algorithm. Assume that the unrestricted MLEs of the experimental arms are in the order $\hat{\pi}_1 \geq \hat{\pi}_2 \geq \cdots \geq \hat{\pi}_I$, where as usual other orders follow by symmetry. We must consider $I+1$ different possibilities which we describe as the following cases.

Case 0: If $\hat{\pi}_1 \leq \hat{\pi}_0$, then $\tilde{\pi}_i = \hat{\pi}_i$ $i = 0, \ldots, I$.

Case 1: If $\hat{\pi}_1 > \hat{\pi}_0$, pool with the control to get $\tilde{\pi}_0^{(1)} = \frac{y_0+y_1}{n_0+n_1}$. If $\hat{\pi}_2 \leq \tilde{\pi}_0^{(1)}$, then $\tilde{\pi}_0 = \tilde{\pi}_1 = \tilde{\pi}_0^{(1)}$ and $\tilde{\pi}_i = \hat{\pi}_i$ for $i = 2, \ldots, I$.

Case 2: If $\hat{\pi}_2 > \tilde{\pi}_0^{(1)}$, then let $\tilde{\pi}_0^{(2)} = \frac{y_0+y_1+y_2}{n_0+n_1+n_2}$. If $\hat{\pi}_3 \leq \tilde{\pi}_0^{(2)}$, then $\tilde{\pi}_0 = \tilde{\pi}_1 = \tilde{\pi}_2 = \tilde{\pi}_0^{(2)}$ and $\tilde{\pi}_i = \hat{\pi}_i$ for $i = 3, \ldots, I$.

⋮

Case J: If $\hat{\pi}_J > \tilde{\pi}_0^{(J-1)}$ then pool to get $\tilde{\pi}_0^{(J)} = \frac{y_0+y_1+\cdots+y_J}{n_0+n_1+\cdots+n_J} = \frac{\sum_{j=0}^{J} y_j}{\sum_{j=0}^{J} n_j}$. If $\hat{\pi}_{J+1} \leq \tilde{\pi}_0^{(J)}$, then $\tilde{\pi}_0 = \tilde{\pi}_1 = \cdots = \tilde{\pi}_J = \tilde{\pi}_0^{(J)}$ and $\tilde{\pi}_i = \hat{\pi}_i$, for $i = J+1, \ldots, I$.

⋮

Case I: If $\hat{\pi}_I > \tilde{\pi}_0^{(I-1)}$ then pool to get $\tilde{\pi}_0 = \tilde{\pi}_1 = \cdots = \tilde{\pi}_I = \frac{y_0+y_1+\cdots+y_I}{n_0+n_1+\cdots+n_I} = \frac{\sum_{j=0}^{I} y_j}{\sum_{j=0}^{I} n_j}$.

We will briefly consider the form of $\tilde{\boldsymbol{\pi}}$ for boundary values of $\hat{\boldsymbol{\pi}}$:

- $\hat{\pi}_0 = 0$ but $\hat{\pi}_1 > 0 \Rightarrow$ Case 0 is not true. Hence, the control will be pooled with other groups and $\tilde{\pi}_0 \neq 0$.

- $\hat{\pi}_0 = \hat{\pi}_1 = \cdots = \hat{\pi}_I = 0 \Rightarrow \tilde{\boldsymbol{\pi}} = \hat{\boldsymbol{\pi}} = \mathbf{0}$ and so the log-likelihood ratio is 0.
- $\hat{\pi}_0 = 1 \Rightarrow \hat{\pi}_1 \leq \hat{\pi}_0$ and we have Case 0, $\tilde{\boldsymbol{\pi}} = \hat{\boldsymbol{\pi}}$.

Now assume $0 < \hat{\pi}_0 < 1$.

- $\hat{\pi}_1 = 0 \Rightarrow \hat{\pi}_2 = \cdots = \hat{\pi}_I = 0$. Then we have Case 0, $\tilde{\boldsymbol{\pi}} = \hat{\boldsymbol{\pi}}$.
- $\hat{\pi}_1 = 1 \Rightarrow$ we do not have Case 0, so we will pool treatment 1 (and any others with $\hat{\pi}_j = 1$), so that $0 < \tilde{\pi}_i < 1 \ \forall i = 0, 1, \ldots, I$.
- $\hat{\pi}_I = 0 \Rightarrow$ we do not have Case I, so we will not pool treatment $I \Rightarrow \tilde{\pi}_I = \hat{\pi}_I = 0$.
- $\hat{\pi}_I = 1 \Rightarrow \hat{\pi}_1 = \cdots = \hat{\pi}_{I-1} = 1 \Rightarrow$ everything is pooled, so that $\tilde{\boldsymbol{\pi}} = \hat{\boldsymbol{\pi}}$.

Thus, $\tilde{\pi}_i$ can be equal to zero or one only if $\tilde{\pi}_i = \hat{\pi}_i$. Hence, we will see that the following results cover the boundary cases as well as the more general cases.

From (5.3), twice log-likelihood ratio is given by

$$\lambda = 2\sum_{i=0}^{I}\left\{y_i \log_+\left(\frac{\hat{\pi}_i}{\tilde{\pi}_i}\right) + (n_i - y_i)\log_+\left(\frac{1-\hat{\pi}_i}{1-\tilde{\pi}_i}\right)\right\}. \tag{5.4}$$

For Case J, i.e. $\hat{\pi}_J > \tilde{\pi}_0^{(J-1)}$, we substitute the values of $\tilde{\pi}_j$ and $\hat{\pi}_j$ to get

$$\begin{aligned}\lambda &= 2\sum_{i=0}^{J}\left[y_i\left\{\log_+\left(\frac{y_i}{n_i}\right) - \log_+\left(\frac{\sum_{j=0}^{J} y_j}{\sum_{j=0}^{J} n_j}\right)\right\}\right.\\ &\quad\left.+(n_i - y_i)\left\{\log_+\left(1 - \frac{y_i}{n_i}\right) - \log_+\left(1 - \frac{\sum_{j=0}^{J} y_j}{\sum_{j=0}^{J} n_j}\right)\right\}\right].\end{aligned}$$

Hence,

$$\begin{aligned}\lambda &= 2\sum_{i=0}^{J}\left[y_i \log_+\left(\frac{y_i}{n_i}\frac{\sum_{j=0}^{J} n_j}{\sum_{j=0}^{J} y_j}\right) + (n_i - y_i)\log_+\left\{\frac{n_i - y_i}{n_i}\frac{\sum_{j=0}^{J} n_j}{\sum_{j=0}^{J}(n_j - y_j)}\right\}\right]\\ &= 2\sum_{i=0}^{J}\left[y_i \log_+ y_i - y_i \log n_i + y_i \log\left(\sum_{j=0}^{J} n_j\right) - y_i \log_+\left(\sum_{j=0}^{J} y_j\right)\right.\\ &\quad + (n_i - y_i)\log_+(n_i - y_i) - (n_i - y_i)\log n_i + (n_i - y_i)\log\left(\sum_{j=0}^{J} n_j\right)\\ &\quad\left. - (n_i - y_i)\log_+\left\{\sum_{j=0}^{J}(n_j - y_j)\right\}\right].\end{aligned}$$

After simplifying,

$$\begin{aligned}\lambda &= 2\sum_{i=0}^{J}\left[y_i \log_+ y_i - y_i \log_+\left(\sum_{j=0}^{J} y_j\right) + (n_i - y_i)\log_+(n_i - y_i) - n_i \log n_i\right.\\ &\quad\left.+n_i \log\left(\sum_{j=0}^{J} n_j\right) - (n_i - y_i)\log_+\left\{\sum_{j=0}^{J}(n_j - y_j)\right\}\right].\end{aligned}$$

Separating the terms in the summation,

$$\begin{aligned}\lambda &= 2\left[\sum_{i=0}^{J} y_i \log_+ y_i - \sum_{i=0}^{J} y_i \log_+ \left(\sum_{j=0}^{J} y_j\right) + \sum_{i=0}^{J} (n_i - y_i) \log_+ (n_i - y_i) - \sum_{i=0}^{J} n_i \log n_i \right.\\ &\left. + \sum_{i=0}^{J} n_i \log \left(\sum_{j=0}^{J} n_j\right) - \sum_{i=0}^{J} (n_i - y_i) \log_+ \left\{\sum_{j=0}^{J} (n_j - y_j)\right\}\right].\end{aligned}$$

Finally, taking common factors and rearranging, we obtain the general form of the LRT statistic as

$$\begin{aligned}\lambda &= 2\left[\sum_{i=0}^{J} y_i \log_+ \left(\frac{y_i}{\sum_{j=0}^{J} y_j}\right) + \sum_{i=0}^{J} (n_i - y_i) \log_+ \left\{\frac{n_i - y_i}{\sum_{j=0}^{J} (n_j - y_j)}\right\}\right.\\ &\left. - \sum_{i=0}^{J} n_i \log \left(\frac{n_i}{\sum_{j=0}^{J} n_j}\right)\right]. \qquad (5.5)\end{aligned}$$

As in Chapter 3, λ does not meet the conditions for the null distribution to have an asymptotically χ^2 distribution, even ignoring the boundary values, due to the restricted MLEs possibly being on the boundary of the parameter space. Its distribution is asymptotically a mixture of χ^2 distributions with unknown mixing probabilities (Barlow et al., 1972; Robertson et al., 1988; Silvapulle and Sen, 2005). As in Chapter 3, we prefer to try to find the exact distribution by numerical methods.

5.5.2 Conditional null distribution of the likelihood ratio test statistic

The null distribution of the LRT statistic λ depends on the unknown values of the π_is and so there is no way it can be obtained in general. Instead, we will follow the usual practice in such circumstances and condition on $\sum_{i=0}^{I} Y_i = \sum_{i=0}^{I} y_i$. We also condition on the sample sizes $n_0, n_1, \ldots, n_I$, i.e. we will fix the marginal totals in the contingency table of the observed data. There are a finite number of tables of dimension $(I + 1) \times 2$ with these fixed marginal totals. We will calculate the exact p-value assuming $\pi_0 = \pi_1 = \cdots \pi_I$, conditional on these margins. The probability of a type I error is maximised when all π_is are equal.

This general structure is illustrated for three arms in Table 5.5. The probability of observed cell frequencies being $r_0, r_1, \ldots, r_I$, conditional on the observed marginal totals $\sum_{i=0}^{I} Y_i = \sum_{i=0}^{I} y_i = Y$ and $n_0, n_1, \ldots, n_I$, is given by

$$\begin{aligned}P(Y_0 = r_0, Y_1 = r_1, \ldots, Y_I = r_I) &= P(Y_0 = r_0) P(Y_1 = r_1 \mid Y_0 = r_0) \cdots \\ &\quad P(Y_I = r_I \mid Y_0 = r_0, Y_1 = r_1, \ldots, Y_{I-1} = r_{I-1}),\end{aligned}$$

Table 5.5: Structure of frequency table for three arms

	Success	Failure	Total
Control	r_0	$n_0 - r_0$	n_0
Treatment 1	r_1	$n_1 - r_1$	n_1
Treatment 2	r_2	$n_2 - r_2$	n_2
Total	Y	$N - Y$	N

where

$$\begin{aligned} r_0 &\in \{\max(0, Y - N + n_0), ..., \min(n_0, Y)\} \\ r_1 &\in \{\max(0, Y - N + n_0 + n_1), ..., \min(n_1, Y - r_0)\} \\ &\vdots \\ r_I &= Y - \sum_{i=0}^{I-1} r_i. \end{aligned}$$

Then it is clear that, if $\pi_0 = \pi_1 = \cdots = \pi_I$,

$$\begin{aligned} P(Y_0 = r_0, Y_1 = r_1, \ldots, Y_I = r_I) &= \frac{\binom{Y}{r_0}\binom{N-Y}{n_0-r_0}}{\binom{N}{n_0}} \times \frac{\binom{Y-r_0}{r_1}\binom{N-Y-n_0+r_0}{n_1-r_1}}{\binom{N-n_0}{n_1}} \times \cdots \times 1 \\ &= \frac{\binom{n_0}{r_0}\binom{n_1}{r_1}\cdots\binom{n_I}{r_I}}{\binom{N}{Y}}. \end{aligned} \tag{5.6}$$

These probabilities, which form what is sometimes known as the multivariate hypergeometric distribution, are the same as those used by Mehta and Patel (1983), Williams (1988), Silvapulle and Sen (2005) and other authors who calculate exact distributions for other test statistics. As noted by Koch and Hothorn (1999), these probabilities remain the same for any test statistic, but generate different distributions for different test statistics.

These probabilities and the corresponding values of λ, can be evaluated for each possible set of values of $r_0, r_1, \ldots, r_I$. The p-value is obtained exactly as the sum of probabilities of values of λ which are greater than or equal to the observed value. Note that if any $\pi_i < \pi_0$, the probabilities will change, but those for large values of λ will be smaller, so that the p-value reported is the *maximum* probability, under H_0, of observing a value of λ at least as large as that observed. Hence, it is a matter of convention whether or not we should refer to this test as *exact*, but we will continue to do so.

Because this is an exact test, there is no need to calculate the size, which can be made exactly α by the use of a randomised test. Although this gives the exact test some advantage over the large-sample tests, whose size is not exactly α, it would also be interesting

Table 5.6: A frequency table for three arms

	Survival	Death	Total
Control	5	25	30
Treatment 1	16	14	30
Treatment 2	18	12	30
Total	39	51	90

to compare the powers of the different methods. This, however, is not as simple as for the normal case studied in Chapter 3. The power depends not just on the true differences between treatments, but on the specific values of $\pi_0, \pi_1, \ldots, \pi_I$ and on the sample sizes $n_0, n_1, \ldots, n_I$, so it is not possible to be comprehensive. Further, the exact conditional test is computationally intensive, so applying it to each of a large number of simulated data sets takes up very large amounts of computer time. It is noticeable that most of the literature on exact tests, e.g. Williams (1988), Piegorsch (1990), Chuang-Stein and Tong (1995) and Koch and Hothorn (1999), focus on comparisons of size and not power. Instead of attempting to study the power in detail, in the next section we present a few examples, with a small power comparison for one of them.

5.6 Examples

5.6.1 Example 1

We use a simple fictitious example to demonstrate the calculation and use of the null distribution of λ. Let there be 39 successes in total in a three arm trial, including a control arm, with 30 subjects in each arm. The observed proportions of successes in the control and treatment arms were 5/30, 16/30 and 18/30 respectively, as shown in Table 5.6. If we test the simple null hypothesis $H_0^* : \pi_0 = \pi_1 = \pi_2$ against the two-sided alternative, the χ^2 test gives a p-value of 0.00129, whereas Fisher's exact test gives 0.00105.

We apply our exact one-sided test as follows. After ordering, $\hat{\pi}_0 = 5/30$, $\hat{\pi}_1 = 18/30$ and $\hat{\pi}_2 = 16/30$. We get the restricted MLEs by pooling as follows. $\hat{\pi}_1 > \hat{\pi}_0$, so pool to get $\tilde{\pi}_0^{(1)} = 23/60$. $\hat{\pi}_2 > \tilde{\pi}_0^{(1)}$, so pool to get $\tilde{\pi}_0 = \tilde{\pi}_1 = \tilde{\pi}_2 = 13/30$. Here there are two violations, so that $J = 2$. Then, from equation (5.5), calculate the observed value of λ for

Table 5.7: Frequency tables conditional on marginal totals with 39 successes.

0	30	30	0	30	30	$\cdots$	30	0	30
9	21	30	10	20	30	$\cdots$	9	21	30
30	0	30	29	1	30	$\cdots$	0	30	30
39	51	90	39	51	90	$\cdots$	39	51	90

this data set as

$$\begin{aligned}
\lambda &= 2\left(5\log\frac{5}{39}+16\log\frac{16}{39}+18\log\frac{18}{39}+25\log\frac{25}{51}+14\log\frac{14}{51}\right.\\
&\quad\left.+\,12\log\frac{12}{51}-30\log\frac{30}{90}-30\log\frac{30}{90}-30\log\frac{30}{90}\right)\\
&= 14.2919.
\end{aligned}$$

We now consider all possible allocations of 39 successes to the 3 treatments, as illustrated in Table 5.7. The total number of tables can be calculated as follows. Clearly, the number of successes on the control arm can be anything from 0 to 30. If there are no successes on the control, the number of successes on treatment 1 must be at least 9 (since there are 39 in total) and can be no more than 30; if there is one success on the control, the number on treatment 1 can be from 8 to 30; $\cdots$; if there are nine successes on the control, the number on treatment 1 can be from 0 to 30; if there are 10 successes on the control, the number on treatment 1 can be from 0 to 29; $\cdots$; if there are 30 successes on the control, the number on treatment 1 can be from 0 to 9. Since the number of successes on treatment 2 is determined by the numbers on treatments 0 and 1, we find the number of possible contingency tables to be $(22+23+\cdots+31)+(30+\cdots+10)=685$.

The probabilities corresponding to each value of λ are calculated from equation (5.6). From these we can construct the exact conditional probability mass function of λ when $\pi_0=\pi_1=\cdots=\pi_I$ and a histogram of this is shown in Figure 5.5. The probability that $\lambda=0$ is 0.3906 and the largest possible value of λ is 73.8. We can see that the distribution is highly skewed to the right. By adding the probabilities for contingency tables which give $\lambda\geq 14.2919$, we find the p-value is 0.000265.

We also use this example to show how to calculate T_2 and, from the simulations of the null distribution done in Chapter 3, we find out the p-values from the large sample approximations to compare them with those from the exact test for the binomial distribution. Of

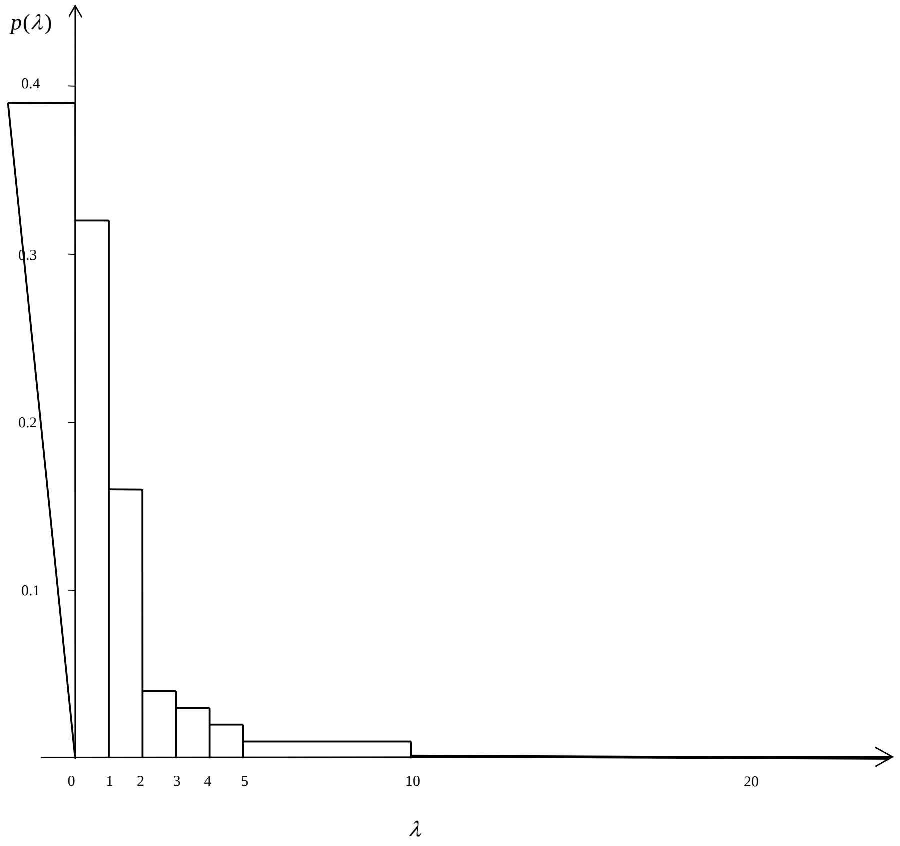

Figure 5.5: Null distribution of λ

course, similar calculations could be done for any other test statistic reported in Chapter 3.

We had $\hat{\pi}_0 = 5/30$, $\hat{\pi}_1 = 16/30$ and $\hat{\pi}_2 = 18/30$. Then $\hat{\pi} = \frac{5+16+18}{30+30+30} = \frac{39}{90}$. For the simple normal approximation to the binomial distribution, from (5.1) we calculate

$$Z_1 = \frac{16/30 - 5/30}{\sqrt{39/90(1-39/90)(\frac{1}{30}+\frac{1}{30})}} = 2.8658$$

and

$$Z_2 = \frac{18/30 - 5/30}{\sqrt{39/90(1-39/90)(\frac{1}{30}+\frac{1}{30})}} = 3.3868.$$

Then

$$T_2 = \sqrt{3.3868^{+2} + \frac{(2.8658 - 0.5 \times 3.3868)^{+2}}{1-0.5^2}} = 3.6474.$$

From the normal distribution simulation we use

$$\frac{\text{no. of simulations with } (T_2 > 3.6474)}{\text{total no. of simulations}}$$

to get the p-value, which turns out to be 0.000352. This is somewhat different from the p-value from the exact test, although the qualitative conclusions do not change.

Similarly in the normal approximation to the log odds ratio we have, from (5.2),

$$Z_1 = \log\left(\frac{16/14}{5/25}\right)\sqrt{\frac{(16 \times 5 \times 14 \times 25) \,/\, (30 \times 30)}{(16 \times 14/30) + (5 \times 25/30)}} = 2.8503$$

and

$$Z_2 = \log\left(\frac{18/12}{5/25}\right)\sqrt{\frac{18 \times 5 \times 12 \times 25/30 \times 30}{18 \times 12/30 + 5 \times 25/30}} = 3.2734.$$

Then

$$T_2 = \sqrt{3.2734^{+2} + \frac{(2.8503 - 0.5 \times 3.2734)^{+2}}{1-0.5^2}} = 3.5608.$$

From the normal distribution simulation we use

$$\text{no. of simulations with } (T_2 > 3.5608)/\text{total no. of simulations}$$

to get the p-value, which turns out to be 0.0004785.

Example 2: just one arm better than control

Here we consider another frequency table where, as in Example 1, the total number of success is 39, but only one experimental arm has more successes than the control. Let the observed proportion of successes in the control and treatments be 12/30, 20/30 and 7/30 respectively. As before we order them as $\hat{\pi}_0 = 12/30$, $\hat{\pi}_1 = 20/30$ and $\hat{\pi}_2 = 7/30$. To get the restricted MLEs by pooling we have $\hat{\pi}_1 > \hat{\pi}_0$, so pool to get $\tilde{\pi}_0^{(1)} = 32/60$. However $\hat{\pi}_2 < \tilde{\pi}_0^{(1)}$, so no more pooling is needed, $\tilde{\pi}_2 = \hat{\pi}_2 = 7/30$ and $J = 1$. Then, from equation (5.5), we calculate the observed value of λ for this data set as

$$\begin{aligned}\lambda &= 2\left(12\log\frac{12}{39} + 20\log\frac{20}{39} + 18\log\frac{18}{51} + 10\log\frac{10}{51}\right. \\ &\quad \left. + -30\log\frac{30}{90} - 30\log\frac{30}{90}\right) \\ &= 4.3392.\end{aligned}$$

Note that since one arm is worse than the control it is not used in λ, i.e. in equation (5.5) the summation is only up to J. The exact p-value in this case is 0.0455.

As before, for the simple normal approximation to the binomial distribution, from (5.1) we calculate

$$Z_1 = \frac{20/30 - 12/30}{\sqrt{39/90(1 - 39/90)(\frac{1}{30} + \frac{1}{30})}} = 2.0842$$

and

$$Z_2 = \frac{7/30 - 12/30}{\sqrt{39/90(1 - 39/90)(\frac{1}{30} + \frac{1}{30})}} = -1.3026.$$

Obviously Z_2 does not contribute, $T_2 = \sqrt{2.0842^2} = 2.0842$ and from the null distribution simulation we obtain the p-value of 0.0376. This is rather more optimistic than the exact p-value.

Similarly in the normal approximation to the log odds ratio we have, from (5.2),

$$Z_1 = \log\left(\frac{20/10}{12/18}\right)\sqrt{\frac{20 \times 12 \times 10 \times 18/30 \times 30}{20 \times 10/30 + 12 \times 18/30}} = 2.0440$$

and

$$Z_1 = \log\left(\frac{7/23}{12/18}\right)\sqrt{\frac{(7 \times 12 \times 23 \times 18)\,/\,(30 \times 30)}{(7 \times 23/30) + (12 \times 18/30)}} = -0.5971,$$

which does not contribute to T_2, so that $T_2 = Z_1 = 2.0440$ for which the p-value is 0.0412. This is also optimistic but slightly closer to the exact p-value.

5.6.2 Example 3: smaller total success rate

Consider another example with $N = 90$ where the success rate is smaller. Let the observed proportions be $\hat{\pi}_0 = 3/30$, $\hat{\pi}_1 = 8/30$ and $\hat{\pi}_2 = 10/30$, respectively. Then $\hat{\pi} = (3 + 8 + 10)/(30 + 30 + 30) = 21/90$. Conditioning on the total number of successes being equal to 21, the LRT statistic gives $\lambda = 5.2984$ and the corresponding exact p-value is 0.0286. Calculating $Z_1 = 1.5262$ and $Z_2 = 2.1366$ for the normal approximation to the binomial

Table 5.8: Power when $\pi_0 = 0.1, \pi_1 = 4/15, \pi_2 = 1/3$ and $N = 90$ with $\delta = 1/3$ for the large sample approximations for T_2 and T_∞ and for the exact conditional test.

	Power (%)					
	T_2			T_∞		
Scale	5%	2.5%	1%	5%	2.5%	1%
Probability	71.358	58.954	41.245	70.270	57.502	40.160
log OR	70.034	54.483	31.324	68.246	50.698	25.871
	λ					
Exact test	69.519	57.877	40.909			

distribution, from (5.1), then $T_2 = 2.2011$ and in this case the second best treatment contributes very little and the p-value is 0.0349. Similarly the normal approximation to log ORs, $Z_1 = 1.6122$ and $Z_2 = 2.0850$ from (5.2), and we have $T_2 = 2.1028$ which gives a p-value of 0.0361. In this case, the approximate p-values are far enough from the exact p-value to cause some concern.

For $T_\infty = 2.1366$ the p-value for the above example for the direct normal approximation is 0.0301 and, with the log OR, $T_\infty = 2.0850$ and the p-value is 0.0341.

5.6.3 Power study for Example 3

We can study the power of the approximate tests for any given values of the π_is and n_is. Here we consider the case where the π_is are all equal to their estimates in Example 3. 100,000 values were simulated from each of $Y_0 \sim Bin(30, 1/10)$, $Y_1 \sim Bin(30, 4/15)$ and $Y_2 \sim Bin(30, 1/3)$ and the proportion which rejected H_0 according to the two asymptotic tests, using the critical values from Chapter 3, and the exact test was counted. The results are shown in Table 5.8.

The results in Section 5.4, shown in Tables 5.3 and 5.4, suggested that both tests were slightly liberal for this sample size, so it is perhaps surprising that the test on the original scale is substantially more powerful than the tests on the log odds scale. We also see that T_2 is more powerful than T_∞. Given that the approximate tests are liberal, we would expect them to have higher power than the exact test. In this sense, the exact test performs remarkably well. Although more experience is needed, this example suggests that the exact test is very promising.

5.6.4 Example 4: four-arm trial

We obtain the exact distribution of λ for 4 arms using the example from Chuang-Stein and Tong (1995). The trial has low, medium and high doses of a new drug, comparing each with placebo. The sample size in each group is 200. For the placebo, $\hat{\pi}_0 = 50/200$, for the low dose, $\hat{\pi}_1 = 45/200$, for the medium dose, $\hat{\pi}_2 = 52/200$ and, for the high dose, $\hat{\pi}_3 = 72/200$. The LRT statistic has observed value $\lambda = 5.7318$ and, after considerable computing, we find the p-value is 0.0274. From their unconditional exact test, based on T_∞, Chuang-Stein and Tong (1995) rejected H_0 at the 5% level of significance, but did not give a p-value.

For this 4-arm example we calculate the large sample approximation both on the original scale and on the log odds scale. For the simple normal approximation to the binomial distribution, $\hat{\pi} = (50 + 45 + 52 + 72)/(4 \times 200) = 219/800$. As before we calculate $Z_1 = -0.5607$ (as the observed low dose is worse than the placebo), $Z_2 = 0.2243$ and $Z_3 = 2.4670$. There is no contribution to the test statistic from the medium or low doses. Then the test statistic is based on the high dose, i.e. $T_2 = Z_3$. From the normal distribution simulations we use

$$\frac{\text{no. of simulations with } (T_2 \geq 2.4670)}{\text{total no. of simulations}}$$

to get the p-value, which turns out to be 0.0229. T_∞ gives a p-value of 0.0184, agreeing with the conclusion of Chuang-Stein and Tong (1995) that H_0 is rejected at the 5% level of significance.

Similarly, using the log OR approximation we have $Z_1 = -0.5873$, $Z_2 = 0.2294$ and $Z_3 = 2.3512$. Then $T_2 = 2.3512$ and the p-value is 0.0305. Meanwhile T_∞ gives a p-value of 0.0249.

We can see in this case, with a large sample size, that the p-values are fairly close for all three methods which use the LRT methods and that methods based on the maximum give smaller p-values.

5.7 Discussion

We have explored a variety of methods for analysing binary data in this chapter. We find that the asymptotic methods work reasonably well with total sample sizes as small as 90, if the success probabilities are close to 0.5. If the success probabilities are 0.1, they are appropriate for large sample sizes, with at least 600 subjects in total. Perhaps surprisingly, the methods based on univariate likelihood ratio test statistics seem less good than those based on the direct normal approximation to the binomial distribution.

We have developed a conditional exact test and would recommend it for small and moderate sized samples. In fact, if the success probabilities are considerably smaller than those we have studied, this will be necessary even for much larger samples. It is probably advisable to use the exact test as long as the computing power available is able to cope with it. Even if it is not, it might be advisable to use a random-permutation test, in which we calculate λ for a large random sample of the possible contingency tables if $n_i\pi_i$ or $n_i(1-\pi_i)$ is less than 60. This is a trivial modification of the method described in Section 5.5. Example 4 is at, or slightly beyond, the limit of the size of data set for which it is possible to do the full set of permutations in a sensible amount of computing time.

Chapter 6

Two-stage adaptive designs

Sequential design of clinical trials is widely acknowledged to be beneficial, when it is possible, due to the possibility of saving resources and exposing fewer patients to experimental treatments, by allowing for possible early stopping. Adaptive design, in which the allocation of treatments to patients can also be changed in the light of data collected at early stages, has recently received much attention in the pharmaceutical industry. As with fixed designs, most of the literature on sequential and adaptive designs concerns two-arm trials, e.g. most of the contents of the books by Whitehead (1997) and Jennison and Turnbull (2000), although multi-arm trials have received attention in the recent literature. Some of this is concerned with all pairwise comparisons, while some deals with many-to-one comparisons with a control, as in this thesis.

The idea of most adaptive designs is to use the information from interim analyses to change the allocation of patients to treatments so that improved estimates or tests can be obtained. Trials comparing several experimental treatments with a control, as considered in this thesis, seem to be a good candidate for adaptive design, since inferior treatments can be dropped at an early stage, in order to get better comparisons of the good treatments with the control. They are usually only viable, however, for responses which are available very soon after treatment.

Adaptive designs for comparing several treatments with a control are considered in this chapter. Attention will be restricted to two-stage procedures, in which a decision about which treatment(s) to drop is made after analysing the data from stage 1. Results will be given for two experimental arms, but the methods are general for any number of arms. We consider equal allocation at each stage, since this is likely to be common in practice, but also briefly study some different pre-assigned allocations. We concentrate mainly on the case where the two stages use equal numbers of patients, but also briefly explore using fewer patients in the first stage. The results are compared, in terms of power and expected loss, with those from fixed trials.

Another possible application of the adaptive designs which allow treatments to be dropped, as considered here, is in seamless phase-II/III trials, which have received some attention recently. In these, rather than running a phase-II trial to select the best experimental treatment (e.g. the optimal dose or formulation), followed by a phase-III trial to establish improved efficacy compared with a control, a single two- (or more-)stage trial is run to simultaneously find the optimal dose and compare it with the control. If this is done using an adaptive two-stage design, stage 1 essentially plays the exploratory role of the phase-II trial in selecting the best treatment or treatments and checking that they are reasonable candidates, while stage 2 plays the confirmatory role of the phase-III trial in comparing them with the control. The benefit is that the comparison can be done using data from both stages, so that overall there is a considerable saving in resources.

In terms of the methodology described in this chapter, a seamless phase-II/III trial and an adaptively designed phase-III trial do not differ. All of the methodology described here can be appropriate to either type of trial. The crucial aspect, as usual, is that the hypotheses we discuss are relevant to the research questions of interest. This chapter only begins the exploration of adaptive designs for the hypothesis $H_0 : \Delta_i \leq 0 \; \forall i$, there being much more scope for research in this area.

The extension of our notation to deal with adaptive designs is described in Section 6.1 and the relevant literature is reviewed in Section 6.2. Several adaptive designs are studied and compared in Section 6.3 in the context of equal allocation to all treatments and stages of equal sizes. A brief study of some designs with unequal stage sizes is described in Section 6.4. Finally, some conclusions are drawn in Section 6.5.

6.1 Notation

The model and notations we defined in Chapter 2 are now reintroduced and extended to two-stage designs. Let N be the total number of patients in all arms in a fixed trial, with $n_i = \delta N$ patients on each experimental arm and $n_0 = (1 - I\delta)N$ patients on the control arm. We assume that we have Z_i, for $i = 1, \ldots, I$, which is a normalised statistic comparing treatment arm i with the control. Since we have large samples, $\mathbf{Z} = (Z_1, \ldots, Z_I)'$ is multivariate normal with $E(Z_i) = \Delta_i/\sigma$, $Var(Z_i) = 1$ and $\rho = Cov(Z_i, Z_j) = \delta/\{1 - (I -$

$1)\delta\}$, $i \neq j$, where $\sigma^2 = \{1-(I-1)\delta\}/\{\delta(1-I\delta)\}$. The Z_is are correlated due to their dependence on a common control arm. For $I=2$,

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \Delta_1/\sigma \\ \Delta_2/\sigma \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right).$$

With equal allocation, i.e. $\delta = 1/3$, the correlation is $\rho = 1/2$. This could be derived by defining Y_{ij} to be the (scaled) response from patient j on treatment $i = 0, 1, \ldots, I$, where $i = 0$ is the control arm. After rescaling to get unit variance, we have $Y_{0j} \sim N(\mu_0, 1)$ and $Y_{ij} \sim N(\mu_i, 1)$, $i = 1, \ldots, I$. All observations are assumed to be independent. Then $\Delta_i = \sqrt{N}(\mu_i - \mu_0)$ and $Z_i = \sqrt{N}\{(\bar{Y}_i - \bar{Y}_0)/\sigma\}$. Although such assumptions are not necessary, they make the derivation of the appropriate test statistics in an adaptive design simpler.

For a two-stage adaptive design, let $Z_{i.k}$ be a normalised univariate test statistic for comparing treatment i with the control obtained from the data in stage k and assume that responses are instantly available. For example, for normally distributed responses, let $Y_{ij.k}$ be the response from the jth patient on the ith treatment $i = 0, 1, \ldots, I$ in the kth stage, $k = 1, 2$. Let there be $N_{.k}$ patients in the kth stage and $n_{i.k} = \delta_{.k} N_{.k}$ patients on each experimental arm in stage k. Let $I_{.k}$ be the number of experimental arms studied at stage k. Then $n_{0.k} = (1 - I_{.k}\delta_{.k})N_{.k}$ and

$$Z_{i.k} = \sqrt{N_{.k}}\{(\bar{Y}_{i.k} - \bar{Y}_{0.k})/\sigma_{.k}\}, \tag{6.1}$$

where

$$\sigma^2_{.k} = \{1 - (I_{.k} - 1)\delta_{.k}\}/\{\delta_{.k}(1 - I_{.k}\delta_{.k})\}. \tag{6.2}$$

As before, it is not necessary to derive $Z_{i.k}$ from the normal case. The following methods apply asymptotically to any normalised univariate test statistics. Note, however, that the asymptotic results now require large samples in every stage.

When combining data from different stages we will define global normalised univariate test statistics Z_i, which will be based on sufficient statistics. They will be obtained from $Z_{i.k}$ in such a way as to ensure that $Var(Z_i) = 1$. To find the optimal way to combine stages, it is simplest to work with $\bar{Y}_{i.k} - \bar{Y}_{0.k}$, although the results generalise as usual.

As in the rest of this thesis, assume that the aim of the trial is to test the null hypothesis $H_0 : \Delta_i \leq 0$, $\forall i \in \{1, \ldots, I\}$ against $H_1 : \Delta_i > 0$ for at least one i and, if H_0 is rejected, to recommend the best treatment, or at least a good treatment. We will also sometimes refer to the simple null hypothesis $H_0^* : \boldsymbol{\Delta} = 0$ and the univariate null hypotheses $H_{0i} : \Delta_i \leq 0$ and $H_{0i}^* : \Delta_i = 0$.

As noted in earlier chapters, since selection of the best treatment is important, it is possible to correctly reject H_0, but to select an inferior treatment. If H_0 is rejected and the treatment selected is no better than the control, we say that a type-III error has been made. If H_0 is rejected and the treatment selected is better than the control, but suboptimal, we say that a type-IV error has been made. These concepts continue to be important in adaptive designs, but with the additional point that selection of the best treatment might be done before the end of the trial, so that a suboptimal treatment might be the only one studied in the second stage of the adaptive procedure.

6.2 Literature review

Although the general area of adaptive design has received a lot of attention, two-stage sequential adaptive design for comparing two or more experimental arms with a control has become an area of interest only relatively recently. Here we focus on that work which is most closely related to the work presented in this chapter.

Thall et al. (1988) suggested an adaptive procedure for binary response data, in which they used the angular transformation to achieve approximate normality. They allocate equal numbers to each arm in stage 1 and, if in stage 1 $\max(Z_{i.1}) < c_1$, for some stopping cutpoint c_1, then they stop the trial and accept $H_0^* : \Delta_i = 0,\ \forall\ i$. If $\max(Z_{i.1}) \geq c_1$, then they select the active arm $i^* = \arg\max(Z_{i.1})$ to be carried forward to stage 2. In stage 2, they allocate equal numbers of patients to control and treatment i^*. After stage 2, they do a Z-test based on data from treatment i^* and the control from both stages. The critical value of this test has to be adjusted to take account of the selection of the active arm with the largest estimate and also to take account of the fact that some trials stopped early. The adjustment is made to achieve size α. The stopping boundary c_1 is chosen in order to minimise the expected sample size for some pre-specified values of the true treatment effects Δ_1 and Δ_2. Clearly, the same procedure can be applied for any of the univariate test statistics Z_i discussed in this thesis.

Schaid et al. (1990) suggested using two-stage adaptive designs for survival data. They tested each active treatment against the control at stage 1 and, for each H_{0i}, they either accepted, rejected or failed to make a decision. They stopped if any treatment were found to be better than the control, or if all H_{0i} were accepted. Otherwise, they continued with all treatments i for which H_{0i} had been neither accepted nor rejected. Unlike most other work in this area, Schaid et al. did not adjust for the multiple comparisons.

Royston et al. (2003) developed a more realistic two-stage approach for survival data, in which the decision to stop or drop treatments after stage 1 is based on a surrogate outcome variable, which is assumed to be available more quickly than the survival outcome and also to be less variable. At stage 1 a one-sided test is carried out for each arm against the control and only those treatments which are significantly better than control are taken to stage 2. After stage 2 a one-sided test of each arm against the control is based on the primary survival outcome. Equal allocation is used at each stage and no adjustment is made for multiple comparisons, although Royston et al. (2003) suggest how such an adjustment can be made if required. The emphasis in this work is on the individual hypotheses H_{0i}. The major technical difficulty dealt with in this paper is the unknown correlation between the primary outcome and the surrogate outcome. This is an important practical consideration for survival data which could be explored further for the designs developed in this chapter.

Hughes (1993) was one of the first to take the approach of sequentially applying multiple comparison procedures in the context of clinical trials with several treatments against a control. However, unlike the work reported in this thesis, his procedure was based on tests of all pairwise comparisons, with adjustments to preserve the global size. At each stage experimental treatment i is dropped either if it is significantly worse than the control, or if it is significantly worse than another experimental treatment j, where the significance levels for testing against the control and against treatment j need not be the same.

Follmann et al. (1994) developed a procedure which, for multiple comparisons with a control, is based on sequentially testing each H_{0i}^* against the two-sided alternative. If H_{0i}^* is rejected in favour of the control, then treatment i is dropped. As soon as any H_{0i}^* is rejected in favour of the active treatment, the trial is stopped and that treatment is selected. If all H_{0i}^* are accepted, or all active arms are dropped then the trial is stopped in favour of the control. Proschan et al. (1994) compared different methods of adjusting for multiple comparisons, which take account of the multiple arms as well as the repeated testing, when using this procedure.

Betensky (1996) developed a fully sequential procedure, i.e. with continuous monitoring of immediately available responses, to compare three treatments, including the case when one was a control, for normally distributed responses. A sequential test of each H_{0i}^* against H_{1i} is carried out for $i = 1, 2$, with Bonferroni adjustment of the significance levels. When H_{0i}^* is accepted or rejected for any $i \in \{1, 2\}$, then treatment i or the control, respectively, is dropped and the procedure continues with a sequential test of the two remaining treatments. Betensky (1997) extended this method to deal with survival data.

In the work reported later in this chapter, we do not allow the control to be dropped and this simplifies matters to some extent due to the fact that less testing is needed.

Hellmich (2001) described group sequential adaptive procedures, to compare several treatments with a control, in which early stopping was allowed with either acceptance or rejection of the null hypothesis, i.e. due to either efficacy or futility, and arms could be dropped due to inferiority. His aim was different from that in this thesis, being to test each individual hypothesis H_{0i}^*, while controlling the overall error rate. This leads to considerable technical difficulties which do not arise in the work reported in this chapter, but Hellmich made considerable progress by using closed testing procedures of the type described in Chapter 1.

Vincent et al. (2002) describe a more flexible multi-stage approach for comparing two experimental arms with a control. In keeping with the approach of Whitehead (1997), they set up any number of interim analyses at pre-specified (not necessarily equally-spaced) expected values of the Fisher information. At the kth interim analysis, each Z_i (calculated from the data collected so far) is used to test each $H_{0i}^* : \Delta_i = 0$, against the one-sided alternative and also to decide whether to continue to use treatment i. When some H_{0i}^* are rejected, we stop and conclude that experimental treatment i is better than the control. If two or more arms are found to be better than the control, the best is chosen on the basis of some point estimate, e.g. the corresponding Z_i. If all experimental arms are dropped, we stop and accept H_0^*. Vincent et al. developed the methodology to work out the stopping boundaries for each interim analysis, in order to control the global size of the test and the power. They worked with the adjusted power, as defined in Chapter 3, and used it to calculate expected sample sizes. This work is more ambitious in scope than what is attempted in this chapter, especially due to the need for testing H_0 at each interim analysis. In keeping with the spirit of this thesis, we avoid multiple testing issues by only carrying out a single hypothesis test and this means that we can avoid some of the complications of Vincent et al.'s work. However, their paper is very close to the work in this thesis in the recognition it gives to the selection of the best treatment, as well as the global hypothesis test. In this chapter, we consider some other test statistics which might be more appropriate than that used by Vincent et al.

The work of Thall et al. (1988) was generalised by Stallard and Todd (2003) to any normally distributed test statistics based on the efficient score, following the methodology of Whitehead (1997), which covers most of our cases. They also allowed any number of experimental arms and any number of interim analyses, although after stage 1, they always select the best treatment and continue with that treatment and the control.

Kelly et al. (2005) extended this work to the more general procedure in which any number of experimental treatments can be dropped at any stage. Testing at each stage is based on Dunnett-type tests. If any H_{0i} is rejected at any stage, the trial is stopped in favour of treatment i, while the decision to drop treatment i at any stage is based on the accumulated Z_i up to that stage being below some boundary.

The methodology in the papers of Vincent et al. (2002), Stallard and Todd (2003) and Kelly et al. (2005) is essentially that of sequential designs for two-arms applied to each pairwise comparison, with adjustment of the significance levels for multiple comparisons with the control based on Dunnett's procedure. Some authors therefore refer to this as *sequential* design, reserving the name *adaptive* design for procedures which allow unplanned changes to the design at interim analyses.

There is another large body of literature on unplanned adaptations to designs, such as sample-size recalculation, at interim analyses. This work, following Bauer (1989) and many other papers by Bauer and his coworkers, was reviewed by Hellmich and Hommel (2004) and by Posch et al. (2005). The important idea is that the stages are first analysed separately and then the p-values combined. A slightly different approach is the conditional error function method of Müller and Schäfer (2001). Jennison and Turnbull (2003) showed that these methods, and any unplanned redesign that maintains the overall type-I error rate whenever the design is modified, must preserve the conditional type-I error rate and hence all such procedures are closely related even though they are described differently.

The topic of unplanned adaptations which preserve the type-I error rate was recently reviewed by Jennison and Turnbull (2006), who showed that, because these methods require the use of nonsufficient statistics, they are inefficient. Jennison and Turnbull recommended that such procedures should only be used as a last resort, when crucial assumptions are clearly contradicted at the interim analysis. We will not consider such methods here, although clearly the trial could be redesigned after stage 1 of the procedures described in this chapter.

Koenig et al. (2008) also used the conditional error function approach and showed, in some limited situations, that an adaptive Dunnett test using this principle could perform well. One of the specific procedures they assessed was to use equal numbers of patients in each of two stages. In stage 1, equal numbers are allocated to two active treatments and a control. In stage 2, equal numbers of patients are allocated to the treatment with the best

Table 6.1: Summary of adaptive design strategies.

Procedure	At end of stage 1
1	Drop all experimental arms except most promising
2	Stop for futility? Drop all experimental arms except most promising
3	Drop between 0 and I experimental arms depending on results $0 =$ nothing dropped, $I =$ stop for futility $I - 1 =$ continue with one experimental arm and control

estimate at the interim analysis and the control. Finally, Dunnett's test is carried out, with suitable adjustment of the nominal significance level, using data from both stages on the arms carried forward to stage 2. We will study a similar procedure below.

Most of the above literature assumes that the null hypothesis will be H_0^* and avoids the issue of this implying a model which does not allow experimental treatments to be worse than the control. The adaptive procedures described in this chapter, like the rest of the work in this thesis, will be based on testing H_0 directly. In the next section, we compare several rules for two-stage designs where treatments can be dropped after the first stage, but H_0 can only be rejected after the second stage.

6.3 Comparisons of adaptive procedures

We consider several different rules for deciding which arms to carry forward to stage 2. We will consider only procedures in which the control treatment is used in each stage and no new treatments are introduced at stage 2. First we consider always forwarding one experimental arm to stage 2, which allows us to assess the benefit of dropping arms. Second, we either stop or forward one experimental arm, so that we can see the possible costs of reducing the expected sample size. Third, we stop or forward a number of arms determined by the results from stage 1, which potentially allows us to obtain the benefits of early stopping or early selection, while also maintaining the possibility of using the advantages of our test statistic T_2 when more than one experimental arm seems promising. As before, we assume that the objective is to test the null hypothesis $H_0 : \Delta_i \leq 0, \ \forall i$, against the alternative $H_1 : \Delta_i > 0$, for at least one i, and, if H_0 is rejected, to select the best experimental treatment. The three procedures considered are summarised in Table 6.1.

In the first procedure considered, selection is always carried out after stage 1 and the test after stage 2. In the second procedure, H_0 might be accepted after stage 1 or accepted or rejected after stage 2, while selection is carried out after stage 1. In the third procedure, H_0 might be accepted after stage 1, or accepted or rejected after stage 2, while selection might be carried out after either stage 1 or stage 2. The details of the procedures are described in the relevant subsections. We emphasise that, unlike much of the work reviewed in the previous section, we do not allow early stopping with rejection of H_0. Also note that arms are dropped not only because there is evidence that they are inferior to the control, but rather because the other active arm seems like a better bet for establishing superiority over the control.

Within each of the next three subsections, we first describe the proposed procedure as it might apply in general and then study the properties of some specific special cases through simulation. The simulation studies are restricted to two experimental arms, equal allocation at each stage and a few values of the decision boundaries. Although these results are not comprehensive, they are enough to compare methods and evaluate how attractive they are for practical use.

In fact, we can define a more general procedure as follows.

Stage 1: Allocate $\delta_{.1}N_{.1}$ patients to each experimental arm and $(1-I\delta_{.1})N$ arms to the control, where $0<\delta_{.1}<1/I$. If, for some $1\leq I_{.2}\leq I$,

$$X_{1.1}>c_1, X_{2.1}>c_2\rho X_{1.1},\ldots,X_{I_{.2}.1}>c_{I_{.2}}\frac{\rho}{1+(I_{.2}-2)\rho}\sum_{j=1}^{I_{.2}-1}X_{j.1},$$

but

$$X_{(I_{.2}+1).1}\leq c_{I_{.2}+1}\frac{\rho}{1+(I_{.2}-1)\rho}\sum_{j=1}^{I_{.2}}X_{j.1},$$

where $X_{j.1}$ is the jth largest of $Z_{1.1},\ldots,Z_{I.1}$ and $c_1,\ldots,c_I$ are known constants, then carry forward the $I_{.2}$ experimental arms corresponding to $X_{1.1},\ldots,X_{I_{.2}.1}$, along with the control.

Stage 2: Allocate a further $\delta_{.2}N_{.2}$ patients to each remaining experimental arm and $(1-I_{.2}\delta_{.2})N_{.2}$ to the control arm, where $0<\delta_{.2}<1/I_2$ and $N_{.1}+N_{.2}=N$. Test using the statistic T_k, for some k, defined in chapter 2 (based on the control plus $I_{.2}$ treatments used at this stage) with a critical value adjusted to take account of the bias induced by the adaptive design.

Recall that k is an index defining which member of a family of test statistics is to be used.

As k is increased, T_k makes less use of the information on the treatments other than that which is estimated to be best.

In this section, we consider the following procedures:

- Procedure 1: $N_{.1} = N/2$, $\delta_{.1} = 1/(I+1)$, $c_1 \to -\infty$, $c_j \to \infty$, $j = 2, \dots, I$, $\delta_{.2} = 1/2$ and T_k (which is the same for any k in this case) is used;
- Procedure 2: $N_{.1} = N/2$, $\delta_{.1} = 1/(I+1)$, $c_j \to \infty$, $j = 2, \dots, I$, $\delta_{.2} = 1/2$ and T_k is used, with various values of c_1;
- Procedure 3: $N_{.1} = N/2$, $\delta_{.1} = 1/(I+1)$, $\delta_{.2} = 1/(I_{.2}+1)$ and T_1, T_2 or T_∞ are used, with various values of c_1 and c_j, for $j = 2, \dots, I$.

In Section 6.4, we will consider:

- Procedure 3 with $I = 2$ as above, but with $N_{.1} = N/4$, with various values of c_1 and c_2.

6.3.1 Procedure 1: always forward one experimental treatment

We begin with the simplest adaptive design, which is one of those studied by Koenig et al. (2008). Allocate half of the patients to each of stage 1 and stage 2 and allocate $\frac{N}{2(I+1)}$ patients to each arm in stage 1. We always carry forward the one experimental arm which gives the best results in stage 1, along with the control, to stage 2 and allocate $\frac{N}{4}$ patients to each. Thus the interim analysis is used for selection but not for testing the null hypothesis. After stage 2, the data on the selected experimental treatment and the control from both stages are combined in order to carry out a test similar to a Z-test to accept or reject H_0, but with the critical value chosen to ensure the correct size in the adaptive design.

We now use the defined notation for $Z_{i.k}$ from Section 6.1 and show how the data from the two stages should be combined in this case.

At stage 1, we allocate $\frac{N}{2(I+1)}$ patients to each arm. Then $n_{.1} = \delta_{.1} N_{.1} = N/\{2(I+1)\}$ and $I_{.1} = I$, so that $\sigma^2_{.1} = \frac{1-1/(I+1)}{1/(I+1)(1-I/(I+1))} = 2(I+1)$ and $Z_{i.1} = \sqrt{N/2}\{(\bar{Y}_{i.1} - \bar{Y}_{0.1})/\sqrt{2(I+1)}\} = \sqrt{N}(\bar{Y}_{i.1} - \bar{Y}_{0.1})/\{2\sqrt{I+1}\}$.

At stage 2, $n_{.2} = \delta_{.2} N_{.2} = \frac{1}{2}\frac{N}{2} = N/4$ and $I_{.2} = 1$, so that $\sigma^2_{.2} = \frac{1}{1/2(1-1/2)} = 4$ and $Z_{i^*.2} = \sqrt{N/2}\{(\bar{Y}_{i^*.2} - \bar{Y}_{0.2})/\sqrt{4}\} = \sqrt{N}(\bar{Y}_{i^*.2} - \bar{Y}_{0.2})/\sqrt{8}$, where treatment i^* was selected after stage 1.

We will combine the estimated differences between the selected experimental treatment and the control from stage 1 and stage 2 as $Z_{i^*} = \gamma_1 Z_{i^*.1} + \gamma_2 Z_{i^*.2}$, for constants γ_1 and γ_2 chosen to give equal weight to each patient such that $Var(Z_{i^*}) = 1$. In calculating these constants, we ignore the selection of treatment i^* at stage 1. In fact, $Var(Z_{i^*.1}) < 1$, since a particular treatment is unlikely to be selected when the corresponding $Z_{i.1}$ turns out to be in the lower tail of its distribution. If, for example, we select treatment 1, we actually work with the distribution of Z_1, whereas strictly speaking we should work with the conditional distribution of $Z_1 | Z_{1.1} > Z_{j.1}\ \forall j = 2, \ldots, I$. We will ignore this complication in the rest of this chapter.

With equal allocation, it is clear that we can write Z_{i^*} as a constant multiple of $\bar{Y}_{i^*} - \bar{Y}_0$. Then writing $\bar{Y}_i$ in terms of $\bar{Y}_{i.k}$ for $i = 0, i^*$, for some constant K, we have

$$\begin{aligned} Z_{i^*} &= K\left(\bar{Y}_{i^*} - \bar{Y}_0\right) \\ &= K\left[\frac{\frac{N}{2(I+1)}\bar{Y}_{i^*.1} + \frac{N}{4}\bar{Y}_{i^*.2}}{\frac{N}{2(I+1)} + \frac{N}{4}} - \frac{\frac{N}{2(I+1)}\bar{Y}_{0.1} + \frac{N}{4}\bar{Y}_{0.2}}{\frac{N}{2(I+1)} + \frac{N}{4}}\right] \\ &= \frac{K}{I+3}\left[2\left(\bar{Y}_{i^*.1} - \bar{Y}_{0.1}\right) + (I+1)\left(\bar{Y}_{i^*.2} - \bar{Y}_{0.2}\right)\right]. \end{aligned}$$

Substituting $\left(\bar{Y}_{i^*.1} - \bar{Y}_{0.1}\right) = \frac{2\sqrt{I+1}Z_{i^*.1}}{\sqrt{N}}$ and $\left(\bar{Y}_{i^*.2} - \bar{Y}_{0.2}\right) = \frac{\sqrt{8}Z_{i^*.2}}{\sqrt{N}}$, we have

$$\begin{aligned} Z_{i^*} &= \frac{K}{I+3}\left[2\frac{2\sqrt{I+1}Z_{i^*.1}}{\sqrt{N}} + (I+1)\frac{\sqrt{8}Z_{i^*.2}}{\sqrt{N}}\right] \\ &= \frac{2K}{(I+3)\sqrt{N}}\left[2\sqrt{I+1}Z_{i^*.1} + (I+1)\sqrt{2}Z_{i^*.2}\right] \\ \therefore Z_{i^*} &= K_2\left(2\sqrt{I+1}Z_{i^*.1} + (I+1)\sqrt{2}Z_{i^*.2}\right), \end{aligned}$$

for some constant K_2.

Now, in order to ensure that $Var\left(Z_{i^*}\right) = 1$, we require

$$Var\left[K_2\left\{2\sqrt{I+1}Z_{i^*.1} + (I+1)\sqrt{2}Z_{i^*.2}\right\}\right] = 1.$$

Note that $Z_{i^*.1}$ and $Z_{i^*.2}$ are independent, since patients in stages 1 and 2 are different and $Var\left(Z_{i^*.k}\right) = 1$. Hence,

$$\begin{aligned} K_2^2\left[Var\left(2\sqrt{I+1}Z_{i^*.1}\right) + Var\left\{(I+1)\sqrt{2}Z_{i^*.2}\right\}\right] &= 1 \\ \Longrightarrow K_2^2\left\{4(I+1)Var\left(Z_{i^*.1}\right) + 2(I+1)^2 Var\left(Z_{i^*.2}\right)\right\} &= 1 \end{aligned}$$

$$\therefore K_2 = \frac{1}{\sqrt{2(I+1)(I+3)}}$$

Table 6.2: Cutpoints of the rejection region for Procedure 1

5%	2.5%	1%
1.8460	2.1515	2.5049

and, therefore,

$$\begin{aligned} Z_{i^*} &= \frac{1}{\sqrt{2(I+1)(I+3)}} \left(2\sqrt{I+1} Z_{i^*.1} + \sqrt{2}(I+1) Z_{i^*.2} \right) \\ &= \frac{1}{\sqrt{I+3}} \left(\sqrt{2} Z_{i^*.1} + \sqrt{I+1} Z_{i^*.2} \right). \end{aligned}$$

Then the test statistic used is

$$T_k = \max\left(0, Z_{i^*}\right),$$

which is the same for any k, since only one arm was forwarded.

For three-arm trials, $I = 2$, $\sigma^2_{.1} = 6$ and

$$Z_{i^*} = \frac{1}{\sqrt{5}} \left(\sqrt{2} Z_{i^*.1} + \sqrt{3} Z_{i^*.2} \right). \tag{6.3}$$

Simulation and results

The null distributions for $I = 2$ were approximated as follows. As in the fixed design, it is clear from Theorem 13 in Chapter 3 that $P(\text{Reject } H_0|H_0 \text{ true})$ is maximised when $\mathbf{\Delta} = 0$. Therefore, to determine the boundary of the rejection region for the test for significance level α, we can simulate under $\mathbf{\Delta} = \mathbf{0}$. For stage 1, we simulated two million values of $Z_{1.1}$ and $Z_{2.1}$ from a bivariate normal distribution with expectations zero, unit variances and correlation $\rho = 0.5$ (i.e. equal allocation). Then, for stage 2, we simulated two million values of $Z_{i^*.2}$ from a univariate standard normal distribution. The stages were combined to give two million simulated values of Z_{i^*}, from which critical values were calculated for 5%, 2.5% and 1% by ensuring that the appropriate proportion of simulations l led to rejection after stage 2. Note that, since our critical values are based on simulations, we expect them to be more accurate than those of Koenig et al. (2008) who used conservative methods.

The cutpoints, as shown in Table 6.2, are smaller than the corresponding critical values for fixed trials, due to the selection bias induced by the adaptive design and also probably because, as noted above, the true variance of Z_{i^*} is less than one.

As in the fixed trial, the power is calculated as the proportion of 100,000 simulations which rejected H_0 using different values of the scaled mean differences between experimental treatments and the control, d_1 and d_2, and expected loss is calculated in a similar way. The same simulation errors were used for all values of d_1 and d_2. Note, however, that from a particular simulation different arms can be forwarded for different values of d_1 and d_2, depending on the results from stage 1.

Similar to the results obtained in fixed designs with different allocations, care is needed in choosing the appropriate scaling of d_1 and d_2 for the adaptive designs. In particular, the standardisations used for $Z_{i.1}$ and $Z_{i.2}$ are different, so we cannot simply use the obvious d_is for each stage. Instead, we fix the scalings of d_i to be those suitable for the fixed design with equal allocation, and rescale them internally to get $d_{i.1}$ and $d_{i.2}$ corresponding to $Z_{i.1}$ and $Z_{i.2}$, as follows.

For the fixed design with equal allocation in Chapter 3, we had $Z_i = \sqrt{N}\left(\bar{Y}_i - \bar{Y}_0\right)/\sigma$ and $E(Z_i) = \Delta_i/\sigma$ where $\Delta_i = \sqrt{N}\left(\mu_i - \mu_0\right)$ and we defined $d_i = \Delta_i/\sigma = \sqrt{N}\left(\mu_i - \mu_0\right)/\sqrt{6}$. For the adaptive design, we define

$$d_{i.1} = \sqrt{N/2}\left(\mu_i - \mu_0\right)/\sigma_{.1} = \sqrt{\frac{N}{12}}\left(\mu_i - \mu_0\right) = \frac{d_i}{\sqrt{2}}$$

and

$$d_{i.2} = \sqrt{N/2}\left(\mu_i - \mu_0\right)/\sigma_{.2} = \sqrt{\frac{N}{8}}\left(\mu_i - \mu_0\right) = \frac{\sqrt{3}d_i}{2},$$

where $\sigma_{.1} = \sqrt{6}$ and $\sigma_{.2} = 2$ as above.

The estimated powers are shown in Table 6.3, along with the corresponding results for fixed trials, with equal allocation, using T_2 and T_∞. Comparing these, there is an overall power gain from using the adaptive design. However, when $\Delta_2 = 0$ the gain is relatively higher than when $\Delta_2 = 0.5\Delta_1$ and the gain is least when $\Delta_2 = \Delta_1$. This is not surprising, since the one of the benefits of the adaptive design is that, by allowing early selection, it improves the comparison of the best treatment with the control. However, when the treatments are equally good, this benefit is lost.

The expected losses for the same situations are shown in Table 6.4, along with the corresponding results from the fixed design. When $\Delta_2 = 0$ the expected loss is smaller than with the fixed trial. When $\Delta_2 = 0.5\Delta_1$ most losses continue to be smaller in a two-stage design although to a lesser extent. There still seems to be some difference between the fixed and two-stage designs when $\Delta_1 = \Delta_2$, although if we refer back to Chapter 3, we find that with T_1, the fixed design is just as good. Since the treatment selection is based on only half the number of patients as in the fixed design, there are more type-III/IV errors when the adaptive design is used. On the other hand, when a correct selection is made, more effort goes into comparing the best treatment with the control, so that fewer type-II

Table 6.3: Power of Procedure 1 and the fixed design at 5, 2.5 and 1%.

		5%			2.5%			1%		
		Adapt.	Fixed		Adapt.	Fixed		Adapt.	Fixed	
d_1	d_2		T_2	T_∞		T_2	T_∞		T_2	T_∞
2.5	0	80.902	71.169	72.242	72.323	59.915	61.565	60.081	45.959	47.891
3	0	92.328	85.477	86.330	87.511	77.348	78.625	79.357	65.451	67.313
3.5	0	97.596	93.956	94.398	95.616	89.460	90.294	91.583	81.522	82.960
2.5	1.25	79.082	74.586	74.349	70.103	63.821	63.574	57.729	49.731	49.513
3	1.5	90.229	87.579	87.490	84.656	79.988	79.871	76.129	68.576	68.493
3.5	1.75	96.022	94.926	94.909	93.343	90.888	90.916	88.542	83.525	83.644
2	2	74.499	71.518	70.028	63.332	60.075	58.011	49.180	45.572	43.020
2.5	2.5	89.215	87.151	85.937	82.188	79.358	77.255	71.014	67.465	64.408
3	3	96.568	95.565	94.877	93.357	91.699	90.340	87.195	84.595	82.177

errors will be made. When $\Delta_2 = 0$, type-III/IV errors are very rare anyway, so this is the situation in which we get most benefit from using this adaptive design. When Δ_1 and Δ_2 have similar values, this adaptive procedure might be undesirable, since it always drops a potentially good treatment at stage 1.

6.3.2 Procedure 2: Stop or forward one experimental treatment

This procedure is similar to Procedure 1, except that we now allow for early stopping if neither of the experimental arms shows sufficient evidence of being better than the control after stage 1. In this case, we will accept H_0 without carrying out stage 2. This has the benefit that, on average, fewer patients will be used in cases where H_0 is accepted. Of course, the corresponding disadvantage is that we will lose power, since we will stop in some cases in which we would have continued and rejected H_0. For this procedure, it is necessary to define the boundary value for deciding whether or not to stop after stage 1. We choose a constant c_1 so that we stop and accept H_0 if $\max(Z_{i.1}) < c_1$. This procedure is essentially that of Thall et al. (1988), except that they considered only the angular transformation for binary responses, or equivalently it is a special case of the procedure of Stallard and Todd (2003), with no early rejection of H_0. Note that as $c_1 \to -\infty$, this procedure converges to Procedure 1 and as c_1 increases we are more likely to stop.

If we stop early, then $T_k = 0$ and no further calculation is needed. If stage 2 is run, then the data from the stages are combined in exactly the same way as for Procedure 1.

Table 6.4: Expected loss of Procedure 1 and the fixed design at 5, 2.5 and 1%.

		5%			2.5%			1%		
		Adapt.	Fixed		Adapt.	Fixed		Adapt.	Fixed	
d_1	d_2		T_2	T_∞		T_2	T_∞		T_2	T_∞
2.5	0	0.4872	0.7260	0.6985	0.6973	1.0044	0.9639	1.0004	1.3533	1.3042
3	0	0.2361	0.4374	0.4116	0.3782	0.6808	0.6423	0.6210	1.0374	0.9812
3.5	0	0.0875	0.2119	0.1964	0.1554	0.3692	0.3400	0.2955	0.6470	0.5966
2.5	1.25	0.6398	0.7170	0.7186	0.8357	0.9705	0.9706	1.1140	1.3041	1.3027
3	1.5	0.4306	0.4490	0.4485	0.5699	0.6664	0.6652	0.7949	0.9955	0.9911
3.5	1.75	0.2815	0.2384	0.2371	0.3544	0.3743	0.3706	0.4943	0.6240	0.6157
2	2	0.5100	0.5696	0.5994	0.7334	0.7985	0.8398	1.0164	1.4841	1.4094
2.5	2.5	0.2696	0.3212	0.3516	0.4453	0.5161	0.5686	0.7247	0.8134	0.8898
3	3	0.1030	0.1330	0.1537	0.1993	0.2490	0.2898	0.3842	0.4622	0.5347

Simulation and results

This procedure was evaluated using the simulation results for Procedure 1, described above. For each simulation run, both for the null distribution and the different values of d_1 and d_2, we checked whether the trial would have been stopped early by comparing $Z_{i^*.1}$ with each value of c_1. If the trial would have stopped, the final value of Z_{i^*} calculated was replaced by zero.

The cutpoints are shown in Table 6.5, again calculated so that the correct proportion of trials is rejected under H_0 using the adaptive procedure. As expected, these become smaller as c_1 increases, i.e. as stopping becomes easier, since some cases in which H_0 would have been rejected are now not forwarded to stage 2. As noted above, as $c_1 \to -\infty$, this procedure converges to Procedure 1 and we see that with $c_1 = -1.5$ it is almost identical.

The percentages of trials stopping early are shown in Table 6.6. We see that if H_0 is true, then we stop early only about 1.8% of the time when $c_1 = -1.5$, while with $c_1 = 0.5$ we stop early more than half of the time. Ideally, we would choose a value of c_1 which makes early stopping likely when H_0 is true, but unlikely when H_0 is false, especially when d_1 is such that the power of a fixed design is large, say greater than 60%. Of course a compromise is necessary and, from Table 6.6, values of c_1 equal to -0.5 or 0 seem like a reasonable compromise. The former stops early almost 1/6 of the time when H_0 is true and no more than 1% of the time for the other values of d_1 and d_2 shown, while the latter stops early 1/3 of the time (theoretically, the value in the table being an estimate) when

Table 6.5: Cutpoints of the rejection region for Procedure 2 with different stopping cutpoints.

c_1	5%	2.5%	1%
-1.5	1.8460	2.1515	2.5049
-1	1.8458	2.1514	2.5048
-.5	1.8442	2.1509	2.5045
0	1.8350	2.1464	2.5023
.5	1.7993	2.1258	2.4919
1	1.6882	2.0590	2.4536

H_0 is true and no more than 4% of the time for the other values of d_1 and d_2.

Note that in this adaptive design the sample size is a random variable, taking value $N/2$ with probability estimated by the values in Table 6.6 and value N otherwise. Hence we can calculated the expected sample sizes under H_0 in each case and these are shown in Table 6.7 as a proportion of the fixed sample size. For example, with $c_1 = -0.5$, the expected sample size is $\frac{N}{2}0.1631 + N(1 - 0.1631) = 0.9185N$. Note that, for $c_1 = 0$, the theoretical value is $5/6 = 0.8333$, so the simulations seem very accurate.

The powers are shown in Tables 6.8-6.10. The power falls as c_1 is increased, indicating that the increase in the probability of rejection implied by the lower cutpoints is less than the decrease in the probability of rejection caused by stopping early in some cases in which we would have gone on to reject H_0. Comparing with the fixed design, shown in Table 6.3, this rule gives higher power when $\Delta_2 = 0$ except for $c_1 = 1$. Because we are allowed to stop early in Procedure 2, and despite the lower cutpoints, the power is smaller than in Procedure 1 and decreases as c_1 increases, since in this case we stop early more often. However, we see that the loss of power is negligible up to $c_1 = -0.5$ and very small for $c_1 = 0$. For positive values of c_1, the power decreases more quickly, but is still not very large and such values could be used if there is particular pressure to reduce the expected sample size.

The corresponding expected losses are shown in Tables 6.11-6.13. Again, we see that the expected loss increases with c_1, but that the increase is very small up to $c_1 = 0$. In some cases, however, we see that the expected loss is much greater for positive values of c_1. This is because the trial has been stopped in too many cases where we would have gone on to reject H_0 had it been continued.

Table 6.6: Percentages of trials stopping early for Procedure 2 with different stopping cutpoints and values of d_1 and d_2.

		c_1					
d_1	d_2	-1.5	-1	-.5	0	.5	1
0	0	1.819	6.229	16.314	33.319	54.624	74.517
2.5	0	0.026	0.174	0.933	3.485	9.727	21.542
3	0	0.007	0.060	0.356	1.534	5.101	12.930
3.5	0	0.001	0.015	0.122	0.607	2.374	6.997
2.5	1.25	0.011	0.083	0.533	2.305	7.315	17.996
3	1.5	0.001	0.026	0.196	0.964	3.701	10.624
3.5	1.75	0.000	0.005	0.063	0.356	1.683	5.631
2	2	0.010	0.084	0.531	2.291	7.647	18.851
2.5	2.5	0.001	0.020	0.160	0.890	3.349	10.150
3	3	0.000	0.002	0.037	0.264	1.331	4.790

Table 6.7: Expected sample sizes under H_0 for Procedure 2 with different stopping cutpoints relative to fixed sample size.

c_1					
-1.5	-1	-.5	0	.5	1
0.9909	0.9680	0.9185	0.8334	0.7269	0.6274

Table 6.8: Power for Procedure 2 at 5% with different stopping cutpoints.

		c_1					
d_1	d_2	-1.5	-1	-.5	0	.5	1
2.5	0	80.901	80.892	80.773	80.210	77.990	71.927
3	0	92.328	92.312	92.228	91.762	89.898	84.456
3.5	0	97.596	97.592	97.550	97.278	96.022	92.163
2.5	1.25	79.082	79.084	79.044	78.815	77.559	73.273
3	1.5	90.229	90.222	90.196	90.004	88.917	85.035
3.5	1.75	96.022	96.021	96.003	95.905	95.152	92.441
2	2	74.499	74.499	74.503	74.450	73.605	70.315
2.5	2.5	89.215	89.215	89.206	89.121	88.431	85.246
3	3	96.568	96.566	96.567	96.518	96.063	93.711

Table 6.9: Power for Procedure 2 at 2.5% with different stopping cutpoints.

		c_1					
d_1	d_2	-1.5	-1	-.5	0	.5	1
2.5	0	72.323	72.317	72.248	71.886	70.419	66.164
3	0	87.511	87.503	87.446	87.111	85.688	81.366
3.5	0	95.616	95.612	95.581	95.368	94.365	90.967
2.5	1.25	70.103	70.099	70.080	69.939	69.139	66.321
3	1.5	84.656	84.653	84.641	84.499	83.716	80.844
3.5	1.75	93.343	93.341	93.334	93.253	92.693	90.347
2	2	63.332	63.333	63.329	63.301	62.900	61.045
2.5	2.5	82.188	82.189	82.185	82.168	81.782	79.691
3	3	93.357	93.356	93.359	93.330	92.990	91.303

Table 6.10: Power for Procedure 2 at 1% with different stopping cutpoints.

		c_1					
d_1	d_2	-1.5	-1	-.5	0	.5	1
2.5	0	60.081	60.080	60.065	59.889	59.106	56.554
3	0	79.357	79.355	79.337	79.147	78.259	75.225
3.5	0	91.583	91.582	91.561	91.403	90.681	88.058
2.5	1.25	57.729	57.728	57.735	57.678	57.273	55.644
3	1.5	76.129	76.128	76.126	76.057	75.585	73.633
3.5	1.75	88.542	88.541	88.534	88.472	88.076	86.360
2	2	49.180	49.180	49.188	49.191	49.059	48.189
2.5	2.5	71.014	71.014	71.022	71.023	70.886	69.765
3	3	87.195	87.195	87.204	87.194	87.013	85.928

Table 6.11: Expected Loss for Procedure 2 at 5% with different stopping cutpoints.

		c_1					
d_1	d_2	-1.5	-1	-.5	0	.5	1
2.5	0	0.4873	0.4875	0.4905	0.5047	0.5606	0.7138
3	0	0.2361	0.2366	0.2392	0.2533	0.3095	0.4740
3.5	0	0.0875	0.0876	0.0892	0.0988	0.1430	0.2787
2.5	1.25	0.6398	0.6398	0.6408	0.6469	0.6791	0.7854
3	1.5	0.4307	0.4309	0.4317	0.4380	0.4708	0.5859
3.5	1.75	0.2815	0.2815	0.2822	0.2861	0.3124	0.4064
2	2	0.5100	0.5100	0.5099	0.5110	0.5279	0.5937
2.5	2.5	0.2696	0.2696	0.2699	0.2720	0.2892	0.3689
3	3	0.1030	0.1030	0.1030	0.1045	0.1181	0.1887

Table 6.12: Expected Loss for Procedure 2 at 2.5% with different stopping cutpoints.

		c_1					
d_1	d_2	-1.5	-1	-.5	0	.5	1
2.5	0	0.6973	0.6975	0.6992	0.7082	0.7450	0.8521
3	0	0.3782	0.3784	0.3801	0.3901	0.4329	0.5632
3.5	0	0.1554	0.1556	0.1567	0.1641	0.1993	0.3186
2.5	1.25	0.8357	0.8358	0.8363	0.8400	0.8606	0.9309
3	1.5	0.5699	0.5700	0.5704	0.5748	0.5991	0.6858
3.5	1.75	0.3545	0.3545	0.3548	0.3580	0.3780	0.4599
2	2	0.7334	0.7333	0.7334	0.7340	0.7420	0.7791
2.5	2.5	0.4453	0.4453	0.4454	0.4458	0.4555	0.5077
3	3	0.1993	0.1993	0.1992	0.2001	0.2103	0.2609

Table 6.13: Expected Loss for Procedure 2 at 1% with different stopping cutpoints.

		c_1					
d_1	d_2	-1.5	-1	-.5	0	.5	1
2.5	0	1.0004	1.0004	1.0008	1.0052	1.0248	1.0888
3	0	0.6210	0.6210	0.6216	0.6273	0.6540	0.7451
3.5	0	0.2955	0.2956	0.2963	0.3018	0.3271	0.4190
2.5	1.25	1.1140	1.1140	1.1139	1.1155	1.1258	1.1668
3	1.5	0.7949	0.7949	0.7950	0.7973	0.8118	0.8714
3.5	1.75	0.4943	0.4943	0.4946	0.4968	0.5111	0.5718
2	2	1.0164	1.0164	1.0162	1.0162	1.0188	1.0362
2.5	2.5	0.7247	0.7247	0.7245	0.7244	0.7279	0.7559
3	3	0.3842	0.3842	0.3839	0.3842	0.3896	0.4222

6.3.3 Procedure 3: Stop or forward one or more experimental treatments

In this procedure, we modify Procedure 2 by allowing more than one experimental arm to be forwarded to stage 2 if it is unclear which is best. If, after stage 1, more than one of the experimental treatments seems to be possibly better than the control and if their estimated effects do not differ by much, then it should be beneficial to forward all of these experimental arms in order to make a better selection after stage 2. We can describe the procedure as follows:

Stage 1: Allocate $\frac{N}{2(I+1)}$ patients to each arm. Carry forward the experimental treatment with the best results in stage 1 if $X_{1.1} > c_1$; also carry forward the treatment with the second best results if $X_{2.1} > c_2 \rho X_{1.1}$; .... We also carry forward the control. Stage 2: Allocate $\frac{N}{2(I_{.2}+1)}$ patients to each remaining arm. Test using T_1, T_2 or T_∞ (based on $I_{.2}$ treatments) with a critical value adjusted to take account of the bias in the test statistics caused by the adaptive design.

This procedure is similar to that of Kelly et al. (2005), except that we do not allow early rejection of H_0 and they considered only the test statistic T_∞.

If we stop early, then $T_k = 0$ for $k = 1, 2, \infty$ and no further calculation is needed. If we continue with one experimental arm, then the data from the two stages are combined in exactly the same way as for Procedure 1. If we continue with both experimental arms, then

$Z_i = (Z_{i.1} + Z_{i.2})/\sqrt{2}$, i.e. Z_i is calculated in the same way as from the fixed design. Note also, that when one arm is forwarded, $d_{i.2}$ is calculated in the same way as for Procedure 1, whereas when two arms are forwarded $d_{i.2} = d_{i.1}$.

Simulation and results

The null distributions of the test statistics were approximated using a new set of simulations. We chose not to reuse the earlier simulations because, depending on whether one or two arms is forwarded to stage 2, there are different numbers of patients on each arm. For stage 1, we simulated two million values of $Z_{1.1}$ and $Z_{2.1}$ from a bivariate normal distribution with correlation $\rho = 0.5$ (i.e. equal allocation), as for Procedure 1. It was noted whether the values of $Z_{1.1}$ and $Z_{2.1}$ led to stopping (Case 0), forwarding one arm (Case 1) or forwarding both arms (Case 2). Then, for stage 2, a loop was created so that for those simulations which led to Case 1 we simulated values of $Z_{i^*.2}$ from a univariate normal distribution, while for those simulations which led to Case 2, we simulated values of $Z_{1.2}$ and $Z_{2.2}$ from a bivariate normal distribution with correlation $\rho = 0.5$. The stages were combined to give two million simulated values of T_1, T_2 and T_∞, from which critical values were calculated at 5%, 2.5% and 1% levels of significance corresponding to the appropriate proportion being rejected. Again, we could have simulated bivariate normal random variables in each case and only used those which were needed. However, we chose not to, since the responses from treatment 1 at stage 2, for example, seem conceptually different if there are different numbers of subjects in different cases, so that reusing the same simulated values would create non-interpretable correlations among the simulation results.

On the basis of the results from Procedure 2, we decided to use only the values -0.5 and 0 for the stopping boundary c_1 and considered four different values of c_2, which determines whether we forward one or two arms to stage 2. The cutpoints, as shown in Table 6.14, indicate that changing c_2 has a clear impact, especially for T_1 and T_2. The difference between the two values of c_1 given is very small.

The probabilities of stopping, forwarding one treatment and forwarding two treatments under the null hypothesis are estimated in Table 6.15. Since the stopping rule is identical to that for Procedure 2, i.e. we stop if $X_{1.1} \leq c_1$, the expected sample sizes under H_0 are the same as those in Table 6.7, apart from simulation error. The probability of forwarding both arms, when H_0 is true, is always smaller than the probability of forwarding one arm, but varies considerably with c_2, where we forward two arms if $X_{2.1} > \frac{c_2}{2} X_{1.1}$.

Table 6.14: Cutpoints for T_1, T_2 and T_∞ for c_1 and c_2 at 5, 2.5 and 1%.

		5%			2.5%			1%		
c_1	c_2	T_1	T_2	T_∞	T_1	T_2	T_∞	T_1	T_2	T_∞
0	1.5	1.9467	1.8837	1.8690	2.2994	2.1985	2.1790	2.7311	2.5640	2.5381
0	1	2.0491	1.9254	1.8977	2.4257	2.2378	2.2034	2.8873	2.5959	2.5538
0	.75	2.0836	1.9367	1.9058	2.4635	2.2457	2.2076	2.9255	2.6065	2.5574
0	.5	2.1069	1.9441	1.9113	2.4905	2.2555	2.2149	2.9572	2.6165	2.5648
-0.5	1.5	1.9565	1.8939	1.8794	2.3029	2.2036	2.1845	2.7311	2.5643	2.5382
-0.5	1	2.0551	1.9321	1.9064	2.4286	2.2424	2.2086	2.8893	2.5993	2.5560
-0.5	.75	2.0893	1.9432	1.9127	2.4717	2.2524	2.2130	2.9377	2.6106	2.5607
-0.5	0.5	2.1146	1.9528	1.9203	2.4946	2.2583	2.2178	2.9576	2.6131	2.5645

Table 6.15: Percentages of Cases 0, 1 and 2 under H_0 for Procedure 3

		Case		
c_1	c_2	0	1	2
0	1.5	33.318	58.968	7.714
0	1	33.386	49.973	16.641
0	.75	33.269	45.439	21.292
0	.5	33.249	41.158	25.594
-.5	1.5	16.321	75.950	7.729
-.5	1	16.363	66.962	16.675
-.5	.75	16.367	62.435	21.198
-.5	.5	16.341	58.038	25.622

As before, the power and expected loss are calculated as the proportion of 100,000 simulations which rejected H_0 using different values of d_1 and d_2. Since different numbers of arms can be forwarded for different values of d_1 and d_2, depending on the results from stage 1, we chose to simulate a new set of errors for each value of d_1 and d_2. Hence, when comparing the estimated powers or expected losses from different values of d_1 and d_2 for this procedure, the variance is larger than for the other simulations presented in this thesis. However, the variance of each single power or expected loss is approximately the same as for any of the other procedures, since it is based on the same number of simulations.

The percentages of trials which stopped early, carried forward one treatment and carried forward two treatments under the alternative hypothesis are shown in Table 6.16. The percentages of trials which stop early are the same as with Procedure 2, the observed differences being due to simulation variance. On the other hand, we see that the number of arms carried forward is very sensitive to the value of d_2, as it should be. When $d_2 = 0$ it is rare to forward both arms, but when $d_2 = d_1$ it is very common to forward both arms. The value of c_2, which gives the decision about how many treatments to forward, is also quite influential, with the smaller values forwarding two treatments too often. From this table, a procedure with $c_1 = 0$ and $c_2 = 1$ or 1.5 looks quite attractive in that it forwards two arms rarely when $\Delta_2 = 0$, but often when $\Delta_2 = \Delta_1$.

Table 6.17 shows the power for all three tests for various values of d_1 when $d_2 = 0$. As we saw in Table 6.16, in this case it is quite rare to forward both arms, except for small c_2. As in the fixed design, T_∞ has the highest power for these configurations, T_2 is only slightly less powerful, but T_1 is worse. The effect of sometimes forwarding two treatments is a loss of power compared with Procedure 2 (cf. Tables 6.8-6.10). We see that $c_2 = 1.5$ always gives the highest power for these configurations and, of course, as $c_2 \to \infty$, this procedure converges to Procedure 2, i.e. we never forward both arms. These results are as expected, since when only one treatment is better than the control, it is clearly wasteful of resources to forward two treatments, since the likelihood of forwarding the inferior treatment if only one is selected is small whenever the power is reasonably high. The power is slightly lower with $c_1 = 0$ than with $c_1 = -0.5$, although of course the expected sample size is smaller.

Table 6.18 shows the power when $d_2 = 0.5d_1$. Again, as in the fixed design, T_2 gives the highest power for this configuration, with T_∞ being slightly less powerful and T_1 slightly less again. The power with $c_1 = -0.5$ is only very slightly higher than with $c_1 = 0$. The best value of c_2 is clearly 1.5 and, in this case, this procedure outperforms Procedure 2. In a few cases, it even has higher power than Procedure 1, despite using fewer patients.

Table 6.16: Percentages of Cases 0, 1 and 2 for Procedure 3 for different values of d_1 and d_2

			$d_2 = 0$			$d_2 = .5d_1$			$d_2 = d_1$		
			Case			Case			Case		
d_1	c_1	c_2	0	1	2	0	1	2	0	1	2
2	0	1.5	6.630	86.376	6.994	4.776	76.115	19.109	2.307	66.785	30.908
2	0	1	6.666	75.564	17.770	4.732	54.691	40.577	2.285	38.311	59.404
2	0	.75	6.678	68.673	24.649	4.699	44.391	50.910	2.320	27.552	70.128
2	0	.5	6.683	60.468	32.849	4.642	35.298	60.060	2.258	19.843	77.899
2	-.5	1.5	2.155	90.907	6.938	1.346	79.363	19.291	0.523	68.602	30.875
2	-.5	1	2.134	80.332	17.534	1.356	57.911	40.733	0.537	39.935	59.528
2	-.5	.75	2.083	72.931	24.986	1.292	47.968	50.740	0.535	29.454	70.011
2	-.5	.5	2.155	64.853	32.992	1.254	39.098	59.648	0.561	21.879	77.560
2.5	0	1.5	3.397	91.730	4.873	2.349	77.144	20.507	0.829	61.701	37.470
2.5	0	1	3.448	82.362	14.190	2.349	53.353	44.298	0.876	29.668	69.456
2.5	0	.75	3.471	75.134	21.395	2.267	41.494	56.239	0.857	19.218	79.925
2.5	0	.5	3.418	66.185	30.397	2.261	31.831	34.092	0.874	12.661	86.465
2.5	-.5	1.5	0.953	94.165	4.882	0.575	78.868	20.557	0.167	62.333	37.500
2.5	-.5	1	0.957	84.992	14.051	0.537	55.159	44.304	0.151	30.444	69.405
2.5	-.5	.75	0.933	77.699	21.368	0.578	43.556	55.866	0.151	19.959	79.890
2.5	-.5	.5	0.944	68.459	30.597	0.492	33.451	66.057	0.160	13.185	86.655
3	0	1.5	1.545	95.466	2.989	0.995	78.237	20.768	0.260	55.499	44.241
3	0	1	1.531	87.883	10.586	1.011	51.852	47.137	0.271	21.582	78.147
3	0	.75	1.596	80.701	17.339	0.960	38.956	60.084	0.293	12.643	87.064
3	0	.5	1.503	70.954	27.543	0.926	28.191	70.883	0.276	7.314	92.410
3	-.5	1.5	0.395	96.566	3.039	0.209	79.120	20.671	0.048	55.598	44.354
3	-.5	1	0.382	88.878	10.740	0.192	52.751	47.057	0.041	21.642	78.317
3	-.5	.75	0.368	81.609	18.023	0.193	39.686	60.121	0.039	12.718	87.243
3	-.5	.5	0.360	72.002	27.638	0.175	29.111	70.714	0.035	7.668	92.297
3.5	0	1.5	0.613	97.719	1.668	0.407	79.247	20.346	0.072	49.295	50.633
3.5	0	1	0.628	91.732	7.640	0.396	51.198	48.406	0.070	15.251	84.679
3.5	0	.75	0.606	85.212	14.182	0.380	36.723	62.897	0.084	7.729	92.187
3.5	0	.5	0.611	74.858	24.531	0.379	24.670	74.951	0.088	3.834	96.078
3.5	-.5	1.5	0.137	98.164	1.699	0.056	79.428	20.516	0.015	49.135	50.850
3.5	-.5	1	0.115	92.371	7.514	0.063	51.214	48.723	0.005	15.420	84.575
3.5	-.5	.75	0.121	85.634	14.245	0.068	36.852	63.080	0.002	7.705	92.293
3.5	-.5	.5	0.136	75.479	24.385	0.083	25.143	74.774	0.013	3.949	96.038

Table 6.17: Power of the tests using Procedure 3 with $\Delta_2 = 0$.

			5%			2.5%			1%		
d_1	c_1	c_2	T_1	T_2	T_∞	T_1	T_2	T_∞	T_1	T_2	T_∞
2.5	0	1.5	78.243	79.664	79.979	67.673	70.753	71.320	51.912	57.839	58.793
2.5	0	1	75.301	78.448	79.056	62.805	68.845	69.935	45.198	55.897	57.349
2.5	0	.75	73.993	77.849	78.524	61.107	68.239	69.373	43.084	54.959	56.749
2.5	0	.5	72.869	77.346	78.108	59.305	67.223	68.451	40.996	53.611	55.466
2.5	-.5	1.5	78.954	80.460	80.784	68.136	71.218	71.798	52.133	58.309	59.230
2.5	-.5	1	75.875	79.053	79.650	63.179	69.298	70.320	45.397	56.252	57.737
2.5	-.5	.75	74.394	78.412	79.177	60.991	68.391	69.645	42.972	54.913	56.670
2.5	-.5	.5	73.343	77.949	78.796	59.724	67.839	69.089	41.268	54.186	55.948
3	0	1.5	90.863	91.639	91.816	84.369	86.535	86.92	72.668	77.550	78.248
3	0	1	89.166	91.003	91.336	81.175	85.409	86.104	66.912	76.206	77.331
3	0	.75	88.315	90.545	90.969	79.598	84.830	85.660	64.650	75.073	76.516
3	0	.5	87.600	90.335	90.817	78.405	84.258	85.143	62.703	74.204	75.821
3	-.5	1.5	91.229	92.066	92.238	84.733	86.828	87.157	72.814	77.903	78.618
3	-.5	1	89.289	91.168	91.536	81.096	85.446	86.157	66.494	75.927	77.121
3	-.5	.75	88.609	90.954	91.391	79.591	84.940	85.819	64.212	75.010	76.462
3	-.5	.5	87.835	90.620	91.118	78.530	84.546	85.382	62.693	74.590	76.065
3.5	0	1.5	96.843	97.167	97.251	94.094	95.043	95.199	87.570	90.576	90.965
3.5	0	1	96.320	97.030	97.184	92.454	94.681	95.009	84.028	89.828	90.443
3.5	0	.75	95.857	96.846	97.031	91.627	94.338	94.717	82.266	89.221	90.036
3.5	0	.5	95.513	96.697	96.892	90.894	93.987	94.430	80.850	88.614	89.525
3.5	-.5	1.5	97.131	97.450	97.525	94.231	95.194	95.375	87.842	90.777	91.145
3.5	-.5	1	96.503	97.317	97.469	92.592	94.761	95.094	84.138	89.955	90.671
3.5	-.5	.75	96.207	97.177	97.361	91.898	94.630	95.011	82.253	89.498	90.356
3.5	-.5	.5	95.671	96.916	97.119	90.977	94.078	94.542	80.881	88.768	89.676

Table 6.18: Power of the tests using Procedure 3 when $\Delta_2 = .5\Delta_1$.

			5%			2.5%			1%		
d_1	c_1	c_2	T_1	T_2	T_∞	T_1	T_2	T_∞	T_1	T_2	T_∞
2.5	0	1.5	78.585	79.062	78.957	68.362	69.672	69.544	53.544	56.644	56.663
2.5	0	1	77.733	78.583	78.353	66.772	68.827	68.500	50.650	55.446	55.257
2.5	0	.75	77.008	78.039	77.886	65.392	67.943	67.669	49.108	53.930	53.727
2.5	0	.5	76.164	77.295	77.048	64.294	66.946	66.670	48.058	52.798	52.707
2.5	-.5	1.5	78.973	79.437	79.374	68.950	70.206	70.091	53.930	57.116	57.017
2.5	-.5	1	77.904	78.711	78.479	66.704	68.781	68.561	50.710	55.357	55.182
2.5	-.5	.75	77.361	78.365	78.171	65.336	68.161	68.028	48.967	54.011	54.039
2.5	-.5	.5	76.578	77.698	77.490	64.515	67.249	67.065	48.091	53.211	52.997
3	0	1.5	90.371	90.660	90.624	84.140	85.036	84.963	73.095	75.843	75.781
3	0	1	89.722	90.165	90.064	82.526	84.156	84.017	69.837	74.358	74.103
3	0	.75	89.116	89.749	89.625	81.435	83.451	83.231	67.998	72.950	72.823
3	0	.5	88.793	89.543	89.363	80.688	82.831	82.629	66.799	72.059	71.954
3	-.5	1.5	90.556	90.827	90.749	84.414	85.290	85.156	73.448	76.077	76.064
3	-.5	1	89.976	90.482	90.314	82.458	84.227	84.119	69.652	74.073	73.940
3	-.5	.75	89.703	90.387	90.246	81.836	84.015	83.752	68.457	73.481	73.301
3	-.5	.5	88.800	89.623	89.457	80.476	82.871	82.648	66.784	72.070	71.873
3.5	0	1.5	96.471	96.567	96.535	93.457	93.981	93.899	87.288	88.920	88.762
3.5	0	1	96.158	96.374	96.321	92.499	93.393	93.243	84.806	87.748	87.552
3.5	0	.75	95.967	96.261	96.150	92.036	93.117	93.016	83.661	87.124	87.030
3.5	0	.5	95.604	96.024	95.941	91.233	92.611	92.440	82.018	85.984	85.935
3.5	-.5	1.5	96.380	96.501	96.468	93.428	93.893	93.798	87.159	88.815	88.729
3.5	-.5	1	96.190	96.432	96.346	92.520	93.362	93.196	84.715	87.652	87.479
3.5	-.5	.75	95.992	96.328	96.250	91.882	93.133	92.989	83.092	86.929	86.799
3.5	-.5	.5	95.711	96.074	95.960	91.302	92.761	92.643	82.097	86.284	86.123

It is interesting to note that the powers in this table are often lower than when $d_2 = 0$, which was never the case with the fixed design. There are probably two contributing factors. First, as can be seen from Table 6.16, when $d_2 = 0.5d_1$, both experimental arms are forwarded to stage 2 considerably more often, and this reduces the number of patients used for comparing treatment 1 with the control. Second, if one treatment is forwarded, it is more likely to be the inferior treatment when $d_2 = 0.5d_1$ than when $d_2 = 0$; when this happens it is likely that we will fail to reject H_0.

Table 6.19 shows the power when $d_2 = d_1$. As in the fixed design, T_∞ has the least power in this case. In the fixed design, T_1 was clearly more powerful than T_2, but the difference seems to be very small here. In some cases at 1%, T_2 is more powerful, though this might be due only to simulation variation. However, there seems to be no practical

Table 6.19: Power of the tests using Procedure 3 when $\Delta_2 = \Delta_1$

			5%			2.5%			1%		
d_1	c_1	c_2	T_1	T_2	T_∞	T_1	T_2	T_∞	T_1	T_2	T_∞
2	0	1.5	74.929	74.791	74.319	63.479	63.672	62.992	48.245	48.992	48.006
2	0	1	74.621	73.885	73.032	63.445	62.567	61.083	48.850	48.161	46.085
2	0	.75	74.579	73.302	72.147	63.559	62.078	60.259	49.180	47.469	45.274
2	0	.5	74.663	73.045	71.762	63.454	61.532	59.645	48.883	46.672	44.334
2	-.5	1.5	74.901	74.820	74.404	63.497	63.621	62.829	48.421	49.152	48.040
2	-.5	1	75.047	74.127	73.131	63.678	62.827	61.403	48.853	48.172	46.207
2	-.5	.75	74.759	73.572	72.428	63.515	62.103	60.462	48.885	47.436	45.316
2	-.5	.5	74.728	73.327	72.058	63.691	61.707	59.794	48.904	47.016	44.458
2.5	0	1.5	89.364	89.215	88.833	82.168	82.179	81.508	70.247	70.941	69.672
2.5	0	1	89.426	88.622	87.939	82.369	81.398	79.893	70.981	69.992	67.309
2.5	0	.75	89.264	88.413	87.526	82.395	80.996	79.328	71.236	69.335	66.738
2.5	0	.5	89.172	88.075	87.147	82.236	80.516	78.695	70.988	68.591	65.764
2.5	-.5	1.5	89.493	89.349	88.984	82.441	82.389	81.685	70.483	71.273	70.067
2.5	-.5	1	89.406	88.673	87.848	82.381	81.325	79.920	70.756	69.809	67.475
2.5	-.5	.75	89.492	88.593	87.642	82.361	80.992	79.329	70.910	69.201	66.661
2.5	-.5	.5	89.379	88.237	87.240	82.221	80.632	78.708	71.035	68.715	65.732
3	0	1.5	96.655	96.635	96.430	93.540	93.451	92.941	86.908	87.273	86.213
3	0	1	96.647	96.306	95.887	93.573	92.971	92.032	87.334	86.476	84.363
3	0	.75	96.492	95.994	95.501	93.233	92.374	91.278	86.956	85.558	83.388
3	0	.5	96.472	95.938	95.367	93.366	92.242	91.046	87.012	85.203	82.875
3	-.5	1.5	96.690	96.608	96.385	93.514	93.407	92.918	86.839	87.236	86.106
3	-.5	1	96.760	96.379	95.905	93.621	92.891	91.940	87.291	86.485	84.464
3	-.5	.75	96.658	96.180	95.646	93.516	92.509	91.457	87.136	85.765	83.626
3	-.5	.5	96.658	96.111	95.526	93.451	92.417	91.222	87.020	85.430	83.097

difference in power between the two tests. In this case, $c_1 = 0$ can even give higher power than $c_1 = -0.5$, even though it leads to more early stopping, although this might be just due to simulation variance. Again $c_2 = 1.5$ seems to be the best value and again this procedure is better than Procedure 2, although only very slightly. Again, it also gives a more powerful test than Procedure 1, despite using fewer patients. As expected, the powers when $d_2 = d_1$ are considerably higher than when $d_2 = 0.5d_1$, because whichever arm or arms are forwarded to stage 2, we are still likely to reject H_0.

Tables 6.20-6.22 show the expected loss when $d_2 = 0$, $d_2 = d_1/2$ and $d_2 = d_1$ respectively. The results are broadly consistent with those for the power. T_∞ is best when $d_2 = 0$, but T_2 is better in the other cases, being very similar to T_1 when $d_2 = d_1$. As with the power, there is very little difference between the values of c_1, with $c_1 = 0$ even giving smaller expected loss than $c_1 = -0.5$ when d_1 and d_2 are both large. A value of 1.5 for c_2 is clearly the best of those studied. When $d_2 = 0$, the expected loss is slightly higher than for Procedure 2, since we sometimes forward two experimental treatments, but end up only using one to calculate the test statistic, and so we lose power due to the wasted patients on the second treatment. In the other two configurations it is somewhat lower than in Procedure 2, because we more often forward the best treatment and so reduce the probability of a type-III/IV error. In many cases, including all of those with $\Delta_2 = \Delta_1/2$, the expected loss is even clearly better than for Procedure 1, which on average uses more patients.

The flexibility of Procedure 3 in allowing a more powerful test when it is clear which experimental treatment is better after stage 1, but allowing better optimal treatment selection when it is not clear which is best, seems to lead to fewer type-III/IV errors and hence to a lower expected losses.

6.3.4 Discussion

To summarise the results from this section, a powerful design from each procedure, including the fixed design, is compared in Table 6.23, using the test statistic T_2 at the 5% level of significance. A design which always drops one treatment after stage 1 (Procedure 1) shows improved performance over the fixed design with the same sample size, due to the extra information gained on the important pairwise comparison. A design which additionally allows early stopping for futility (Procedure 2) leads to some savings in expected sample size, up to about $N/6$, without much loss of performance. A design which additionally allows two arms to be forwarded if it is unclear which is best (Procedure 3), although the most complex to manage, allows identical savings in terms of expected numbers of patients to Procedure 2, while achieving power and expected loss, which are as good as Procedure 1.

Overall, the message from the simulations reported in this chapter is very clear. When feasible, an adaptive design should be used in preference to the fixed design. The best adaptive design studied seems to be Procedure 3, with $c_2 = 1.5$ and $c_1 = 0$. This design saves, on average, at least 1/6 of the patients if H_0 is true, but, if H_1 is true, achieves a similar power to Procedure 1, which has a fixed sample size and which, in turn, has higher power than the fixed design. Note also that the increase in power over the fixed design can be as much as 6%, whereas the optimal allocation for the fixed design found in Chapter 4 increased the power over equal allocation by no more than 3%.

Table 6.20: Expected loss of the tests using Procedure 3 when $\Delta_2 = 0$

			5%			2.5%			1%		
d_1	c_1	c_2	T_1	T_2	T_∞	T_1	T_2	T_∞	T_1	T_2	T_∞
2.5	0	1.5	0.5501	0.5145	0.5065	0.8118	0.7347	0.7204	1.2048	1.0557	1.0315
2.5	0	1	0.6230	0.5440	0.5283	0.9340	0.7820	0.7545	1.3730	1.1047	1.0680
2.5	0	.75	0.6557	0.5584	0.5410	0.9766	0.7971	0.7684	1.4257	1.1279	1.0829
2.5	0	.5	0.6839	0.5712	0.5514	1.0217	0.8222	0.7912	1.4775	1.1612	1.1145
2.5	-.5	1.5	0.5329	0.4951	0.4867	0.8010	0.7235	0.7089	1.1993	1.0445	1.0212
2.5	-.5	1	0.6087	0.5284	0.5130	0.9244	0.7705	0.7446	1.3677	1.0954	1.0579
2.5	-.5	.75	0.6463	0.5452	0.5256	0.9801	0.7940	0.7620	1.4290	1.1295	1.0850
2.5	-.5	.5	0.6732	0.5566	0.5349	1.0119	0.8077	0.7758	1.4714	1.1473	1.1025
3	0	1.5	0.2767	0.2533	0.2480	0.4706	0.4056	0.3940	0.8208	0.6745	0.6535
3	0	1	0.3297	0.2744	0.2634	0.5668	0.4436	0.4222	0.9961	0.7167	0.6800
3	0	.75	0.3526	0.2855	0.2727	0.6138	0.4565	0.4314	1.0618	0.7487	0.7051
3	0	.5	0.3739	0.2915	0.2769	0.6493	0.4733	0.4466	1.1199	0.7745	0.7259
3	-.5	1.5	0.2659	0.2410	0.2358	0.4597	0.3968	0.3869	0.8166	0.6638	0.6422
3	-.5	1	0.3230	0.2666	0.2555	0.5684	0.4376	0.4162	1.0060	0.7228	0.6868
3	-.5	.75	0.3440	0.2735	0.2601	0.6141	0.4532	0.4266	1.0750	0.7506	0.7069
3	-.5	.5	0.3671	0.2834	0.2683	0.6458	0.4651	0.4398	1.1207	0.7633	0.7187
3.5	0	1.5	0.1111	0.0998	0.0969	0.2071	0.1740	0.1685	0.4353	0.3301	0.3164
3.5	0	1	0.1294	0.1046	0.0991	0.2646	0.1866	0.1751	0.5594	0.3562	0.3347
3.5	0	.75	0.1454	0.1108	0.1043	0.2934	0.1985	0.1852	0.6210	0.3774	0.3489
3.5	0	.5	0.1575	0.1160	0.1091	0.3191	0.2108	0.1952	0.6705	0.3987	0.3669
3.5	-.5	1.5	0.1015	0.0903	0.0877	0.2025	0.1689	0.1626	0.4259	0.3232	0.3102
3.5	-.5	1	0.1231	0.0946	0.0892	0.2597	0.1839	0.1722	0.5555	0.3518	0.3267
3.5	-.5	.75	0.1334	0.0994	0.0930	0.2841	0.1884	0.1750	0.6215	0.3678	0.3377
3.5	-.5	.5	0.1519	0.1084	0.1013	0.3161	0.2077	0.1914	0.6695	0.3933	0.3616

Table 6.21: Expected loss of the tests using Procedure 3 when $\Delta_2 = .5\Delta_1$

			5%			2.5%			1%		
d_1	c_1	c_2	T_1	T_2	T_∞	T_1	T_2	T_∞	T_1	T_2	T_∞
2.5	0	1.5	0.6296	0.6172	0.6188	0.8638	0.8304	0.8317	1.2122	1.1328	1.1295
2.5	0	1	0.6476	0.6223	0.6252	0.9060	0.8477	0.8514	1.2899	1.1629	1.1628
2.5	0	.75	0.6639	0.6331	0.6338	0.9401	0.8691	0.8708	1.3296	1.1993	1.1985
2.5	0	.5	0.6856	0.6508	0.6535	0.9677	0.8931	0.8949	1.3561	1.2273	1.2239
2.5	-.5	1.5	0.6194	0.6075	0.6084	0.8492	0.8170	0.8182	1.2024	1.1215	1.1214
2.5	-.5	1	0.6417	0.6181	0.6214	0.9058	0.8486	0.8500	1.2880	1.1642	1.1630
2.5	-.5	.75	0.6542	0.6238	0.6255	0.9402	0.8621	0.8606	1.3319	1.1965	1.1902
2.5	-.5	.5	0.6757	0.6411	0.6425	0.9622	0.8856	0.8852	1.3557	1.2179	1.2172
3	0	1.5	0.3841	0.3762	0.3767	0.5542	0.5279	0.5289	0.8660	0.7836	0.7828
3	0	1	0.3922	0.3774	0.3787	0.5974	0.5454	0.5467	0.9649	0.8238	0.8263
3	0	.75	0.4081	0.3858	0.3869	0.6293	0.5637	0.5659	1.0202	0.8631	0.8616
3	0	.5	0.4196	0.3931	0.3955	0.6542	0.5831	0.5844	1.0583	0.8913	0.8889
3	-.5	1.5	0.3745	0.3671	0.3686	0.5423	0.5164	0.5191	0.8528	0.7733	0.7715
3	-.5	1	0.3827	0.3657	0.3687	0.5976	0.5412	0.5414	0.9690	0.8305	0.8296
3	-.5	.75	0.3905	0.3671	0.3694	0.6173	0.5468	0.5508	1.0068	0.8484	0.8487
3	-.5	.5	0.4185	0.3901	0.3925	0.6591	0.5813	0.5841	1.0578	0.8918	0.8919
3.5	0	1.5	0.2061	0.2034	0.2043	0.3007	0.2834	0.2858	0.5027	0.4469	0.4503
3.5	0	1	0.2011	0.1933	0.1944	0.3235	0.2911	0.2943	0.5851	0.4789	0.4823
3.5	0	.75	0.2044	0.1929	0.1956	0.3378	0.2970	0.2980	0.6243	0.4984	0.4977
3.5	0	.5	0.2193	0.2027	0.2038	0.3677	0.3156	0.3186	0.6831	0.5391	0.5369
3.5	-.5	1.5	0.2098	0.2062	0.2072	0.3019	0.2871	0.2895	0.5062	0.4498	0.4513
3.5	-.5	1	0.1995	0.1905	0.1925	0.3220	0.2907	0.2947	0.5874	0.4813	0.4832
3.5	-.5	.75	0.2040	0.1909	0.1925	0.3427	0.2965	0.2992	0.6436	0.5054	0.5059
3.5	-.5	.5	0.2129	0.1984	0.2008	0.3630	0.3087	0.3100	0.6787	0.5274	0.5288

Table 6.22: Expected loss of the tests using Procedure 3 when $\Delta_2 = \Delta_1$

			5%			2.5%			1%		
d_1	c_1	c_2	T_1	T_2	T_∞	T_1	T_2	T_∞	T_1	T_2	T_∞
2	0	1.5	0.5014	0.5042	0.5136	0.7304	0.7266	0.7402	1.0351	1.0202	1.0399
2	0	1	0.5076	0.5223	0.5394	0.7311	0.7487	0.7783	1.0230	1.0368	1.0783
2	0	.75	0.5084	0.5340	0.5571	0.7288	0.7584	0.7948	1.0164	1.0506	1.0945
2	0	.5	0.5067	0.5391	0.5648	0.7309	0.7694	0.8071	1.0223	1.0666	1.1133
2	-.5	1.5	0.5020	0.5036	0.5119	0.7301	0.7276	0.7434	1.0316	1.0170	1.0392
2	-.5	1	0.4991	0.5175	0.5374	0.7264	0.7435	0.7719	1.0229	1.0366	1.0759
2	-.5	.75	0.5048	0.5286	0.5514	0.7297	0.7579	0.7908	1.0223	1.0513	1.0937
2	-.5	.5	0.5054	0.5335	0.5588	0.7262	0.7659	0.8041	1.0219	1.0597	1.1108
2.5	0	1.5	0.2659	0.2696	0.2792	0.4458	0.4455	0.4623	0.7438	0.7265	0.7582
2.5	0	1	0.2644	0.2845	0.3015	0.4408	0.4651	0.5027	0.7255	0.7502	0.8173
2.5	0	.75	0.2684	0.2897	0.3119	0.4401	0.4751	0.5168	0.7191	0.7666	0.8316
2.5	0	.5	0.2707	0.2981	0.3213	0.4441	0.4871	0.5326	0.7253	0.7852	0.8559
2.5	-.5	1.5	0.2627	0.2663	0.2754	0.4390	0.4403	0.4579	0.7379	0.7182	0.7483
2.5	-.5	1	0.2649	0.2832	0.3038	0.4405	0.4669	0.5020	0.7311	0.7548	0.8131
2.5	-.5	.75	0.2627	0.2852	0.3090	0.4410	0.4752	0.5168	0.7273	0.7700	0.8335
2.5	-.5	.5	0.2655	0.2941	0.3190	0.4445	0.4842	0.5323	0.7241	0.7821	0.8567
3	0	1.5	0.1004	0.1010	0.1071	0.1938	0.1965	0.2118	0.3928	0.3818	0.4136
3	0	1	0.1006	0.1108	0.1234	0.1928	0.2109	0.2390	0.3800	0.4057	0.4691
3	0	.75	0.1052	0.1202	0.1350	0.2030	0.2288	0.2617	0.3913	0.4333	0.4984
3	0	.5	0.1058	0.1219	0.1390	0.1990	0.2327	0.2686	0.3896	0.4439	0.5138
3	-.5	1.5	0.0993	0.1018	0.1085	0.1946	0.1978	0.2125	0.3948	0.3829	0.4168
3	-.5	1	0.0972	0.1086	0.1229	0.1914	0.2133	0.2418	0.3813	0.4055	0.4661
3	-.5	.75	0.1003	0.1146	0.1306	0.1945	0.2247	0.2563	0.3859	0.4271	0.4912
3	-.5	.5	0.1003	0.1167	0.1342	0.1965	0.2275	0.2633	0.3894	0.4371	0.5071

Table 6.23: Summary of properties of different procedures using T_2 at the 5% level of significance.

	Procedure			
	Fixed	1	2 ($c_1 = 0$)	3 ($c_1 = 0, c_2 = 1.5$)
Expected no. patients ×100/N	100	100	83.33	83.33
Power ($d_1 = 3, d_2 = 0$)	85.477	92.328	91.762	91.639
Power ($d_1 = 3, d_2 = 1.5$)	87.579	90.229	90.004	90.660
Power ($d_1 = 2.5, d_2 = 2.5$)	87.151	89.215	89.121	89.215
$E(Loss)$ ($d_1 = 3, d_2 = 0$)	0.4374	0.2361	0.2533	0.2533
$E(Loss)$ ($d_1 = 3, d_2 = 1.5$)	0.4490	0.4306	0.4380	0.3762
$E(Loss)$ ($d_1 = 2.5, d_2 = 2.5$)	0.3212	0.2696	0.2720	0.2696

Further refinement of these values could be used to find an optimum design for specific situations, but the criterion of optimality would have to balance costs against benefits. Because the main benefit of Procedure 3 is its ability to forward a different number of treatments depending on the results from stage 1, the benefits will be even greater if more arms are used. For example, with a large number of experimental treatments, this design will forward a larger or smaller number depending on the results of stage 1, thus allowing resources to be concentrated on comparing a few treatments against the control if only a few are beneficial, or using the information from many treatments if many are beneficial.

6.4 Procedure 3 with unequal stage sizes

So far we have considered only stages of equal sizes. For any of the procedures, it might be better to use fewer patients in stage 1, but this is particularly true for Procedure 3, since it has the flexibility to drop a clearly inferior arm, but continue with all arms which remain good candidates. There is little advice in the literature on how to choose the stage sizes in adaptive multi-arm trials, though Stallard and Todd (2003) showed that in a four-arm trial with their procedure, it was optimal to use slightly fewer than $N/5$ patients in stage 1. Here we consider Procedure 3 with a total of $N/4$ patients in stage 1 and $3N/4$ patients in stage 2, keeping balanced allocation within stages. Again, this is just a preliminary study and no attempt is made to optimise the size of the stages.

The possible advantage of this method is that, when a clearly inferior treatment is dropped after stage 1, only $N/12$ patients have been used for it, whereas with equal stage sizes $N/6$

patients would have been used. As well as possible ethical benefits, this should lead to more efficient comparisons of the best treatment with the control. When both treatments are good and we continue to use them in stage 2, the total numbers allocated to each treatment are the same in this design as in the design with equal stage sizes. A possible disadvantage of this procedure is that, with less information available at stage 1, a bad decision, either to drop a promising treatment, or to continue with a poor treatment, is more likely to be made.

As with equal stage sizes, we need to combine the univariate test statistics from the two stages in the appropriate way. As before, this depends on whether one (Case 1) or two (Case 2) treatments are forwarded to the second stage. From stage 1, $\sigma_{.1}^2 = 6$, so that

$$\begin{aligned} Z_{i.1} &= \sqrt{\frac{N}{4}}\left(\bar{Y}_{i.1} - \bar{Y}_{0.1}\right)/\sqrt{6} \\ \Rightarrow \bar{Y}_{i.1} - \bar{Y}_{0.1} &= \frac{Z_{i.1}\sqrt{24}}{\sqrt{N}}. \end{aligned}$$

In Case 1, there are only two treatments at stage 2, so that $\sigma_{.2}^2 = 4$ and

$$\begin{aligned} Z_{i.2} &= \sqrt{\frac{3N}{4}}\left(\bar{Y}_{i.2} - \bar{Y}_{0.2}\right)/\sqrt{4} \\ \Rightarrow \bar{Y}_{i.2} - \bar{Y}_{0.2} &= \frac{Z_{i.2}\sqrt{16}}{\sqrt{3N}}. \end{aligned}$$

Combining stages, we have

$$\bar{Y}_i = \frac{2\bar{Y}_{i.1} + 9\bar{Y}_{i.2}}{11}$$

and

$$\bar{Y}_0 = \frac{2\bar{Y}_{0.1} + 9\bar{Y}_{0.2}}{11},$$

so that, writing $Z_i = k_1\left(\bar{Y}_i - \bar{Y}_0\right)$, for a constant k_1 chosen to ensure unit variance, and rewriting in terms of $Z_{i.k}$, we obtain

$$Z_i = \frac{1}{\sqrt{11}}\left(\sqrt{2}Z_{i.1} + 3Z_{i.2}\right).$$

In Case 2, there are three treatments at stage 2, so that $\sigma_{.2}^2 = 6$ and

$$\begin{aligned} Z_{i.2} &= \sqrt{\frac{3N}{4}}\left(\bar{Y}_{i.2} - \bar{Y}_{0.2}\right)/\sqrt{6} \\ \Rightarrow \bar{Y}_{i.2} - \bar{Y}_{0.2} &= \frac{Z_{i.2}\sqrt{8}}{\sqrt{N}}. \end{aligned}$$

Combining stages, we have

$$\bar{Y}_i = \frac{\bar{Y}_{i.1} + 3\bar{Y}_{i.2}}{4}$$

and

$$\bar{Y}_0 = \frac{\bar{Y}_{0.1} + 3\bar{Y}_{0.2}}{4},$$

Table 6.24: Cutpoints for T_1, T_2 and T_∞ for Procedure 3 with unequal stage sizes.

		5%			2.5%			1%		
c_1	c_2	T_1	T_2	T_∞	T_1	T_2	T_∞	T_1	T_2	T_∞
0	1.5	1.8478	1.8048	1.7942	2.2009	2.1315	2.1173	2.6206	2.5022	2.4829
0	1	1.9418	1.8523	1.8324	2.3199	2.1786	2.1515	2.7775	2.5516	2.5173
-0.5	1.5	1.8714	1.8303	1.8214	2.2157	2.1481	2.1347	2.6316	2.5191	2.5007
-0.5	1	1.9641	1.8778	1.8587	2.3305	2.1939	2.1688	2.7790	2.5577	2.5245

so that, writing $Z_i = k_2\left(\bar{Y}_i - \bar{Y}_0\right)$, for a constant k_2 chosen to ensure unit variance, and rewriting in terms of $Z_{i.k}$, we obtain

$$Z_i = \frac{1}{2}\left(Z_{i.1} + \sqrt{3}Z_{i.2}\right).$$

The relationships between $d_{i.k}$ and d_i are also worked out in the same way as before. We obtain $d_{i.1} = d_i/2$. In Case 1, $d_{i.2} = 3d_i/2\sqrt{2}$ and, in Case 2, $d_{i.2} = \sqrt{3}d_i/2$.

6.4.1 Simulations and results

The simulations were carried out in the same way as in Section 6.3.3, but only for the stopping rules $c_1 = 0, -0.5$ and for the rules to carry forward two arms $c_2 = 1.5, 1$. Note, however, that these values are not directly comparable with the values of c_1 and c_2 for equal stage sizes, since scaling of $Z_{i.1}$ is different. The percentages of trials stopping early (Case 0), carrying forward one arm (Case 1) and carrying forward two arms (Case 2) are the same as those shown in Table 6.15, except for simulation variance. However, because the first stage uses fewer patients, the expected sample sizes are smaller, being $0.75N$ when $c_1 = 0$ and approximately $0.163N/4 + 0.837N = 0.878N$ when $c_1 = -0.5$. Therefore, in terms of the expected sample size under H_0, the results here with $c_1 = -0.5$ are roughly comparable with those for $c_1 = 0$ when the stage sizes are equal.

The cutpoints for rejecting H_0 using T_1, T_2 and T_∞ are shown in Table 6.24. These are smaller than the corresponding results with equal stage sizes, indicating that the tails of the null distributions are less extreme. This might be expected to lead to more powerful tests.

The simulations under H_1 were carried out, as before, with new errors simulated for different values of d_1, d_2, c_1 and c_2. Table 6.25 shows the percentages stopping early and forwarding one or two arms to stage 2 under H_1. The trial stops early here more often

Table 6.25: Percentages of Cases 0, 1 and 2 for Procedure 3 with unequal stage sizes for different values of d_1 and d_2

			$d_2 = 0$			$d_2 = .5d_1$			$d_2 = d_1$		
			Case			Case			Case		
d_1	c_1	c_2	0	1	2	0	1	2	0	1	2
2	0	1.5	12.690	78.283	9.027	9.640	73.707	16.653	6.220	70.633	23.147
2	0	1	12.727	66.608	20.665	9.831	55.317	34.852	6.136	47.324	46.540
2	-.5	1.5	4.676	86.312	9.012	3.226	80.378	16.396	1.774	75.049	23.177
2	-.5	1	4.744	74.677	20.579	3.224	61.844	34.932	1.789	52.134	46.077
2.5	0	1.5	8.669	83.327	8.004	6.398	75.114	18.488	3.456	68.774	27.770
2.5	0	1	8.779	72.048	19.173	6.251	55.195	38.554	3.446	41.922	54.632
2.5	-.5	1.5	2.988	89.142	7.870	1.858	79.848	18.294	0.912	71.244	27.844
2.5	-.5	1	2.986	77.950	19.082	1.894	59.500	38.606	0.892	44.767	54.341
3	0	1.5	5.717	87.830	6.453	4.045	76.157	19.798	1.821	65.396	32.783
3	0	1	5.769	77.295	16.945	4.020	54.235	41.745	1.840	36.074	62.086
3	-.5	1.5	1.749	91.753	6.498	1.018	79.377	19.605	0.415	66.998	32.587
3	-.5	1	1.775	81.580	16.645	1.064	57.049	41.887	0.384	37.647	61.969
3.5	0	1.5	3.550	91.499	4.951	2.388	77.108	20.504	0.882	62.020	37.098
3.5	0	1	3.511	82.043	14.446	2.296	53.322	44.382	0.895	29.975	69.130
3.5	-.5	1.5	0.965	93.934	5.101	0.585	78.912	20.503	0.185	62.279	37.536
3.5	-.5	1	0.976	84.714	14.310	0.530	55.057	44.413	0.154	30.818	69.028

than with equal stage sizes, due to the high variance in stage 1. When $d_2 = 0$ this design seems to forward both experimental arms to the second stage more often than the design with equal stage sizes, since with higher variance it is more difficult to decide which is best. When $d_2 = 0.5d_1$ and when $d_2 = d_1$, it seems to forward one arm more than the other case. All of these features are undesirable and might counteract the narrower tails in the null distribution to make this design less powerful.

The powers for various configurations are shown in Tables 6.26-6.28 and the corresponding expected losses are in Tables 6.29-6.31. The first thing to note is that it is much less clear than with equal allocation which value of c_2 is best. When Δ_1 is not so large, $c_2 = 1.5$ gives higher power and lower expected loss, but as Δ_1 increases, $c_2 = 1$ becomes better. Overall, perhaps we would recommend $c_2 = 1$, as the difference can become quite large, e.g. more than 1% difference in power for $d_1 = 3.5$ and $d_2 = 1.75$.

As noted earlier, the design described here with $c_1 = -0.5$ is roughly equivalent, in terms of expected sample size under H_0, to the design described earlier with $c_1 = 0$. Comparing

Table 6.26: Power of the tests using Procedure 3 with unequal stage sizes when $\Delta_2 = 0$.

			5%			2.5%			1%		
d_1	c_1	c_2	T_1	T_2	T_∞	T_1	T_2	T_∞	T_1	T_2	T_∞
2	0	1.5	58.224	59.144	59.366	47.655	49.519	49.862	34.136	37.515	38.057
2	0	1	56.314	58.391	58.826	44.409	48.237	48.935	29.166	35.711	36.706
2	-0.5	1.5	60.544	61.545	61.757	49.266	51.329	51.715	34.937	38.289	38.850
2	-0.5	1	58.524	60.783	61.224	45.792	49.977	50.708	30.032	36.785	37.795
2.5	0	1.5	75.760	76.408	76.561	68.021	69.551	69.848	55.759	59.383	59.949
2.5	0	1	74.676	76.259	76.601	64.940	68.646	69.308	49.812	57.203	58.187
2.5	-0.5	1.5	78.498	79.225	79.397	70.155	71.877	72.204	57.057	60.586	61.178
2.5	-0.5	1	77.041	78.773	79.144	66.860	70.622	71.324	51.456	58.944	59.987
3	0	1.5	87.204	87.518	87.604	83.027	83.952	84.141	74.839	77.407	77.782
3	0	1	86.836	87.706	87.870	81.392	83.663	84.054	70.653	76.380	77.143
3	-0.5	1.5	89.895	90.267	90.336	85.421	86.346	86.548	76.829	79.451	79.837
3	-0.5	1	89.580	90.484	90.707	83.762	86.152	86.566	72.529	78.413	79.194
3.5	0	1.5	93.518	93.659	93.691	91.942	92.309	92.399	88.329	89.540	89.731
3.5	0	1.	93.903	94.209	94.278	91.615	92.633	92.806	85.970	89.219	89.597
3.5	-0.5	1.5	95.762	95.904	95.917	94.043	94.455	94.528	90.113	91.351	91.559
3.5	-0.5	1	95.829	96.206	96.289	93.311	94.460	94.641	87.359	90.782	91.224

Table 6.27: Power of the tests using Procedure 3 with unequal stage sizes when $\Delta_2 = .5\Delta_1$.

			5%			2.5%			1%		
d_1	c_1	c_2	T_1	T_2	T_∞	T_1	T_2	T_∞	T_1	T_2	T_∞
2	0	1.5	59.111	59.459	59.433	48.238	49.000	48.945	34.900	36.633	36.607
2	0	1	58.810	59.395	59.254	47.226	48.560	48.393	32.790	35.648	35.534
2	-0.5	1.5	60.607	60.933	60.850	49.429	50.317	50.263	35.465	37.177	37.197
2	-0.5	1	60.041	60.679	60.599	48.018	49.548	49.349	33.297	36.317	36.288
2.5	0	1.5	74.965	75.283	75.224	66.522	67.253	67.208	54.503	56.230	56.237
2.5	0	1	75.388	75.889	75.715	66.034	67.413	67.271	52.107	55.616	55.461
2.5	-0.5	1.5	76.802	77.105	77.033	68.079	68.807	68.772	55.337	57.143	57.056
2.5	-0.5	1	77.263	77.778	77.676	67.673	69.082	68.821	53.195	56.777	56.716
3	0	1.5	85.678	85.843	85.828	80.542	81.057	81.003	71.960	73.336	73.233
3	0	1	86.818	87.138	87.063	81.084	82.072	81.921	70.537	73.473	73.339
3	-0.5	1.5	87.684	87.864	87.839	82.298	82.838	82.771	73.306	74.749	74.709
3	-0.5	1	88.297	88.593	88.511	82.239	83.246	83.056	71.588	74.356	74.313
3.5	0	1.5	92.214	92.385	92.380	89.443	89.753	89.716	84.538	85.394	85.358
3.5	0	1	93.267	93.445	93.398	90.271	90.833	90.701	84.313	86.037	85.875
3.5	-0.5	1.5	93.367	93.495	93.465	90.439	90.750	90.711	85.414	86.217	86.130
3.5	-0.5	1	94.519	94.715	94.657	91.406	91.988	91.903	85.011	87.006	86.833

Table 6.28: Power of the tests using Procedure 3 with unequal stage sizes when $\Delta_2 = \Delta_1$

			5%			2.5%			1%		
d_1	c_1	c_2	T_1	T_2	T_∞	T_1	T_2	T_∞	T_1	T_2	T_∞
2	0	1.5	70.373	70.545	70.346	59.311	59.597	59.241	44.872	45.749	45.163
2	0	1	70.960	70.637	70.095	59.912	59.693	58.849	45.731	45.862	44.516
2	-0.5	1.5	71.246	71.349	71.059	59.871	60.101	59.709	45.145	45.806	45.198
2	-0.5	1	72.112	71.870	71.273	60.742	60.651	59.744	46.332	46.400	45.070
2.5	0	1.5	85.284	85.375	85.255	77.873	78.132	77.830	65.975	67.067	66.414
2.5	0	1	85.759	85.736	85.363	78.608	78.615	77.837	67.058	67.573	66.093
2.5	-0.5	1.5	86.360	86.470	86.236	78.542	78.829	78.471	66.235	67.045	66.321
2.5	-0.5	1	86.837	86.786	86.366	79.176	79.235	78.308	67.351	67.796	66.306
3	0	1.5	93.818	93.929	93.876	89.805	90.143	89.933	82.286	83.191	82.655
3	0	1	94.262	94.226	94.005	90.653	90.728	90.240	83.477	84.091	82.909
3	-0.5	1.5	94.682	94.771	94.692	90.552	90.849	90.614	82.818	83.733	83.232
3	-0.5	1	95.096	95.128	94.920	91.278	91.427	90.926	84.017	84.698	83.532
3.5	0	1.5	97.752	97.813	97.782	96.098	96.287	96.198	92.528	93.129	92.897
3.5	0	1	97.957	97.962	97.887	96.515	96.657	96.416	93.132	93.601	92.912
3.5	-0.5	1.5	98.311	98.369	98.339	96.613	96.801	96.705	92.822	93.473	93.186
3.5	-0.5	1	98.551	98.577	98.478	97.022	97.142	96.917	93.661	94.067	93.384

Table 6.29: Expected loss of the tests using Procedure 3 with unequal stage sizes when $\Delta_2 = 0$

			5%			2.5%			1%		
d_1	c_1	c_2	T_1	T_2	T_∞	T_1	T_2	T_∞	T_1	T_2	T_∞
2	0	1.5	0.8513	0.8332	0.8286	1.0558	1.0183	1.0112	1.3214	1.2536	1.2425
2	0	1	0.8874	0.8452	0.8361	1.1197	1.0421	1.0277	1.4212	1.2892	1.2690
2	-0.5	1.5	0.8050	0.7849	0.7805	1.0235	0.9820	0.9739	1.3054	1.2379	1.2265
2	-0.5	1	0.8428	0.7967	0.7873	1.0923	1.0073	0.9922	1.4039	1.2676	1.2469
2.5	0	1.5	0.6185	0.6026	0.5990	0.8057	0.7674	0.7599	1.1085	1.0180	1.0039
2.5	0	1	0.6408	0.6013	0.5929	0.8809	0.7883	0.7717	1.2570	1.0722	1.0473
2.5	-0.5	1.5	0.5491	0.5314	0.5272	0.7522	0.7095	0.7013	1.0766	0.9883	0.9734
2.5	-0.5	1	0.5818	0.5387	0.5294	0.8333	0.7390	0.7214	1.2163	1.0288	1.0024
3	0	1.5	0.3922	0.3834	0.3810	0.5130	0.4859	0.4804	0.7564	0.6797	0.6685
3	0	1	0.3995	0.3740	0.3689	0.5609	0.4928	0.4810	0.8818	0.7098	0.6868
3	-0.5	1.5	0.3113	0.3006	0.2987	0.4414	0.4138	0.4079	0.6965	0.6182	0.6068
3	-0.5	1	0.3163	0.2894	0.2828	0.4894	0.4178	0.4053	0.8255	0.6488	0.6252
3.5	0	1.5	0.2317	0.2273	0.2263	0.2844	0.2719	0.2688	0.4095	0.3673	0.3607
3.5	0	1	0.2154	0.2051	0.2029	0.2943	0.2590	0.2530	0.4914	0.3779	0.3647
3.5	-0.5	1.5	0.1537	0.1492	0.1488	0.2116	0.1974	0.1950	0.3472	0.3041	0.2968
3.5	-0.5	1	0.1481	0.1352	0.1323	0.2350	0.1953	0.1890	0.4428	0.3230	0.3077

Table 6.30: Expected loss of the tests using Procedure 3 with unequal stage sizes when $\Delta_2 = .5\Delta_1$

			5%			2.5%			1%		
d_1	c_1	c_2	T_1	T_2	T_∞	T_1	T_2	T_∞	T_1	T_2	T_∞
2	0	1.5	0.8985	0.8915	0.8916	1.0910	1.0754	1.0755	1.3358	1.3003	1.2994
2	0	1	0.8994	0.8863	0.8877	1.1116	1.0818	1.0828	1.3814	1.3195	1.3185
2	-0.5	1.5	0.8670	0.8605	0.8617	1.0664	1.0481	1.0481	1.3241	1.2891	1.2875
2	-0.5	1	0.8770	0.8624	0.8624	1.0971	1.0629	1.0642	1.3725	1.3074	1.3052
2.5	0	1.5	0.7289	0.7226	0.72389	0.9113	0.8943	0.8948	1.1852	1.1426	1.1411
2.5	0	1	0.7062	0.6934	0.6963	0.9191	0.8830	0.8844	1.2473	1.1563	1.1573
2.5	-0.5	1.5	0.6832	0.6771	0.6784	0.8731	0.8561	0.8562	1.1643	1.1200	1.1209
2.5	-0.5	1	0.6610	0.6480	0.6494	0.8793	0.8432	0.8472	1.2207	1.1283	1.1268
3	0	1.5	0.5493	0.5466	0.5472	0.6736	0.6616	0.6629	0.9021	0.8636	0.8655
3	0	1	0.4858	0.4775	0.4792	0.6397	0.6110	0.6141	0.9375	0.8489	0.8502
3	-0.5	1.5	0.4901	0.4869	0.4878	0.6222	0.6094	0.61088	0.8612	0.8214	0.8219
3	-0.5	1	0.4457	0.4383	0.4404	0.6082	0.5798	0.5836	0.9090	0.8250	0.8237
3.5	0	1.5	0.3946	0.3914	0.3920	0.4650	0.4578	0.4591	0.6072	0.5823	0.5839
3.5	0	1	0.3207	0.3166	0.3183	0.4102	0.3940	0.3978	0.6025	0.5444	0.5484
3.5	-0.5	1.5	0.3556	0.3536	0.3548	0.4313	0.4246	0.4260	0.5779	0.5543	0.5570
3.5	-0.5	1	0.2797	0.2751	0.2767	0.3728	0.3551	0.3579	0.5792	0.5125	0.5169

the results here with those in Section 6.3.3, Tables 6.17-6.22, we find that this design does not succeed in improving the properties under H_1. Although the power and expected loss are similar, they are mostly slightly inferior and the expected sample size under H_0, although also similar, is slightly higher.

6.5 Conclusions

In this chapter, we have shown that a two-stage adaptive design is very promising for three-arm trials with a control. In particular, Procedure 3 seems to give power and expected loss which are better than the fixed design, but with a smaller expected sample size. The results obtained here are from balanced designs, with equal allocation to each treatment. These are the only adaptive designs used in practice and commonly described in the literature, although a few authors, such as Koenig et al. (2008) have suggested that unbalanced designs might be beneficial.

We have only looked at two-stage designs, with a single interim analysis. However, methodologically, there is no restriction to the number of interim analyses which can be performed.

Table 6.31: Expected loss of the tests using Procedure 3 with unequal stage sizes when $\Delta_2 = \Delta_1$

			5%			2.5%			1%		
d_1	c_1	c_2	T_1	T_2	T_∞	T_1	T_2	T_∞	T_1	T_2	T_∞
2	0	1.5	0.5925	0.5891	0.5931	0.8138	0.8081	0.8152	1.1026	1.0850	1.0967
2	0	1	0.5808	0.5873	0.5981	0.8018	0.8061	0.8230	1.0854	1.0828	1.1097
2	-0.5	1.5	0.5751	0.5730	0.5788	0.8026	0.7980	0.8058	1.0971	1.0839	1.0960
2	-0.5	1	0.5578	0.5626	0.5745	0.7852	0.7870	0.8051	1.0734	1.0720	1.0986
2.5	0	1.5	0.3679	0.3656	0.3686	0.5532	0.5467	0.5543	0.8506	0.8233	0.8397
2.5	0	1	0.3560	0.3566	0.3659	0.5348	0.5346	0.5541	0.8236	0.8107	0.8477
2.5	-0.5	1.5	0.3410	0.3383	0.3441	0.5365	0.5293	0.5382	0.8441	0.8239	0.8420
2.5	-0.5	1	0.3291	0.3304	0.3409	0.5206	0.5191	0.5423	0.8162	0.8051	0.8424
3	0	1.5	0.1855	0.1821	0.1837	0.3059	0.2957	0.3020	0.5314	0.5043	0.5203
3	0	1	0.1721	0.1732	0.1799	0.2804	0.2782	0.2928	0.4957	0.4773	0.5127
3	-0.5	1.5	0.1595	0.1569	0.1592	0.2834	0.2745	0.2816	0.5155	0.4880	0.5030
3	-0.5	1	0.1471	0.1462	0.1524	0.2617	0.2572	0.2722	0.4795	0.4591	0.4940
3.5	0	1.5	0.0787	0.0765	0.0776	0.1366	0.1230	0.1331	0.2615	0.2405	0.2486
3.5	0	1	0.0715	0.0713	0.0740	0.1220	0.1170	0.1254	0.2404	0.2240	0.2481
3.5	-0.5	1.5	0.0591	0.05709	0.0581	0.1185	0.1120	0.1153	0.2512	0.2284	0.2385
3.5	-0.5	1	0.0507	0.0498	0.0533	0.1042	0.1000	0.1079	0.2219	0.2077	0.2316

If Procedure 3 is used with several interim analyses, one could start with a small value of c_2 and increase it through the analyses, so that it is very likely that one treatment would eventually be dropped, but only after sufficient information had been gained to make a good treatment selection. The benefits of this should be even greater with more than two experimental treatments. However, one of the commonly cited benefits of having many interim analyses is that they allow sequential hypothesis testing with early rejection of H_0. Although it has been avoided in this chapter, early rejection based on the test statistic T_2 is another area of research which remains to be explored.

Of course, there are many practical situations in which the use of adaptive designs is not possible, e.g. because it takes too long to obtain responses from the patients, but, when they can be used, they seem to have clear benefits, both in terms of the expected numbers of patients used and in terms of the amount of information obtained. We hope that opportunities for their application will arise in the near future.

Chapter 7

Conclusions and Further Work

7.1 Conclusions

This thesis has considered clinical trials for comparing several treatments with a control, in which the aim is to establish evidence for efficacy of at least one experimental treatment and, if it is established, to select a treatment to be recommended for practical use. Most important is to find a treatment which is better than the control and of secondary importance is to find the best such treatment. Efficacy is established by testing the null hypothesis $H_0 : \Delta_i \leq 0 \; \forall i \in \{1, \ldots, I\}$ against the alternative $H_1 : \Delta_i > 0$, for at least one $i \in \{1, \ldots, I\}$. If H_0 is rejected, then the experimental treatment with the best estimated response is selected.

The most important conclusion from this work is that, if this is the aim of the trial, then a test which is suitable for this hypothesis must be used. Many methods available in the literature for similar, but not identical, hypotheses have been used, but these can have undesirable, and even very poor, properties. This conclusion is not original, but it is not widely recognised, at least in applications to clinical trials, that with multiple treatments, when the alternative hypothesis is one-sided, the distinction between a null hypothesis of equality and one of inequality is of vital importance.

The emphasis in this thesis has been on developing likelihood ratio tests (LRTs) and studying their properties. In doing so, we have had to give a new definition of the size of a test, to take account of the probability of type-III errors, although this seems to have little impact in practice. For the common large-sample case, we conclude that the test statistic based on the LRT, T_2, has good properties, but so do the test statistic based on Dunnett's procedure, T_∞, and that based on Hochberg's procedure. If equal numbers of patients are allocated to each arm, it is difficult to convincingly recommend any one of these procedures for three-arm trials. Hochberg's procedure is a little more awkward to use than the others and does not seem to offer any particular advantages, so the choice seems to be between T_2 and T_∞. For more than three arms, the differences are somewhat greater and there are many configurations for which T_2 is somewhat better than T_∞.

Several authors have considered the so-called least favourable configuration (mainly for choosing a design) and on this basis, T_∞ would be preferred. However, this argument is not completely convincing, since it implies that it is more important to detect a difference of the type $(\Delta^*, 0, \ldots, 0)$ than one of the type $(\Delta^* - \varepsilon, \ldots, \Delta^* - \varepsilon)$ for a small value of ε. It also takes a minimax approach to the choice of statistic. Over a broad range of configurations, T_2 performs slightly better than T_∞, especially when the number of experimental arms is large. Also, in trials in which the experimental treatments are similar, e.g. different doses of the same drug, two drugs for the same target, or one drug either alone or in combination with another, then we would often expect *a priori* that either H_0 is true or Δ_1 and Δ_2 are both positive. Using T_2 is also consistent with using the χ^2 statistic for two-sided tests, rather than combining the results from I z-tests, which gives it a certain logic. On balance, it is justified to recommend T_2 as the most appropriate test statistic for this problem, but we recognise that T_∞ is also reasonably acceptable.

If these results were the strongest argument for T_2, the conclusions from this thesis might be rather disappointing. However, when different allocations of patients to treatments are possible, the benefit of T_2 becomes greater. With T_∞, other authors have shown that the so-called square root allocation, which for two experimental arms gives slightly more than 40% of patients to the control, maximises the power of the test. We have shown that with T_2 the optimal proportion allocated to the control is similar, but slightly higher, and that similar conclusions hold for the expected loss we have used, as well as power. When close to optimal allocations are used, T_2 has a slight, but clear, benefit over T_∞ in terms of power and expected loss. Given these results, it does not seem to be sensible to recommend different test statistics for different allocations, so our overall recommendation must be for T_2.

The large sample results developed will be adequate for many clinical trials but sometimes, especially with binary response data, they will be insufficiently precise. In this case, we recommend that the normal approximation be abandoned and a conditional exact test, using the LRT statistic for the binomial distribution, should be used. As far as its properties have been studied in this thesis, this test appears to show great promise, although its power has not been thoroughly studied, and its use is consistent with that of Fisher's exact test in the case of two-sided alternative hypotheses.

If a two-stage adaptive design can be used, it is possible to simultaneously achieve higher power if H_1 is true and a smaller expected sample size if H_0 is true, by allowing inferior treatments to be dropped after the first stage. A design which allows early stopping for futility and allows either one or two experimental arms to be forwarded to the second stage, along with the control, shows considerable advantages over the fixed design. In this situation, T_2 performs somewhat better than T_∞.

In conclusion, the likelihood ratio tests used and developed in this thesis have been shown to have good properties and can be recommended for use in practice whenever the aims of the trial dictate that the null hypothesis to be tested is $H_0 : \Delta_i \leq 0 \ \forall i$, there are no restrictions on the model and no prior information is to be used in the analysis. These methods now await practical use in clinical trials.

7.2 Further work

As well as requiring successful applications, before the methods presented here will be widely accepted, there are many further methodological developments that could be made. Some of these have been mentioned in passing throughout the thesis, but here we summarise a few ideas for further research.

The implementations describe here have been largely simulation based. This is probably the best that can be done and the general idea was strongly recommended in order restricted inference by Silvapulle and Sen (2005), in place of the approximations used in the older texts of Barlow et al. (1972) and Robertson et al. (1988). However, there might be a place for further approximations to be developed and compared with the simulation results. These would be particularly useful for further theoretical studies of the properties of the tests, where simulation takes a considerable time.

After defining the corrected size of a test to take account of the probability of a type-III, we found that this had no consequences for the usual significance levels. However, we have not dealt with how to make a similar adjustment to calculated p-values. Conceptually, there is no difficulty. The corrected p-value is the smallest $\alpha^\dagger$ such that H_0 is rejected at corrected significance level $\alpha^\dagger$. However computing this from the simulations presented here is rather tedious, since we would have to consider, in the three arm case, the probability of rejecting H_0 in favour of treatment 2 over a range of values of d_1 with $d_2 = 0$. It would be particularly useful to try to derive expressions giving the value of d_1 which maximises the probability of a type-III error, or at least bounds on this value. A separate, but related problem, is to work out a suitable correction for the p-value in the exact test for binary responses.

The emphasis in this thesis has been on large-sample normal approximations. Although these will be used in many applications, there will be some trials in which they cannot be relied upon. The exact conditional test for binary responses, developed in Chapter 5, could be extended to ordered (or unordered) categorical responses. Silvapulle and Sen (2005) developed an exact conditional test for the null hypothesis of equality in an order restricted model, which was a direct extension of the corresponding test for binary responses. We would expect that a direct extension of our exact conditional test could be developed, although we would require either an assumption of proportional odds, or a further refinement of the null hypothesis.

Small-sample tests for continuous data are a more difficult case to consider. If normality can be assumed, it will be straightforward to modify the testing procedure for unknown, but equal, variances and indeed Robertson et al. (1988) and Silvapulle and Sen (2005) show how the LRT statistic can be obtained. An open question is whether this test statistic can be expressed in a neat form similar to T_2.

If it is not reasonable to assume normality, an obvious form of test is a permutation test. However, a straightforward permutation test fails, because under H_0, the responses from different treatments are not necessarily exchangeable (the means from each treatment need not be equal). One possibility would be to adjust for the restricted maximum likelihood estimates under H_0 and use permutations of the residuals. Similarly, the residuals could be bootstrapped. The difference between these two methods is that the former samples from the residuals without replacement, while the latter samples with replacement. The permutation test relies less on large sample approximations, but the results of bootstrapping, if valid, are applicable to the population of patients and not just those in the trial. These methods deserve further investigation.

As noted in Chapter 6, this thesis has only just scratched the surface of adaptive and sequential designs for multi-arm trials. An enormous amount of work remains to be done in this area and it is quite possible that this will become a topic of major interest in the near future. The first priority in this area is probably a more detailed study of the sizes of stages and the balance between treatments within stages, to optimise the properties of two-stage designs. Given that no improvement was obtained from the designs with unequal stage sizes studied, further exploration of this might not lead to great improvements. It might, therefore, be more beneficial to consider designs with more than two stages.

The topic of estimation has not been discussed in this thesis, although it will almost always be needed in clinical trials, even if only as a secondary analysis after testing. When H_0 is rejected and a treatment selected, the simple estimator of the difference between that treatment and the control is positively biased, due to the effect of selection. This is most obvious in the case of $\Delta_1 = \Delta_2$, where the larger of two estimates will be used. The development of corrections to the estimator to make it approximately unbiased is an essential area of future work and could be a major area of research in itself. We would expect such corrections would make use of the estimates of other treatment differences. If the selected treatment is estimated to be considerably better than any of the other treatments, the bias will be trivially small, whereas if there are several nearly equally good treatments, it will be more substantial. In adaptive designs, there is an additional selection bias caused by the selection of treatments after stage 1 and the decision to continue to stage 2.

In conclusion, we hope that research in the use of order restricted inference in clinical trials will continue and that it will prove beneficial in their design and analysis.

Bibliography

Abelson, R. P. and J. W. Tukey (1963). Efficient utilization of non-numeric information in quantitative analysis: general theory and the case of a simple order. *Annals of Mathematical Statistics 34*, 1347–1369.

Agresti, A. and B. A. Coull (1996). Order-restricted tests for stratified comparisons of binomial proportions. *Biometrics 52*, 1103–1111.

Appel, L. J., T. J. Moore, E. Obarzanek, W. M. Vollmer, L. P. Svetkey, F. M. Sacks, G. A. Bray, T. M. Vogt, J. A. Cutler, M. M. Windhauser, P.-H. Lin, and N. Karanja (1997). A clinical trial of the effects of dietary patterns on blood pressure. *The New England Journal of Medicine 336*, 1117–1124.

Armitage, P., G. Berry, and J. N. S. Matthews (2002). *Statistical Methods in Medical Research* (4th ed.). Blackwell.

Atkinson, A. C., A. N. Donev, and R. Tobias (2007). *Optimum Experimental Designs, with SAS.* Oxford University Press.

Barlow, R. E., D. J. Bartholomew, J. M. Bremner, and H. D. Brunk (1972). *Statistical Inference Under Order Restrictions.* Wiley.

Bauer, P. (1989). Multistage testing with adaptive designs (with discussion). *Biometrie und Informatik in Medizin und Biologie 20*, 130–148.

Bayarri, M. J. and J. O. Berger (2004). The interplay of bayesian and frequentist analysis. *Statistical Science 19*, 58–80.

Beal, S. L. (1987). Asymptotic confidence intervals for the difference between two binomial parameters for use with small samples. *Biometrics 43*, 941–950.

Bechhofer, R. E. (1969). Optimal allocation of observations when comparing several treatments with a control. In *Multivariate Analysis*, Volume II, pp. 463–473. Academic Press Inc.

Bechhofer, R. E. and D. J.-M. Nocturne (1972). Optimal allocation of observations when comparing several treatments with a control, II: 2-sided comparisons. *Technometrics 14*, 423–436.

Bechhofer, R. E. and A. C. Tamhane (1983). Design of experiments for comparing treatments with a control: Tables of optimal allocations of observations. *Technometrics 25*, 87–95.

Betensky, R. A. (1996). An O'Brien-Fleming sequential trial for comparing three treatments. *The Annals of Statistics 24*, 1765–1791.

Betensky, R. A. (1997). Sequential analysis of censored survival data from three treatment groups. *Biometrics 53*, 807–822.

Bofinger, E. (1985). Multiple comparisons and type III errors. *Journal of the American Statistical Association 80*, 433–437.

Chen, J. and S. K. Sarkar (2004). Multiple testing of response rates with a control: a bayesian stepwise approach. *Journal of Statistical Planning and Inference 125*, 3–16.

Chuang-Stein, C. and D. M. Tong (1995). Multiple comparisons procedures for comparing several treatments with a control based on binary data. *Statistics in Medicine 14*, 2509–2522.

Cummings, S. R., S. Eckert, K. A. Krueger, D. Grady, T. J. Powells, J. A. Cauley, L. Norton, T. Nickelson, N. H. Bjarnason, M. Morrow, M. E. Lippman, D. Black, J. E. Glusman, A. Costa, and V. C. Jordan (1999). The effect of raloxifene on risk of breast cancer in postmenopausal women: results from the MORE randomized trial. *The Journal of the American Medical Association 281*, 2189–2197.

Cuzick, J. (1982). The efficiency of the proportions test and the logrank test for censored survival data. *Biometrics 38*, 1033–1039.

Dmitrienko, A., A. C. Tamhane, and F. Bretz (2010). *Multiple Testing Problems in Pharmaceutical Statistics.* CRC Press.

Dunnett, C. W. (1955). A multiple comparisons procedure for comparing several treatments with a control. *Journal of the American Statistical Association 50*, 1096–1121.

Dunnett, C. W. (1964). New tables for multiple comparisons with a control. *Biometrics 20*, 482–491.

Dunnett, C. W. (1984). Selection of the best treatment in comparison to a control with application to a medical trial. In T. J. Santner and A. C. Tamhane (Eds.), *Design of Experiments: Ranking and Selection*, pp. 47–66. Marcel Dekker.

Dunnett, C. W., M. Horn, and R. Vollandt (2001). Sample size determination in step-down and step-up multiple tests for comparing treatments with a control. *Journal of Statistical Planning and Inference 97*, 367–384.

Dunnett, C. W. and A. C. Tamhane (1992). A step-up multiple test procedure. *Journal of the American Statistical Association 87*, 162–170.

Finney, D. J. (1952). *Statistical Methods in Biological Assay.* Hafner.

Fleiss, J. L. (1986). *The Design and Analysis of Clinical Experiments.* Wiley.

Follmann, D. A., M. A. Proschan, and N. L. Geller (1994). Monitoring pairwise comparisons in multi-armed clinical trials. *Biometrics 50*, 325–336.

Harter, H. L. (1957). Error rates and sample sizes for range tests in multiple comparisons. *Biometrics 13*, 511–536.

Hellmich, M. (2001). Monitoring clinical trials with multiple arms. *Biometrics 57*, 892–898.

Hellmich, M. and G. Hommel (2004). *Multiple Testing in Adaptive Designs - A Review.* Institute of Mathematical Statistics Lecture Notes - Monograph Series.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika 75*, 800–802.

Hochberg, Y. and D. Rom (1995). Extensions of multiple testing procedures based on Simes' test. *Journal of Statistical Planning and Inference 48*, 141–152.

Hochberg, Y. and A. C. Tamhane (1987). *Multiple Comparison Procedures.* Wiley.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6*, 65–70.

Hommel, G. (2001). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika 75*, 383–386.

Horn, M. and C. W. Dunnett (2004). Power and sample size comparisons of stepwise FWE and FDR controlling test procedures in the normal many-one case. In *Recent Developments in Multiple Comparison Procedures*, Volume 47 of *Lecture Notes - Monograph Series*, pp. 48–64. Institute of Mathematical Statistics.

Horn, M. and R. Vollandt (1998). Sample sizes for comparisons of k treatments with a control based on different definitions of the power. *Biometrical Journal 40*, 589–612.

Hsu, J. C. (1989). Sample size computation for designing multiple comparison experiments. *Computational Statistics & Data Aanalysis 7*, 79–91.

Hsu, J. C. (1996). *Multiple Comparisons.* Chapman & Hall.

Hughes, M. D. (1993). Stopping guidelines for clinical trials with multiple treatments. *Statistics in Medicine 12*, 901–915.

Hwang, J. T. G. and S. D. Peddada (1994). Confidence interval estimation subject to order restrictions. *The Annals of Statistics 22*, 67–93.

Jennison, C. and B. W. Turnbull (2000). *Group Sequential Methods with Applications to Clinical Trials.* CRC Press.

Jennison, C. and B. W. Turnbull (2003). Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine 22*, 971–993.

Jennison, C. and B. W. Turnbull (2006). Adaptive group sequential tests. *Biometrika 93*, 1–21.

Kelly, P. J., N. Stallard, and S. Todd (2005). An adaptive group sequential design for phase II/III clinical trials that select a single treatment from several. *Journal of Biopharmaceutical Statistics 12*, 641–658.

Kieser, M. and T. Friede (2007). Planning and analysis of three-arm non-inferiority trials with binary endpoints. *Statistics in Medicine 26*, 253–273.

Koch, H.-F. and L. A. Hothorn (1999). Exact unconditional distributions for dichotomous data in many-to-one comparisons. *Journal of Statistical Planning and Inference 82*, 83–99.

Koenig, F., W. Brannath, F. Bretz, and M. Posch (2008). Adaptive Dunnett tests for treatment selection. *Statistics in Medicine 27*, 1612–1625.

Krzanowski, W. J. and F. H. C. Marriott (1994). *Multivariate Analysis, Part 1.* Edward Arnold.

Lee, P. M. (2004). *Bayesian Statistics: An Introduction* (3rd ed.). Arnold.

Liu, W. (1997). On sample size determination of Dunnett's procedure for comparing several treatments with a control. *Journal of Statistical Planning and Inference 62*, 255–261.

Marschner, I. C. (2007). Optimal design of clinical trials comparing several treatments with a control. *Pharmaceutical Statistics 6*, 23–33.

Mead, R. (1988). *Design of Experiments.* Cambridge University Press.

Mehta, C. R. (1994). The exact anylysis of contingency tables in medical research. *Statistical Methods in Medical Research 3*, 135–156.

Mehta, C. R. and N. R. Patel (1983). A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. *Journal of the American Statistical Association 78*, 427–434.

Mosteller, F. (1948). A k-slippage test for an extreme population. *The Annals of Mathematical Statistics 19*(1), 58–65.

Mukerjee, H., T. Robertson, and F. T. Wright (1985). *Advances in Order Restricted Statistical Inference.* Springer-Verlag.

Mukerjee, H., T. Robertson, and F. T. Wright (1987). Comparisons of several treatments with a control using multiple contrasts. *Journal of the American Statistical Association 82*, 902–910.

Müller, H.-H. and H. Schäfer (2001). Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approches. *Biometrics 57*, 886–891.

Paulson, E. (1952). On the comparison of several experimental categories with a control. *The Annals of Mathematical Statistics 23*, 239–246.

Peddada, S. D., J. K. Haseman, X. Tan, and G. Travlos (2006). Tests for a simple tree order restriction with application to dose-response studies. *Applied Statistics 55*, 493–506.

Peddada, S. D., K. E. Prescott, and M. Conaway (2001). Tests for order restrictions in binary data. *Biometrics 57*, 1219–1227.

Piegorsch, W. W. (1990). One-sided singnificance tests for generalized linear models under dichotomous response. *Biometrics 46*, 309–316.

Piegorsch, W. W. (1991). Multiple comparisons for analyzing dichotomous response. *Biometrics 47*, 45–52.

Pigeot, I., J. Schäfer, J. Röhmel, and D. Hauschke (2007). Assessing non-inferiority of a new treatment in a three-arm clinical trial including a placebo. *Statistics in Medicine 26*, 253–273.

Pocock, S. J. (1983). *Clinical Trials.* Wiley.

Posch, M., F. Koenig, M. Branson, and W. Brannath (2005). Testing and estimation in flexble group sequential designs with adaptive treatment selection. *Statistics in Medicine 24*, 3697–3714.

Proschan, M. A. and D. A. Follmann (1995). Multiple comparisons with control in a single experiment versus separate experiments: Why do we feel differently? *The American Statistician 49*, 144–149.

Proschan, M. A., D. A. Follmann, and N. L. Geller (1994). Monitoring multi-armed trials. *Statistics in Medicine 13*, 1441–1452.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing.

Robertson, T., F. T. Wright, and R. L. Dykstra (1988). *Order Restricted Statistical Inference.* Wiley.

Royston, P., M. K. B. Parmar, and W. Qian (2003). Novel designs for multi-arm clinical trials with survival outcomes with an application in ovarian cancer. *Statistics in Medicine 22*, 2239–2256.

Schaafsma, W. and L. J. Smid (1966). Most stringent somewhere most powerful tests against alternatives restricted by a number of linear inequalities. *Annals of Mathematical Statistics 37*, 1161–1172.

Schaid, D. J., S. Wieand, and T. M. Therneau (1990). Optimal two-stage screening designs for survival comparisons. *Biometrika 77*, 507–13.

Sen, P. K. and M. J. Silvapulle (2001). An appraisal of some aspects of statistical inference under inequality constraints. *Journal of Statistical Planning and Inference 107*, 3–43.

Shaffer, S. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology 46*, 561–584.

Silvapulle, M. J. and P. K. Sen (2005). *Constrained Statistical Inference.* Wiley.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika 73*, 751–754.

Stallard, N. and S. Todd (2003). Sequential designs for phase III clinical trials incorporating treatment selection. *Statistics in Medicine 22*, 689–703.

Suissa, S. and J. J. Shuster (1985). Exact unconditional sample sizes for the 2x2 binomial trial. *Journal of the Royal Statistical Society, Series A 148*, 317–327.

Tang, D.-I. and S. P. Lin (1997). A approximate likelihood ratio test for comparing several treatments to a control. *Journal of the American Statistical Association 92*, 103–117.

Thall, P. F., R. Simon, and S. S. Ellenberg (1988). Two-stage selection and testing designs for comparative clinical trials. *Biometrika 75*, 303–310.

The ATAC Trialists' Group (2002). Anastrozole alone or in combination with tamoxifen versus tamoxifen alone for adjuvant treatment of postmenopausal women with early breast cancer: first results of the ATAC randomised trial. *The LANCET 359*, 2131–2139.

Thompson, W. A. (1962). The problem of negative estimates of variance components. *The Annals of Mathematical Statistics*, 273–289.

Vincent, E., S. Todd, and J. Whitehead (2002). A sequential procedure for comparing two experimental treatments with a control. *Journal of Biopharmaceutical Statistics 12*, 249–265.

Whitehead, J. R. (1997). *The Design and Analysis of Sequential Clinical Trials* (2nd ed.). Wiley.

Williams, D. A. (1988). Tests for differences between several small proportions. *Applied Statistics 37*, 421–434.

Zhao, H. B. (2007). Comparing several treatments with a control. *Journal of Statistical Planning and Inference 137*, 2996–3006.