A RESTRICTED CLASS OF GROUP SEQUENTIAL DESIGNS FOR MONITORING CLINICAL TRIALS

by

Larry Douglas Case

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Mimeo Series No. 1824T

May 1987

A RESTRICTED CLASS OF GROUP SEQUENTIAL DESIGNS
FOR MONITORING CLINICAL TRIALS

by

Larry Douglas Case

A Dissertation submitted to the faculty of The University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill

1987

Approved by:

Advisor

Reader

Reader

Reader

Reader

LARRY DOUGLAS CASE. A Restricted Class of Group Sequential Designs for Monitoring Clinical Trials. (Under the direction of C. E. DAVIS.)

A feature of most group sequential designs is that the critical value at the final stage does not equal the critical value corresponding to a fixed sample design. This can lead to confusion when a seemingly significant sample statistic at the final stage fails to reject the null hypothesis. Two and three-stage designs are presented which allow acceptance, rejection, or continuation at the first stage (or first two stages) while retaining the fixed sample critical value at the final stage. The designs are optimized under this restriction to minimize either the maximum expected sample size or a weighted average of the expected sample sizes under the null and alternative hypotheses. These restricted designs are almost fully efficient when compared to optimal unrestricted designs. In addition, these designs provide a substantial proportion of the savings in sample size possible using a fully sequential design. In some cases, more than 50% of the savings can be realized with only one interim look at the data and 75% of the possible savings can be realized with just two interim looks. The greatest savings per stage occur in going from a fixed sample to a two-stage design. Approximately 40% to 50% of the savings can be realized with this one additional look. Approximately 15% to 25% of the remaining savings can be realized with the second interim look. These restricted

designs provide an efficient and practical method for interpreting hypothesis tests during interim and final analyses. Examples are given which illustrate their use in the design and analysis of clinical trials.

ACKNOWLEDGEMENTS

I wish to thank my advisor, Dr. Ed Davis, for his friendly and encouraging leadership, and the committee members, Dr. Michael Symons, Dr. Herman Tyroler, Dr. P. K. Sen, and particularly Dr. Tim Morgan, for their helpful comments and discussions. I appreciate Dr. Sen pointing out alternative methods for calculating multidimensional integrals. Sincere thanks go to Don Hire for programming the software which allowed greek letters and mathematical symbols to be included easily into the text. I am indebted Dr. H. Bradley Wells, a friend and mentor, for the "gentle" nudging which led to my graduate studies and for his continued support and encouragement. I am also grateful to Dr. Wells and Dr. Curt Furberg for providing free computer time. Special thanks go to my wife, Janet, and my children, John and Jane, for their constant love and understanding.

TABLE OF CONTENTS

I. INTRODUCTION AND LITERATURE REVIEW

1.1 Introduction

Investigators conducting clinical trials involving human subjects typically review accumulating data one or more times before the study is complete to ensure that no group of patients is receiving an inferior therapy. Such interim reviews are ethical since real differences in treatment efficacy should lead to early termination of clinical trials. Hypothesis tests are usually performed during these interim reviews to assist in deciding whether or not to continue the trial. Investigators do not want to continue trials in which one treatment is clearly superior; nor do they want to continue trials which have little hope of showing a treatment benefit. Unfortunately, interim hypotheses are sometimes rejected at a critical value appropriate for a fixed sample design. This increases the probability of falsely declaring a statistically significant treatment effect during the course of the trial, much as multiple comparisons increase the probability of finding a significant contrast in analysis of variance. The usual approach to this interim analysis problem is to use a smaller nominal significance level for each review so that the significance level over all analyses will equal the corresponding fixed sample value. With this approach, one might be faced with the unenviable task of explaining why a 'significant' p-value at the final stage does not lead to rejection of the null hypothesis. It may be difficult to understand why final results should be interpreted differently just because of multiple looks, as it can be argued that the final data will be the same whether or not they are reviewed periodically.

An objective of this research is to develop group sequential designs which allow hypothesis tests at the final analysis to be interpreted as though a fixed sample design had been used. These designs should satisfy the ethical need to monitor accumulating data and eliminate some incredulity at the final analysis. The features of the proposed designs are as follows.

a) The overall level of significance is α.

b) The critical value at the final stage is equivalent to the critical value of a fixed sample design with the same significance level.

The plans developed give stopping rules which serve as guidelines in monitoring clinical trials for early termination. However, no stopping rule is a panacea for all the difficulties inherent in closing a study. Many factors are involved in such a decision, and usually many endpoints are being evaluated. Pocock (1980,1983), Friedman, Furberg, and DeMets (1981), and Demets (1984) discuss many of the issues to be considered. These papers poignantly illustrate the demanding and sometimes controversial decisions which must often be made by monitoring committees.

Throughout this discussion, emphasis will be placed on the design and analysis of clinical trials. However, the designs presented herein can also be used in sampling inspection and drug screening. In fact, they may be more appropriate in such settings since human ethics are not involved and usually only one endpoint is considered.

1.2 Literature Review

1.2.1 Repeated Significance Tests

As originally pointed out by Feller (1940) and later discussed by Robbins (1952) and Anscombe (1954), repeated significance tests on accumulating data can increase the probability of obtaining a significant result when the null hypothesis of no treatment effect is true. Armitage, McPherson, and Rowe (1969) examined this problem for response variables having binomial, normal, and exponential distributions. Some results for the normal case are presented in Table 1.1. These values are calculated under the assumption that the difference in response between two treatments has a normal distribution with known variance.

Table 1.1. Probability of significant result at some stage of a K-stage study when there is no difference between two treatments. (*)

Significance level per test	Number of repeated tests (K) 1	2	3	4	5	10	25	50	100
.01	.01	.018	.024	.029	.033	.047	.070	.088	.107
.05	.05	.083	.107	.126	.142	.193	.266	.320	.374
.10	.10	.160	.202	.234	.260	.342	.449	.524	.592

* Reference: Armitage et al (1969). Normal responses with known variance; two-sided tests carried out at equal accrual intervals.

These results indicate that the probability of falsely rejecting the null hypothesis can be much larger than the nominal significance level used for each test. This probability is more than double the nominal significance level for only three repeated tests, and it increases to unity as the number of tests approaches infinity. McPherson and Armitage (1971) and McPherson (1974) add additional insight on this issue. Investigators who plan on analyzing their data more than once should use designs which control the overall significance level. Sequential and group sequential designs have been proposed which allow for multiple looks at the data while maintaining the overall significance level at some prespecified value. Such designs not only control the overall significance level, but require, on average, a smaller sample size than required for fixed sample designs.

1.2.2 Sequential Designs

The decision to stop or continue a fully sequential trial is made after each patient or pair of patients is accrued. These designs thus involve the greatest possible number of interim analyses. The theory of sequential analysis was developed by Wald (1947) in the 1940's. This methodology created much excitement because, in expectation, it led to a large savings in sample size. Wald and Wolfowitz (1948) showed that fully (item-by-item) sequential designs yield the lowest expected sample size under both the null and the specified alternative hypotheses. Despite the initial enthusiam, several aspects of sequential design proved impractical and few clinical trials have used them in practice. Constant data monitoring is necessary since decisions are made after each accrual. Rapid response measures are required since the results of one patient need to be evaluated prior to the accrual of another. Most sequential designs require matched pairs for treatment comparison and have no maximum sample size. Matched pairs are difficult to obtain in practice, and open-ended designs can create administrative difficulties.

Extensions and variations of sequential designs have been developed by several investigators. Armitage (1975) developed closed sequential designs based on repeated significance testing. Additionally, Armitage (1978) suggested that sequential methods could be applied to the analysis of survival data using either the logrank test (Mantel, 1966) or Cox's (1972) proportional hazards regression model. Jones and Whitehead (1979a) and Sellke and Siegmund (1983) developed the necessary theory to support these contentions. These advances add to the applicability of sequential designs in actual practice; however, it still appears that few trials utilize these methods. An excellent account of the theory and application of sequential designs is given by Whitehead (1983).

1.2.3 Group Sequential Designs

Data are analyzed after groups of patients are entered into a group sequential study. Group sequential plans are generally more practical, and they give much of the savings possible from sequential designs. Many of these designs are based on the work of Armitage et al (1969) on repeated significance testing and were developed to match the theory of clinical trials to its actual practice. Jones and Whitehead (1979b) and Pocock (1979) argue the relative merits of sequential and group sequential methods. Basically, the group sequential designs are less affected by the problems of sequential designs mentioned above. On the other hand, group sequential designs generally require more patients. However, the increase in the number of patients is usually not great, and as few as five interim analyses can realize most of the sample size savings expected from sequential designs.

The two-stage plan is the simplest form of a group sequential design and is important from its historical prespective (it predated the sequential designs of Wald), its relative simplicity, and the fact that it can lead to major savings in sample size. McPherson (1982) determined the optimal number of interim analyses based on prior distributions which describe the possible treatment effect. He found that one look was best for well-defined priors centered about zero, while additional looks were warranted for poorly defined priors. Two looks were almost always justified, and they accounted for much of the possible savings in sample size. Much of the work to follow in this thesis is devoted to developing new two-stage designs, so their history will be discussed separately, followed by a discussion of multi-stage designs.

1.2.3.1 <u>Two-stage Designs</u>

In general a two-stage design (also called double sample test) proceeds as follows. At the first stage, a sample of n_1 patients is obtained, and a statistic based on these observations is calculated. If the statistic is less than C_1, the trial is stopped and the null hypothesis accepted. If the statistic is greater than C_2, the trial is stopped and the null hypothesis rejected. If the statistic lies between C_1 and C_2, a second sample of n_2 patients is obtained. A new statistic based on the $n_1 + n_2$ patients is then calculated. If the new statistic is greater than C_3, the null hypothesis is rejected; otherwise it is accepted. Some designs presented in the literature allow only rejection of H_0 at the first stage while other designs allow either decision at the first stage.

The reference usually given as providing the first discussion of a two-stage design is that of Dodge and Romig (1929) who developed plans to determine the acceptability of individual lots submitted for inspection. Their plans proceeded in two stages, but they involved only a single sample. At the first stage, a sample of n_1 items was randomly selected from the lot of n items. If C or fewer items were defective the lot was considered acceptable; otherwise the entire lot was inspected. These same authors (1941) also developed double sample designs. Here a first sample was selected, and the lot was accepted if C_1 or fewer items were defective and rejected if more than C_2 items were defective; otherwise a second sample was obtained. If the number of defectives for both samples combined was C_2 or less the lot was accepted; otherwise the entire lot was inspected.

Hamaker and Van Strik (1955) compared the efficiency of the double sample tests proposed by Dodge and Romig (1941) against equivalent fixed sample tests. They used the following two definitions of equivalency:

1) points and slopes of the Operating Characteristic (OC) curves match at an acceptance probability of 50%, and

2) OC curves match at 10% and 95% acceptance probabilities.

The majority of their work was obtained using equivalency definition (1), but they showed that the use of definition (2) would give similar results. They pointed out that the choice of $C_2 = C_3$ used by Dodge and Romig (1941), although convenient, does not give the most efficient tests.

These early papers laid the groundwork for many of the developments to follow. The majority of the research on two-stage designs has been done for normal and binomial response variables.

Owen (1953) describes three different two-stage tests for one-sided hypotheses about a normal mean with known variance. For his first test, a sample of size n_1 is obtained at the first stage. A decision based on these initial observations is then made to accept H_0, reject H_0, or continue to the second stage. If a second stage is necessary, n additional observations are obtained and the fixed sample test of the hypothesis is made based only on these last n observations. This test is clearly not optimally efficient since no use is made of the initial sample at the second stage. The first stage of Owen's second test proceeds as before. However, if a second stage is necessary, a second sample of size n_2 is obtained and the test of significance is made using all $n_1 + n_2$ observations.

Owen constrains the total sample size ($n_1 + n_2$) to be the sample size of the fixed sample test with some loss in power. The third test presented by Owen is an approximation of the second test for cases in which the bivariate normal distribution is not properly tabulated. This is no longer a problem since numerical procedures are available which evaluate bivariate normal integrals.

Colton (1963) developed optimal two-stage designs for drug screening and compared these designs to fixed sample and fully sequential designs. His designs did not allow acceptance at the first stage since he felt this was not "numerically feasible" and since there was sceptism towards accepting new drugs based on a single test. He also investigated the case where $n_1 = n_2$ and found that this restriction never increased the expected sample size (ESS) more than 3% for the plans he presented. Colton's designs minimized a weighted average of the ESS under the null and alternative hypotheses, where the weight could be estimated by the previous proportion of effective drugs. As an aside, he allowed acceptance at the first stage when the type I and II errors were the same, making the assumption that the most efficient designs were symmetric. In fact, these plans are not the most efficient for that optimality criterion as will be shown in Chapter 2. Colton did admit that more efficient nonsymmetric plans might exist. Colton and McPherson (1976) later extended these results to the clinical trial setting (which was really nothing more than changing terminology).

Methods for determining the operating characteristic curves and the accept-reject regions for two-stage screening procedures were given by Roseberry and Gehan (1964). They limited themselves to the case of equal sample sizes at each stage and allowed acceptance of H_0 at the final stage only since they felt that multiple tests of effective products was the usual decision, additional testing could be done without having to plan another study, and equivalent testing would have been done on all drugs passing the screen.

Elder (1974) considered the problem of acceptance sampling and presented an optimal symmetric two-stage plan for $\alpha = \beta = 0.05$. He stated that the efficiency is not very sensitive to differences in the mean. It will be seen that this is not necessarily true in general. Hald (1975) determined optimal two-stage designs using minimax and Bayes weighted average optimality criteria. He gave only symmetric plans for $\alpha = \beta$ and nonsymmetric plans for $\beta = 2*\alpha$. His designs are extended to other values of α and β in Chapter 4 and compared to designs developed in Chapter 2.

Elashoff and Reedy (1984) presented two-stage designs intermediate to other proposed plans. They considered the cases where one-half or two-thirds of the patients were accrued at the first stage. These authors felt that optimizing a stopping rule for a particular alternative was artificial, and they proposed designs which performed well over a range of alternatives.

Colton and McPherson (1976) determined optimal two-stage designs for binomial variates which did not allow acceptance of the null hypothesis at the first stage. They felt there was no ethical advantage in terminating clinical trials early and concluding no difference, and the mathematics were simplified. Equal sample sizes at each stage were used, and the type I and II errors were allowed to be within $\pm$ 10% (and sometimes 20%) of their nominal levels.

Elashoff and Beal (1976) presented two-stage designs for screening potential carcinogens where the outcome was the presence or absence of a tumor. They allowed acceptance at each stage and minimized a weighted average of the ESS under the null and alternative hypotheses. The majority of their results were found using a weight of 0.1, but they showed that changing the weight to 0.5 had little effect on the design parameters. These authors also showed that the expected sample size obtained using a normal approximation to the binomial was similar to the exact solution; however optimal parameters were not always "well determined" by the approximation, and the approximation took more computer time.

Spurrier (1978) extended the results of Colton and McPherson (1976) by allowing acceptance of the null hypothesis at the first stage. In addition, he introduced randomized tests so that exact α and β levels could be attained. Designs were given for α = 0.01 and 0.05. He matched the power curve of his design to that of the fixed sample design at three points - α, 0.5, and 0.9 - and sought the design which minimized the maximum ESS. His designs were restricted to equal sample sizes at each stage for n_1 and n_2 between .53n and .63n, where n is the fixed sample size. He found that his designs gave a smaller ESS for powers less than 0.8 and a slightly larger ESS for powers greater than 0.8 compared to those of Colton and McPherson.

De With (1982), using optimality criteria previously considered by Hald (1975), derived optimal designs for binomial variates which allowed acceptance and rejection at the first stage. He chose error levels less than or equal to the nominal values pointing out that only one of 29 Colton-McPherson plans checked had type I or II error levels less than the nominal values. The majority of De With's results were for the case of equal sample sizes at each stage, and he showed there was not much lost with this assumption when $\alpha = 0.05$ and $\beta = 0.10$. He found that the Colton-McPherson plans could be improved and the choice of a plan did not depend heavily on the optimality criterion.

Various extensions to the two-stage design have been proposed in the literature. Spurrier and Hewett (1975) extended previous results to the case of unknown variance using the bivariate t-distribution. They presented designs for $\alpha = 0.01$ and 0.05. These authors matched the power of the corresponding fixed sample t-test at α, 0.5, and 0.9, and they presented designs with constraints on n_1 and n_2.

These same authors (1976) also proposed two-stage Wilcoxon signed rank and rank sums tests. They showed that the limiting distribution for both these tests was the bivariate normal, so that designs developed for normally distributed data could be used for these nonparametric tests with moderate to large sample sizes.

1.2.3.2 Multi-stage Designs

The first multi-stage designs were simply ad hoc stopping rules which allowed for multiple reviews of data collected in fixed sample designs without substantially affecting the type I and II error levels. Haybittle (1971) and Peto et al (1976) both suggested using a Z statistic of 3.0 for interim analyses and the usual test at the final review. This allowed for early termination if extreme differences occurred while practically maintaining the simplicity of a fixed sample design. However, this approach has few other redeeming features. Only huge differences lead to closure of the trial during interim reviews. This may present ethical problems since real but less substantial differences would not lead to early termination. In addition, these designs are not efficient in terms of minimizing the ESS.

Pocock (1977) devised multi-stage designs for normal means with known variance. He assumed equal samples would be accrued at each stage and a constant significance level would be used for each test. His designs, which allow acceptance of H_0 at the final stage only, are based on the results of Armitage, McPherson, and Rowe (1969) mentioned previously. Pocock showed via simulations that these designs provide good approximations for other types of response variables (eg, binomial, exponential, or normal with unknown variance). The binomial case required larger sample sizes for the designs to be appropriate. He also indicated that slight departures from equal samples at each stage would not have much effect on the error levels.

O'Brien and Fleming (1979) developed plans for comparing two treatments with dichotomous outcomes using Pearson's chi-square as the test statistic. Like Pocock, they allowed acceptance of H_0 at the final stage only. However, they did not use constant nominal significance levels at each stage. They allowed rejection of H_0 at the i^{th} stage if $(i/K)\chi^2(i) > P(K,\alpha)$, where K denotes the maximum number of tests, and $P(K,\alpha)$ was determined by numerical simulations such that the overall level of significance was α. They found that $P(K,\alpha)$ could be approximated by $\chi^2(1,1-\alpha)$ without increasing the type I error substantially. Pocock (1982) compared these designs to his equal α_i designs. He found that his designs had smaller ESS for trials with reasonable power (> 80%). On the other hand, the O'Brien-Fleming designs had smaller ESS for low powers.

DeMets and Ware (1980) extended the group sequential methods of Pocock (1977) to the case of one-sided hypotheses. They felt these designs were appropriate for testing a new therapy versus the standard in which case it might be unethical to continue a trial to see if the new treatment is actually worse than the standard. Three methods were presented. The first was a simple modification of Pocock's plans, rejecting H_0 if the upper boundary was crossed, accepting H_0 if the lower boundary was crossed, and continuing otherwise. (Recall that Pocock's two-sided plans rejected H_0 if the upper or lower boundary was crossed.) For their second method, they used constant significance levels at each stage with different upper and lower boundaries. The choice of the lower boundary was arbitrary and was selected "to reflect the strength of the negative evidence required." Their third approach was motivated by the early work of Wald (1947), and they referred to the method as the 'constant likelihood approach'. The lower boundary was arbitrarily set to that proposed by Wald, and the other design parameters were chosen to satisfy the α and β error restrictions. They thought this method was the best of the three if early acceptance of H_0 was important. These same authors (1982) combined design features of the plans proposed by O'Brien and Fleming (1979) and DeMets and Ware (1980). They used the upper boundaries of O'Brien-Fleming, while the lower boundaries were essentially those they proposed earlier. These designs maintain the advantages of the O'Brien-Fleming plans while allowing for early termination if the new therapy is no better or actually worse than the standard.

Fleming (1982) simply modified the O'Brien-Fleming design to be used for one-sample one-sided clinical trials. He proposed using the O'Brien-Fleming upper boundary for rejection and the lower boundary for acceptance of H_0. The maximum sample size was that required by the fixed sample design, and the fixed sample critical value was used at the final stage. Like the DeMets-Ware (1982) plans, these allowed for early acceptance or rejection of H_0 while substantially, though not exactly, maintaining the power and level of significance of a fixed sample design.

A major endpoint in many clinical trials is survival time. For example, a study may be undertaken to determine if the addition of a new drug to a standard chemotherapy regimen increases survival in cancer patients. Recent work has extended group sequential theory to the design of trials in which survival is the major endpoint. For most survival analyses, the sample size denotes the expected number of deaths required rather than the actual number of patients (although they could be the same).

Canner (1977) used simulation techniques to develop stopping rules for long-term clinical trials analyzed at K fixed points in time. He assumed exponential lifetimes and used the test statistic:

$$Z_k = (h_1 - h_2) / [var(h_1 - h_2)]^{0.5},$$

where h_i denotes the estimate of the hazard rate for the i^{th} group and is based using all data accrued through the k^{th} stage, $k = 1,\ldots,K$. He considered two cases, namely:

(a) all patients were entered together, and

(b) uniform entry over time.

Canner also considered two stopping rules:

(i) equal α_i at each stage, and

(ii) linearly sloping boundaries.

Note that (i) is similar to the Pocock plan and (ii) is similar to the O'Brien and Fleming plan. His results for (i) using analyses at fixed time points were almost identical to those of Pocock where analyses were done after equal sample sizes. This provides evidence that Pocock's plans give adequate approximations for nonnormal response variables and are robust against modest departures from the assumption of equal samples at each stage. Canner found that his own designs were robust against the pattern of accrual, the study length, and the choice of a test statistic, but they were not robust against large departures from the model.

Tsiatis (1981,1982) has provided much of the theory for group sequential survival analysis. He determined the asymptotic joint distribution of the logrank statistic (1981) and later a class of sequentially computed statistics (1982) when analyses are undertaken at specified points in time. Gail, DeMets, and Slud (1982) present the results of simulation studies which applied the group sequential stopping rules of Haybittle, Pocock, and O'Brien and Fleming to the case of censored survival data using the logrank test statistic with analyses undertaken after a specified number of deaths. They found that the normal model performed adequately, especially under the null hypothesis. Trends over time such that healthier patients enter later and a strong treatment effect slightly increase the size of Pocock's designs. The other plans were more robust to these effects. Tsiatis, Rosner, and Tritchler (1985) show that the test for treatment effect in Cox's (1972) proportional hazards model has the same asymptotic distribution as the usual logrank test when data are analyzed at prespecified times. Their simulation results also indicate that group sequential designs developed using normal theory can be used when results are analyzed after fixed numbers of deaths rather than time.

Several extensions to group sequential designs have been developed. Methods have been proposed which allow the incorporation of cost (Gould and Pecore, 1982) and clinical opinion (Freedman, Lowe, and Macaskill, 1984) into the development of group sequential designs. Fairbanks and Madsen (1982) provide a method for determining overall p-values in trials involving repeated significance tests. Similarly, Peace and Schriver discuss p-value and power calculations in group sequential designs. Armitage, Stratton, and Worthington (1985) discuss the extension of group sequential methods to the case where the increments in accumulated responses are not independent.

Lan and DeMets (1983) proposed a method of determining group sequential boundaries in which the number of interim analyses need not be specified in advance. This method is based on a function f(t) which is increasing in t such that $f(1) = \alpha$. Note that f(t) represents the rate at which the type I error is used. Fleming, Harrington, and O'Brien (1984) developed group sequential designs similar to those of O'Brien and Fleming in that the critical value at the final stage is similar to that of a fixed sample design. However, the initial stage critical values are not as large. They chose boundaries such that $P[|Z_1| < r_1, \ldots , |Z_{j-1}| < r_{j-1}, |Z_j| > r_j] = p_j$ where $p_1 = p_2 = \ldots = p_{k-1} = p$, $p = u\alpha/(K-1)$, and Z_i is the standard normal variate calculated through the i^{th} stage. Here u represents the proportion of the type I error used up before the final review. These designs also allow the flexibility of increasing the number of interim analyses.

1.2.4 Other Approaches

A criticism sometimes leveled against the frequentist approach in designing clinical trials is that final conclusions are dependent upon the particular stopping rule chosen. For example, the nominal significance levels for both stages of a two-stage design are listed in Table 1.2 for several of the designs mentioned above. A p-value of 0.04 at each stage of the trial would lead to rejection of H_0 using two of the designs and acceptance using the other two.

Table 1.2. Nominal Significance Levels Two-Stage Designs (*)

Method	Nominal Significance Level α_1	α_2
Pocock (1977)	.0294	.0294
Canner Equal α_i (1977)	.0316	.0316
O'Brien - Fleming (1979)	.0051	.0475
Elashoff - Reedy (1984)	.0150	.0418

* Approximately equal sample sizes at each stage, overall level of significance = 0.05, acceptance of H_0 at final stage only.

This dependency on the stopping rule is anti-intuitive and disconcerting to many statisticians (and non-statisticians). Authors such as Birnbaum (1962), Anscombe (1954,1963), Edwards (1972), and Cornfield (1966a,1966b,1976) suggest that inference in clinical trials be based on the Likelihood Principle. They argue that relative merits of various hypotheses based on the likelihood function are determined entirely by the observed data and not on possibilities that might have occurred, but did not. That is, results based on the likelihood function are not affected by the choice of a stopping rule. However, it has been pointed out (eg, see Cornfield, 1966a) that repeated analysis using the likelihood criterion also leads to ultimate rejection of H_0. Some Bayesians suggest placing a nonzero prior probability on H_0 to eliminate this possibility (Cornfield, 1966a, 1966b). The problem then becomes one of choosing an appropriate (and defensible) prior.

Although in agreement with the likelihood approach, Dupont (1983) discards it as not being "part of the standard vocabulary of medical science." He recommends that the fixed sample p-value be used when deciding to stop trials, along with consideration of the medical plausibility of the hypotheses and the degree to which they were predicted in advance. From the discussion of his paper, there did not appear to be much support for his proposal even though this approach is probably common in practice.

Several other methods are available for monitoring the progress of clinical trials. Deterministic curtailment is a method whereby nonsequentially planned clinical trials can be terminated early. If during an interim analysis, the results indicate that the hypothesis is sure to be rejected (or accepted) regardless of the outcome of the patients yet to be accrued, then the trial can be terminated without affecting the error rates. Stochastic curtailment is an extension of deterministic curtailment whereby trials are terminated early if, with high probability, the rejection or acceptance of H_0 is assured. This technique has been discussed by Lan, Simon, and Halperin (1982) and Halperin et al (1982). Ware, Muller, and Braundwald (1985) apply this idea to terminating trials for 'futility', where the futility index is simply the probability that H_0 will be accepted given the current results.

Rubinstein and Gail (1982) give methods for monitoring survival studies in which accrual can be stopped, but follow-up continues on those already accrued until the prespecified number of deaths is reached. They show that the accrual can be dramatically decreased while only slightly increasing the trial duration.

Chatterjee and Sen (1973) developed theory which enabled the assessment of treatment effect via linear rank statistics in survival studies. They assumed a progressive censoring scheme and simultaneous entry of all patients. In most clinical trials, patients enter randomly over time.

Since later entries can alter the ranks of the failure times, the theory developed for simultaneous entry is not appropriate. However, Sen (1976), Majumdar and Sen (1978), and Bangdiwala and Sen (1985) extended the linear rank statistic to cover the staggered entry situation.

Descriptions and reviews of these and other methods can be found in Gail (1982), DeMets and Lan (1984), and Lan, DeMets, and Halperin (1984).

1.3 Discussion

For ethical and cost considerations, interim analyses are performed while data accumulate during a clinical trial. When these interim analyses are done using tests appropriate for fixed sample designs, the probability of falsely declaring a statistically significant treatment effect at some review is substantially increased. To adjust for multiple looks, smaller significance levels are often used for each review. This can lead to confusion when a seemingly significant sample statistic at the final stage fails to reject the null hypothesis. Designs are developed in this dissertation which permit multiple looks while allowing the fixed sample test to be used at the final stage.

Also for ethical and cost considerations, the number of patients required to adequately test the trial hypothesis should be kept as small as possible. To this end, it appears that one should use fully sequential designs. However, in actual practice one must weigh the advantage of a smaller expected sample size against the complications of continuous monitoring. Group sequential designs offer a good compromise. The group sequential designs presented in this research are those plans which minimize the expected sample size (ESS) using minimax and Bayes weighted average optimality criteria. Additionally, we require the restriction that the fixed sample critical value be used at the final analysis.

The number of unknown parameters in a K stage design is 3K-1. With two error restrictions, the number of free parameters is 3(K-1). With each additional stage, the numerical computations become more and more burdensome. A major emphasis will be on the development of two-stage designs. These plans illustrate the basic approach and are numerically feasible. Additionally, they can account for a large savings in sample size over the fixed sample design, and they can provide much of the savings possible from sequential designs (McPherson, 1982). Optimal restricted two-stage designs are presented in Chapter 2. An extension of these designs to three stages is given in Chapter 3. These two and three-stage designs will be compared to fixed sample and other group sequential designs in Chapter 4. Examples which illustrate the use of the proposed plans in the design and analysis of clinical trials are given in Chapter 5. A discussion of some possible extensions is given in Chapter 6.

II. OPTIMAL RESTRICTED TWO-STAGE DESIGNS

2.1 Introduction

Clinical trials are often designed to estimate a parameter θ and test the null hypothesis H_0: $\theta = \theta_0$ versus the alternative H_1: $\theta > \theta_0$ (or $\theta \neq \theta_0$). In this chapter, two-stage designs are developed for testing this hypothesis. The estimator of θ will be denoted by $\hat{\theta}$ and its variance by $\sigma^2_{\hat{\theta}} = \sigma^2/f(n)$, where σ^2 is a constant and $f(n)$ is a strictly increasing function of n. The majority of the results will be presented for $f(n) = n$; an example will be given in Chapter 5 for $f(n)$ not equal to n. The parameter θ might be the difference between two means or the coefficient of treatment effect in a proportional hazards or logistic regression model. The results in this chapter are based on normal theory and are exact if $\hat{\theta}$ is normally distributed with known variance; the results are approximate if $\hat{\theta}$ is asymptotically normal with variance proportional to $1/f(n)$.

2.2 Method

A two-stage design was described briefly in Chapter 1. In that chapter, the test statistic was left unspecified as it could be a cumulative sum, a mean, or any other appropriate statistic. The following description details a two-stage design which uses the normal (Wald) statistic on all data accumulated through the i^{th} stage, $i=1,2$. The two-stage design proceeds as follows:

Stage I - Accrue n_1 patients and calculate $Z_1 = (\hat{\theta} - \theta_0)/\sigma_{\hat{\theta}}$, where $\hat{\theta}$ is computed from data on the first n_1 patients. "Accept" H_0 if $Z_1 < C_1$; reject H_0 if $Z_1 > C_2$; otherwise, continue to the next stage.

Stage II - Accrue an additional n_2 patients. Let $n = n_1 + n_2$ and calculate $Z = (\hat{\theta} - \theta_0)/\sigma_{\hat{\theta}}$, where $\hat{\theta}$ is computed from data on all n patients. "Accept" H_0 if $Z < C_3$; otherwise, reject H_0.

Z_1 and Z are distributed normally with zero means and unit variances, and their joint distribution is bivariate normal with zero means, unit variances, and correlation coefficient $(n_1/n)^{1/2}$. The maximum sample size for the two-stage design is n and is realized whenever a second stage is necessary. Let $P_S(\theta)$ denote the probability that the trial will be stopped at the first stage. Then the expected sample size (ESS) of the two-stage design is given by:

$$ESS(\theta) = n_1 + n_2[1 - P_S(\theta)] = n[1 - (1 - p)P_S(\theta)], \qquad (2.1)$$

where $p = n_1/n$. For any given trial, the sample size will be n_1 with probability $P_S(\theta)$ and n with probability $1-P_S(\theta)$.

There are five unknown parameters in the two-stage design, namely: n_1, n_2, C_1, C_2, and C_3. The critical value at the second stage, C_3, will be set to equal that of the fixed sample test. The other four parameters of interest are chosen to satisfy the two equations:

$$\alpha = 1 - \Phi(C_2) + B(C_1, C_2; C_3, \infty; p) \tag{2.2}$$

$$1-\beta = 1 - \Phi(C_2 - u\sqrt{p}) + B(C_1 - u\sqrt{p}, C_2 - u\sqrt{p}; C_3 - u, \infty; p) \tag{2.3}$$

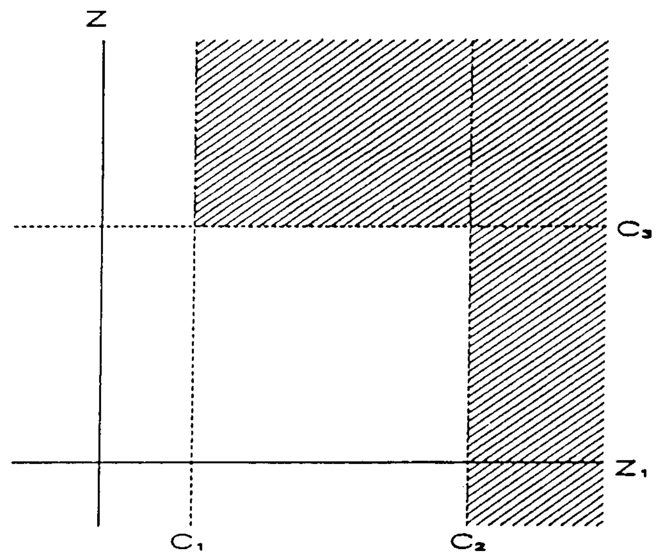
where $B(a,b;c,d;p)$

$= (1/2\pi\sqrt{1-p})\int_a^b\int_c^d \exp[-(1/2(1-p))(y^2 - 2\sqrt{p}yz + z^2)]dydz$,

$u = \sqrt{n}(\theta_1 - \theta_0)/\sigma$, and $\Phi(x)$ represents the standard normal distribution function. Equation (2.2) is simply the probability of rejecting H_0 at the first stage plus the probability of continuing the trial and rejecting H_0 at the second stage, assuming H_0 is true. This rejection region is shown as the shaded area in Figure 2.1(a). Equation (2.3) is this same probability under the alternative hypothesis. Let A denote the probability of accepting H_0 at the first stage and rejecting it at the second stage, had the trial continued. Let B denote the probability of rejecting H_0 at the first stage and accepting it at the second stage, had the trial continued. The restriction that $C_3 = \Phi^{-1}(1-\alpha)$ is equivalent to setting A equal to B. These areas are shown shaded in Figure 2.1(b).

Equations (2.2) and (2.3) are solved iteratively by numerical integration of the bivariate normal distribution using a double precision function written by Donnelly (1973). With five parameters and only three constraints,

(a)



(b)

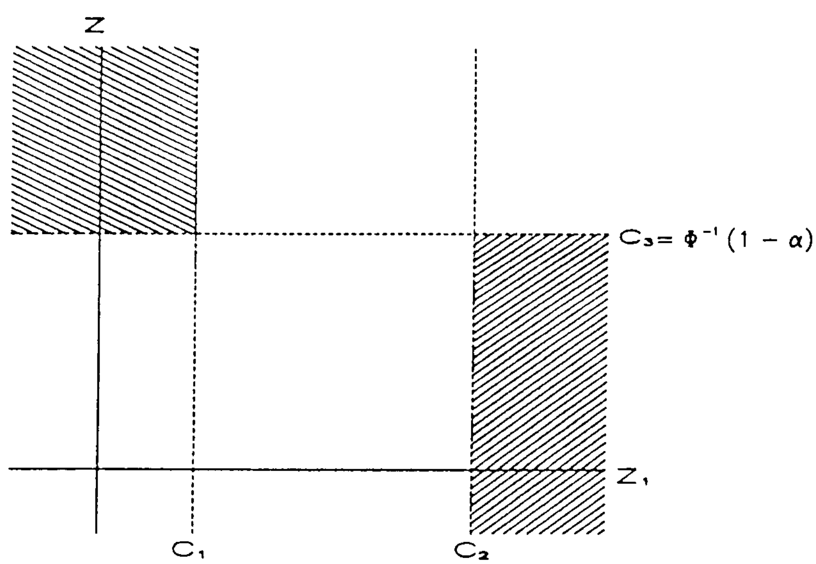


Figure 2.1. The shaded area in (a) represents the rejection region of a two-stage design for testing one-sided hypotheses. The shaded areas in (b) represent the probability of accepting H_0 at the first stage and rejecting it at the second, had the trial continued (\\\\\\\\\\) and the probability of rejecting H_0 at the first stage and accepting it at the second, had the trial continued (//////////). The restriction that $C_3 = \Phi^{-1}(1 - \alpha)$ is equivalent to setting these two areas equal to one another.

optimality criteria are used to determine the parameter values. The two optimality criteria used by Hald (1975) and De With (1983) are considered, namely:

(1) Minimax criterion - minimize the maximum ESS, i.e. minimize max ESS(θ) with respect to the five design parameters. We denote by θ_{max} that value of θ for which ESS(θ) is maximum. For the two-stage design presented above, $\theta_{max} = \theta_0 + (C_1 + C_2)\sigma/2\sqrt{n_1}$ (see section 2.7).

(2) Bayes weighted average criterion - minimize a weighted average of the ESS under H_0 and the ESS under H_1, i.e. minimize $ESS_w(\theta) = (1-w)ESS(\theta_0) + wESS(\theta_1)$. Using a weight of 0 for this criterion gives the the most efficient designs if the null hypothesis is true while a weight of 1 gives the most efficient designs if the specified alternative is true.

The probability of stopping at the first stage varies depending on the design parameters and the true value of θ. In general, $P_S(\theta) = \Phi(C_1 - v) + 1 - \Phi(C_2 - v)$, where $v = \sqrt{n_1}(\theta - \theta_0)/\sigma$. The algorithm used to obtain the optimal parameter values is outlined in section 2.3.

2.3 Algorithm

The type I and II errors (α and β) are initially fixed by the investigator, and $C_3 = \Phi^{-1}(1 - \alpha)$ is determined by inverse evaluation of the standard normal distribution. The parameters p, C_1, and C_2 are then chosen to obtain a type I error of $\alpha \pm \varepsilon$, where ε is arbitrarily set to 1×10^{-7}. For a given p, C_1 is selected and C_2 is found by a simple bisection bracketing method to obtain $\alpha \pm \varepsilon$ (Johnson and Riess, 1977). The value of $n\Delta^2 = n(\theta_1 - \theta_0)^2/\sigma^2$ necessary to achieve the required power ($\pm\ \varepsilon$) is also found by the bisection bracketing method. This gives the maximum sample size required for a particular Δ^2. The expected sample size (times Δ^2) is found as $[1 - (1-p)P_S(\theta)]n\Delta^2$, where $P_S(\theta)$ is defined in section 2.2. This function is minimized using a combination golden section search and parabolic interpolation (GSPI) subroutine (Brent, 1973) giving the optimal C_1 and C_2 for the particular p. The same GSPI algorithm is used to minimize the expected sample size over all p, thus giving optimal values for all the unknown parameters. This algorithm is shown schematically in Table 2.1. An interactive FORTRAN program which provides the optimal parameters for the restricted two-stage design is given in the Appendix.

Table 2.1. Outline of Parameter Selecting Algorithm

Step	Process
(1)	Choose α and β.
(2)	Calculate $C_3 = \Phi^{-1}(1 - \alpha)$.
(3)	Select bounds on p.
(4)	Determine starting value for p.
(5)	Select bounds on C_1.
(6)	Determine starting value for C_1.
(7)	Find C_2 to obtain a type I error of $\alpha \pm \varepsilon$.
(8)	Find $n\Delta^2$ to obtain a type II error of $\beta \pm \varepsilon$.
(9)	Calculate $ESS \times \Delta^2$ for particular criterion.
(10)	Minimum $ESS \times \Delta^2$ for given p? If no, go to (11). If yes, go to (12).
(11)	Determine new value for C_1; go to (7).
(12)	Minimum $ESS \times \Delta^2$ over all p? If no, go to (13). If yes, go to (14).
(13)	Determine new value for p; go to (5).
(14)	Print out optimal design parameters.

2.4 Results

The design parameters obtained using the different criteria are listed in Tables 2.2 - 2.5 for several values of α and β. Note that the sample sizes are all in units of standardized differences. The sample size required for a given clinical trial is obtained by dividing the tabled values by $\Delta^2 = [(\theta_1-\theta_0)/\sigma]^2$. The following notation is used: n_f denotes the sample size required using a fixed sample design; n denotes the maximum sample size required using the two-stage design; $R_m = ESS(\theta_{max})/n_f$; $R_0 = ESS(\theta_0)/n_f$; and $R_1 = ESS(\theta_1)/n_f$. The value of $P_S(\theta)$ listed in the tables is the probability of stopping at the first stage assuming that $\theta = \theta_{max}$, θ_0, or θ_1, depending on the particular criterion. For example, when minimizing the ESS under H_0, $P_S(\theta)$ is the probability of stopping early if $\theta = \theta_0$. As mentioned previously, the sample size for a particular trial will be n_1 with probability $P_S(\theta)$ and n with probability $1 - P_S(\theta)$.

Several observations can be made regarding Tables 2.2 - 2.5. The expected decrease in sample size obtained by using two-stage minimax designs remains relatively constant over the range of α and β considered. These designs result approximately in an expected 13% savings in sample size. In general, the greatest expected decrease in sample size occurs for designs which minimize the ESS under H_0, assuming

Table 2.2. Optimal Restricted Two-Stage One-Sided Designs (Minimax Criterion)

α	$1-\beta$	p	C_1	C_2	C_3	$P_S(\theta)$	n_f^a	n/n_f	R_m	R_0	R_1
.005	.7	.575	1.661	2.901	2.576	.535	9.611	1.121	.866	.666	.861
	.8	.593	1.654	2.919	2.576	.527	11.679	1.108	.870	.678	.849
	.9	.618	1.649	2.943	2.576	.518	14.879	1.093	.877	.696	.820
	.95	.640	1.646	2.963	2.576	.511	17.814	1.081	.882	.710	.793
	.99	.675	1.643	2.998	2.576	.498	24.031	1.064	.892	.735	.759
.01	.7	.564	1.437	2.669	2.326	.538	8.127	1.128	.863	.671	.858
	.8	.583	1.431	2.687	2.326	.530	10.036	1.114	.868	.683	.846
	.9	.610	1.424	2.712	2.326	.520	13.017	1.097	.875	.701	.815
	.95	.632	1.422	2.732	2.326	.513	15.770	1.084	.880	.715	.789
	.99	.670	1.419	2.766	2.326	.500	21.648	1.066	.890	.741	.755
.025	.7	.549	1.115	2.331	1.960	.543	6.172	1.139	.859	.688	.854
	.8	.570	1.107	2.349	1.960	.534	7.849	1.123	.864	.700	.840
	.9	.598	1.098	2.374	1.960	.523	10.507	1.104	.872	.717	.809
	.95	.622	1.094	2.394	1.960	.516	12.995	1.090	.878	.731	.782
	.99	.663	1.087	2.430	1.960	.502	18.372	1.069	.888	.756	.750
.05	.7	.535	0.838	2.043	1.645	.547	4.706	1.149	.856	.711	.849
	.8	.557	0.830	2.060	1.645	.538	6.183	1.131	.862	.722	.835
	.9	.588	0.819	2.086	1.645	.527	8.564	1.110	.869	.739	.803
	.95	.614	0.814	2.105	1.645	.519	10.822	1.095	.875	.752	.776
	.99	.657	0.808	2.139	1.645	.506	15.770	1.072	.886	.776	.745
.10	.7	.521	0.523	1.713	1.282	.552	3.261	1.160	.853	.747	.844
	.8	.544	0.514	1.730	1.282	.543	4.508	1.141	.858	.757	.829
	.9	.577	0.502	1.755	1.282	.531	6.569	1.117	.866	.771	.796
	.95	.604	0.493	1.776	1.282	.521	8.564	1.100	.873	.783	.770
	.99	.650	0.483	1.808	1.282	.508	13.017	1.076	.885	.804	.740

[a] Divide each value by Δ^2.

Table 2.3. Optimal Restricted Two-Stage One-Sided Designs (Weighted Average Criterion, W = 0)

α	$1-\beta$	p	C_1	C_2	C_3	$P_S(\theta)$	n_f^a	n/n_f	R_m	R_0	R_1
.005	.7	.215	1.002	2.881	2.576	.844	9.611	1.420	1.033	.479	1.019
	.8	.246	0.955	2.931	2.576	.832	11.679	1.343	1.016	.501	1.016
	.9	.289	0.901	3.000	2.576	.818	14.879	1.265	1.000	.529	.970
	.95	.322	0.864	3.057	2.576	.807	17.814	1.218	.992	.551	.906
	.99	.380	0.810	3.158	2.576	.792	24.031	1.156	.984	.589	.766
.01	.7	.235	0.897	2.660	2.326	.819	8.127	1.399	.994	.522	.989
	.8	.267	0.852	2.710	2.326	.806	10.036	1.325	.982	.542	.980
	.9	.311	0.799	2.778	2.326	.791	13.017	1.251	.973	.569	.930
	.95	.344	0.764	2.832	2.326	.780	15.770	1.206	.968	.590	.867
	.99	.403	0.712	2.929	2.326	.763	21.648	1.148	.965	.625	.737
.025	.7	.270	0.722	2.340	1.960	.775	6.172	1.358	.943	.590	.943
	.8	.303	0.681	2.387	1.960	.761	7.849	1.294	.939	.607	.929
	.9	.347	0.634	2.451	1.960	.744	10.507	1.227	.936	.631	.876
	.95	.381	0.601	2.502	1.960	.732	12.995	1.187	.936	.649	.817
	.99	.441	0.554	2.589	1.960	.715	18.372	1.134	.938	.681	.705
.05	.7	.306	0.554	2.067	1.645	.730	4.706	1.319	.907	.651	.907
	.8	.338	0.517	2.110	1.645	.715	6.183	1.264	.908	.666	.889
	.9	.382	0.474	2.168	1.645	.697	8.564	1.205	.910	.686	.838
	.95	.416	0.444	2.214	1.645	.685	10.822	1.169	.913	.702	.784
	.99	.475	0.400	2.292	1.645	.666	15.770	1.122	.919	.729	.689
.10	.7	.353	0.340	1.749	1.282	.673	3.261	1.274	.877	.719	.872
	.8	.384	0.311	1.785	1.282	.659	4.508	1.230	.881	.730	.853
	.9	.427	0.274	1.835	1.282	.641	6.569	1.181	.886	.747	.805
	.95	.460	0.249	1.873	1.282	.629	8.564	1.150	.892	.760	.758
	.99	.518	0.208	1.940	1.282	.609	13.017	1.108	.901	.783	.685

[a] Divide each value by Δ^2.

Table 2.4. Optimal Restricted Two-Stage One-Sided Designs (Weighted Average Criterion, W = 1/2)

α	$1-\beta$	p	C_1	C_2	C_3	$P_S(\theta)$	n_f^a	n/n_f	R_m	R_0	R_1
.005	.7	.343	1.075	2.980	2.576	.601	9.611	1.177	.913	.512	.913
	.8	.389	1.110	3.011	2.576	.611	11.679	1.147	.907	.539	.899
	.9	.431	1.143	3.039	2.576	.650	14.879	1.129	.908	.567	.856
	.95	.449	1.137	3.061	2.576	.689	17.814	1.124	.916	.583	.811
	.99	.557	1.049	3.120	2.576	.757	24.031	1.124	.941	.603	.721
.01	.7	.363	0.982	2.740	2.326	.610	8.127	1.185	.899	.551	.899
	.8	.404	1.005	2.767	2.326	.621	10.036	1.157	.896	.574	.883
	.9	.441	1.015	2.797	2.326	.657	13.017	1.138	.900	.599	.841
	.95	.454	0.993	2.822	2.326	.692	15.770	1.132	.910	.612	.797
	.99	.462	0.888	2.891	2.326	.750	21.648	1.127	.935	.632	.711
.025	.7	.389	0.810	2.390	1.960	.616	6.172	1.195	.881	.611	.880
	.8	.424	0.816	2.413	1.960	.627	7.849	1.169	.883	.629	.863
	.9	.453	0.801	2.446	1.960	.657	10.507	1.149	.891	.649	.822
	.95	.465	0.764	2.478	1.960	.687	12.995	1.140	.901	.661	.780
	.99	.470	0.645	2.559	1.960	.733	18.372	1.128	.926	.683	.698
.05	.7	.410	0.637	2.090	1.645	.615	4.706	1.200	.869	.665	.865
	.8	.440	0.630	2.113	1.645	.625	6.183	1.175	.873	.679	.848
	.9	.465	0.601	2.149	1.645	.650	8.564	1.154	.883	.695	.808
	.95	.474	0.555	2.186	1.645	.674	10.822	1.142	.893	.707	.768
	.99	.479	0.424	2.279	1.645	.710	15.770	1.125	.918	.729	.689
.10	.7	.434	0.409	1.748	1.282	.606	3.261	1.201	.859	.726	.852
	.8	.459	0.394	1.771	1.282	.615	4.508	1.177	.864	.737	.834
	.9	.478	0.352	1.810	1.282	.635	6.569	1.156	.875	.750	.795
	.95	.486	0.298	1.853	1.282	.654	8.564	1.142	.885	.761	.756
	.99	.489	0.159	1.960	1.282	.677	13.017	1.119	.909	.784	.681

[a] Divide each value by Δ^2.

Table 2.5. Optimal Restricted Two-Stage One-Sided Designs (Weighted Average Criterion, W = 1)

α	$1-\beta$	p	C_1	C_2	C_3	$P_S\theta$	n_f^a	n/n_f	R_m	R_0	R_1
.005	.7	.586	1.683	2.899	2.576	.553	9.611	1.117	.866	.675	.861
	.8	.597	1.661	2.919	2.576	.576	11.679	1.106	.870	.681	.849
	.9	.585	1.571	2.957	2.576	.618	14.879	1.101	.878	.669	.818
	.95	.565	1.469	2.996	2.576	.661	17.814	1.100	.887	.655	.784
	.99	.517	1.240	3.078	2.576	.744	24.031	1.110	.918	.631	.711
.01	.7	.575	1.459	2.666	2.326	.556	8.127	1.124	.863	.679	.858
	.8	.585	1.433	2.687	2.326	.579	10.036	1.113	.867	.683	.846
	.9	.573	1.344	2.729	2.326	.621	13.017	1.106	.876	.675	.813
	.95	.554	1.246	2.769	2.326	.664	15.770	1.106	.886	.664	.778
	.99	.507	1.026	2.857	2.326	.747	21.648	1.115	.918	.649	.705
.025	.7	.559	1.132	2.327	1.960	.562	6.172	1.135	.860	.694	.853
	.8	.567	1.102	2.351	1.960	.583	7.849	1.123	.864	.698	.840
	.9	.556	1.016	2.395	1.960	.625	10.507	1.116	.873	.693	.806
	.95	.536	0.922	2.439	1.960	.668	12.995	1.116	.884	.687	.771
	.99	.494	0.712	2.537	1.960	.750	18.372	1.123	.917	.687	.696
.05	.7	.543	0.850	2.040	1.645	.566	4.706	1.145	.857	.715	.849
	.8	.551	0.819	2.063	1.645	.587	6.183	1.134	.862	.719	.835
	.9	.540	0.737	2.111	1.645	.629	8.564	1.126	.871	.718	.800
	.95	.522	0.647	2.159	1.645	.672	10.822	1.125	.883	.718	.764
	.99	.482	0.445	2.266	1.645	.753	15.770	1.130	.918	.730	.689
.10	.7	.526	0.532	1.709	1.282	.572	3.261	1.159	.853	.749	.844
	.8	.532	0.498	1.736	1.282	.592	4.508	1.146	.859	.753	.829
	.9	.521	0.418	1.787	1.282	.634	6.569	1.138	.869	.757	.792
	.95	.504	0.332	1.840	1.282	.676	8.564	1.135	.882	.762	.755
	.99	.467	0.140	1.958	1.282	.756	13.017	1.139	.918	.786	.680

[a] Divide each value by Δ^2.

H_0 is true. Such designs lead to savings in sample size of approximately 25% to 50% for the α and β considered with expected savings greater for smaller α and larger β. These designs can also lead to a large maximum sample size for a single trial, in some cases giving sample sizes almost one and one-half times the fixed sample size. Additionally, designs optimized under H_0 sometimes result in expected sample sizes larger than the fixed sample size when $\theta = \theta_{max}$ or θ_1. The optimal choice of p is usually smaller and the probability of stopping early is usually greater for designs optimized under H_0. Minimizing the ESS under H_1 produces designs similar to those obtained using the minimax criterion for larger β. The ESS decreases under H_1 with decreasing β and increasing α. It is interesting that the ESS is usually smaller when $\theta = \theta_0$ even for designs optimized under H_1. The ESS is smaller for $\theta = \theta_1$ rather than θ_0 when α is large and β is small. Not surprisingly, the designs obtained using the weighted average criterion with w = 1/2 give values intermediate to those obtained using w = 0 and w = 1. These designs do nearly as well as the designs optimized under H_0 without resulting in such large possible sample sizes for individual trials. When $\theta = \theta_0$, the ESS for these designs is never more than 4% greater than the ESS for designs optimized under H_0. Additionally, the expected sample sizes when $\theta = \theta_1$ and $\theta = \theta_{max}$ are almost always smaller than their respective values for designs optimized under θ_0.

2.5 Other Restrictions

The choice of p is sometimes determined by factors unrelated to optimal designs. For some studies it might be practical to choose equal samples at each stage. For studies in which severe toxicity is possible, a small value of p might be chosen so that the data can be reviewed early in the trial. As mentioned by Elfring and Schultz (1973), some studies have secondary objectives for which a fixed number of patients is needed. For these studies p might have to be large so that the other goals can be reached. The expected and maximum sample sizes, obtained from designs optimized for different values of p, are plotted relative to the fixed sample size in Figure 2.2 for various α and β. For all of these designs, there is a range of p for which the ESS is fairly constant, affording some flexibility in the choice of this parameter without much loss in efficiency. The greatest maximum sample size occurs for p somewhat less than its optimal value and drops sharply thereafter with increasing p. A choice of p slightly larger than the optimal value gives almost the same ESS and a smaller maximum sample size for any given trial.

Tables 2.6 - 2.8 give optimal design parameters for p = 0.25, 0.5, and 0.75. Most of the same observations made earlier regarding Tables 2.2 - 2.5 can also be made for these tables. In addition, for the minimax designs and the designs optimized under H_1, the expected reduction in sample size is much greater for p = 0.5 and p = 0.75 compared to p

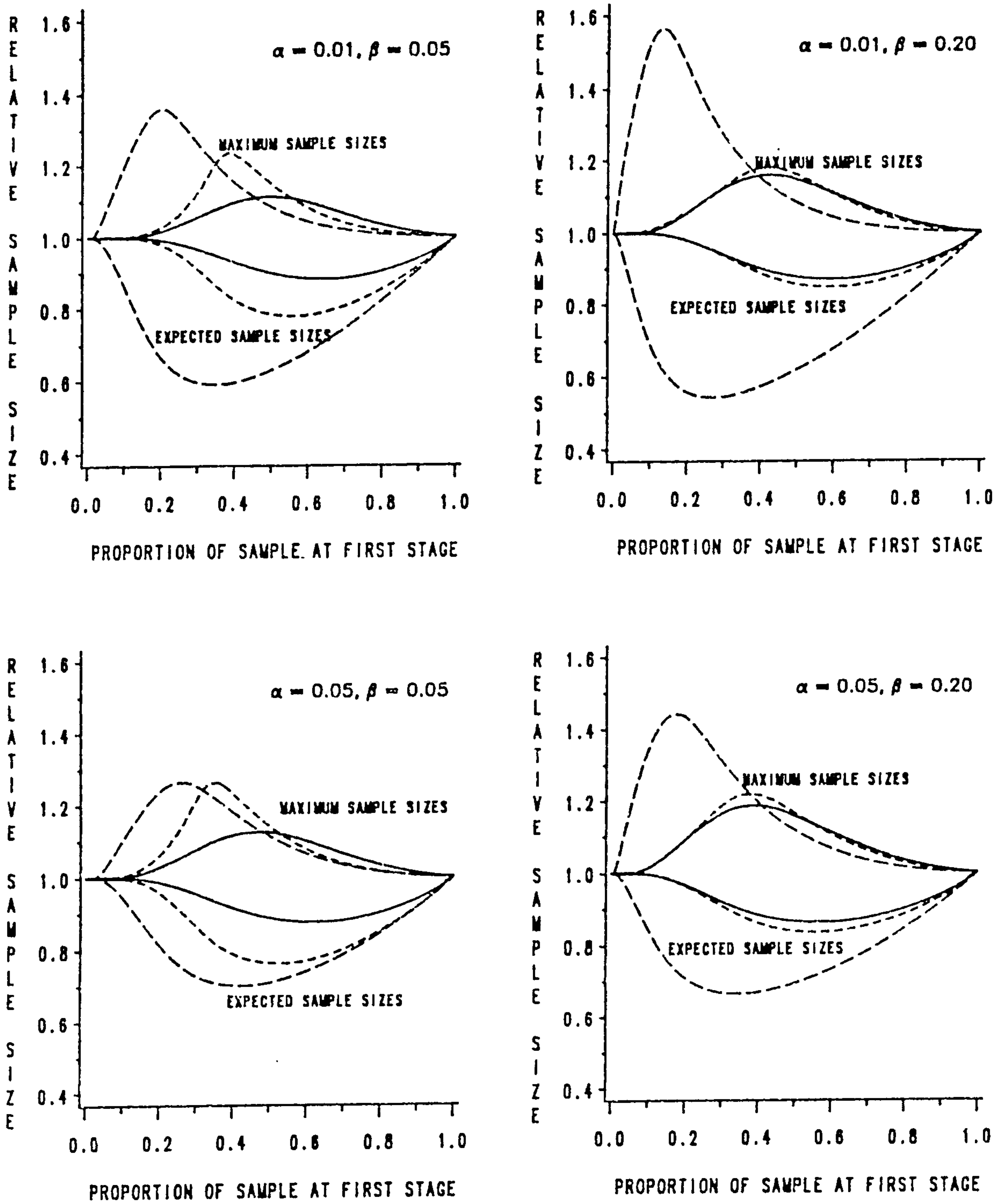


Figure 2.2. Expected and maximum sample sizes relative to the fixed sample size as functions of p, the proportion of the sample at the first stage.
—— minimax design assuming $\theta = \theta_{max}$;
---- design optimized for θ_1 assuming $\theta = \theta_1$;
— — design optimized for θ_0 assuming $\theta = \theta_0$.

Table 2.6. Optimal Restricted Two-Stage One-Sided Designs (p = .25)

α	$1-\beta$	p	C_1	C_2	C_3	$P_S(\Theta)$	n_f^a	n/n_f	R_m	R_0	R_1
Minimax Criterion											
.01	.8	.25	0.209	3.014	2.326	.161	10.036	1.091	.959	.613	.959
	.9	.25	-0.058	3.174	2.326	.106	13.017	1.056	.972	.678	.960
	.95	.25	-0.284	3.320	2.326	.072	15.770	1.036	.981	.734	.953
.025	.8	.25	0.070	2.672	1.960	.193	7.849	1.112	.951	.669	.949
	.9	.25	-0.197	2.841	1.960	.129	10.507	1.070	.966	.729	.947
	.95	.25	-0.420	2.992	1.960	.088	12.995	1.046	.977	.780	.938
.05	.8	.25	-0.062	2.382	1.645	.222	6.183	1.132	.943	.721	.938
	.9	.25	-0.324	2.558	1.645	.150	8.564	1.082	.961	.775	.934
	.95	.25	-0.555	2.723	1.645	.101	10.822	1.053	.973	.822	.924
Weighted Average Criterion (W = 0)											
.01	.8	.25	0.830	2.702	2.326	.800	10.036	1.358	1.002	.543	1.001
	.9	.25	0.693	2.763	2.326	.759	13.017	1.345	1.042	.579	1.010
	.95	.25	0.555	2.830	2.326	.713	15.770	1.325	1.072	.617	.992
.025	.8	.25	0.616	2.374	1.960	.740	7.849	1.380	.987	.614	.981
	.9	.25	0.459	2.452	1.960	.684	10.507	1.344	1.022	.654	.980
	.95	.25	0.298	2.539	1.960	.623	12.995	1.303	1.047	.695	.957
.05	.8	.25	0.406	2.105	1.645	.675	6.183	1.380	.971	.681	.961
	.9	.25	0.229	2.204	1.645	.604	8.564	1.323	1.002	.723	.952
	.95	.25	0.047	2.313	1.645	.529	10.822	1.266	1.021	.763	.926
Weighted Average Criterion (W = 1)											
.01	.8	.25	0.209	3.014	2.326	.161	10.036	1.091	.959	.613	.959
	.9	.25	0.003	3.136	2.326	.133	13.017	1.066	.972	.665	.960
	.95	.25	-0.091	3.195	2.326	.142	15.770	1.063	.983	.693	.950
.025	.8	.25	0.076	2.668	1.960	.197	7.849	1.114	.951	.668	.949
	.9	.25	-0.096	2.775	1.960	.176	10.507	1.091	.967	.711	.946
	.95	.25	-0.167	2.821	1.960	.195	12.995	1.093	.982	.736	.934
.05	.8	.25	-0.040	2.368	1.645	.235	6.183	1.130	.943	.718	.938
	.9	.25	-0.185	2.463	1.645	.222	8.564	1.118	.963	.755	.933
	.95	.25	-0.253	2.508	1.645	.245	10.822	1.123	.982	.781	.917

[a] Divide each value by Δ^2.

Table 2.7. Optimal Restricted Two-Stage One-Sided Designs (p = .50)

α	$1-\beta$	p	C_1	C_2	C_3	$P_S(\Theta)$	n_f^a	n/n_f	R_m	R_0	R_1
Minimax Criterion											
.01	.8	.50	1.279	2.706	2.326	.476	10.036	1.146	.874	.629	.853
	.9	.50	1.175	2.762	2.326	.428	13.017	1.126	.885	.629	.822
	.95	.50	1.083	2.815	2.326	.386	15.770	1.111	.896	.632	.785
.025	.8	.50	0.998	2.366	1.960	.494	7.849	1.153	.868	.663	.845
	.9	.50	0.906	2.421	1.960	.449	10.507	1.135	.880	.666	.811
	.95	.50	0.813	2.478	1.960	.405	12.995	1.118	.891	.671	.774
.05	.8	.50	0.750	2.078	1.645	.507	6.183	1.157	.864	.699	.837
	.9	.50	0.667	2.130	1.645	.464	8.564	1.141	.876	.705	.802
	.95	.50	0.583	2.186	1.645	.423	10.822	1.125	.887	.712	.765
Weighted Average Criterion (W = 0)											
.01	.8	.50	1.052	2.833	2.326	.856	10.036	1.081	.880	.619	.861
	.9	.50	1.014	2.856	2.326	.847	13.017	1.082	.889	.624	.828
	.95	.50	0.975	2.880	2.326	.837	15.770	1.082	.900	.629	.789
.025	.8	.50	0.836	2.464	1.960	.805	7.849	1.101	.872	.657	.849
	.9	.50	0.795	2.490	1.960	.793	10.507	1.100	.882	.664	.814
	.95	.50	0.754	2.516	1.960	.781	12.995	1.100	.892	.671	.775
.05	.8	.50	0.638	2.150	1.645	.754	6.183	1.117	.866	.696	.840
	.9	.50	0.595	2.178	1.645	.739	8.564	1.116	.877	.704	.804
	.95	.50	0.548	2.210	1.645	.722	10.822	1.113	.887	.711	.765
Weighted Average Criterion (W = 1)											
.01	.8	.50	1.310	2.690	2.326	.525	10.036	1.157	.874	.632	.853
	.9	.50	1.253	2.720	2.326	.576	13.017	1.153	.886	.635	.821
	.95	.50	1.189	2.755	2.326	.634	15.770	1.146	.898	.639	.783
.025	.8	.50	1.024	2.351	1.960	.548	7.849	1.163	.868	.665	.845
	.9	.50	0.957	2.390	1.960	.595	10.507	1.154	.880	.670	.811
	.95	.50	0.888	2.432	1.960	.649	12.995	1.144	.892	.675	.773
.05	.8	.50	0.768	2.066	1.645	.562	6.183	1.165	.864	.700	.837
	.9	.50	0.700	2.109	1.645	.609	8.564	1.153	.876	.706	.802
	.95	.50	0.631	2.154	1.645	.662	10.822	1.142	.887	.713	.764

[a] Divide each value by Δ^2.

Table 2.8. Optimal Restricted Two-Stage One-Sided Designs (p = .75)

α	$1-\beta$	p	C_1	C_2	C_3	$P_S(\Theta)$	n_f^a	n/n_f	R_m	R_0	R_1
Minimax Criterion											
.01	.8	.75	1.662	2.672	2.326	.614	10.036	1.049	.888	.798	.871
	.9	.75	1.643	2.684	2.326	.603	13.017	1.048	.890	.798	.850
	.95	.75	1.625	2.697	2.326	.592	15.770	1.047	.892	.798	.831
.025	.8	.75	1.325	2.326	1.960	.617	7.849	1.049	.887	.809	.869
	.9	.75	1.310	2.337	1.960	.608	10.507	1.048	.889	.809	.848
	.95	.75	1.294	2.348	1.960	.598	12.995	1.048	.891	.809	.830
.05	.8	.75	1.037	2.027	1.645	.621	6.183	1.050	.887	.821	.868
	.9	.75	1.022	2.039	1.645	.611	8.564	1.049	.889	.821	.847
	.95	.75	1.005	2.051	1.645	.601	10.822	1.048	.890	.822	.829
Weighted Average Criterion (W = 0)											
.01	.8	.75	1.322	2.918	2.326	.909	10.036	1.012	.904	.782	.882
	.9	.75	1.313	2.925	2.326	.907	13.017	1.012	.906	.782	.855
	.95	.75	1.307	2.930	2.326	.906	15.770	1.012	.906	.783	.830
.025	.8	.75	1.069	2.518	1.960	.863	7.849	1.017	.898	.797	.876
	.9	.75	1.061	2.524	1.960	.861	10.507	1.017	.899	.798	.850
	.95	.75	1.051	2.532	1.960	.859	12.995	1.017	.900	.799	.827
.05	.8	.75	0.847	2.175	1.645	.816	6.183	1.023	.893	.814	.871
	.9	.75	0.837	2.184	1.645	.813	8.564	1.022	.895	.815	.846
	.95	.75	0.825	2.193	1.645	.810	10.822	1.022	.896	.815	.824
Weighted Average Criterion (W = 1)											
.01	.8	.75	1.619	2.701	2.326	.658	10.036	1.042	.888	.794	.870
	.9	.75	1.544	2.753	2.326	.715	13.017	1.033	.892	.789	.848
	.95	.75	1.468	2.807	2.326	.772	15.770	1.025	.896	.786	.827
.025	.8	.75	1.284	2.355	1.960	.665	7.849	1.042	.888	.805	.869
	.9	.75	1.208	2.411	1.960	.720	10.507	1.032	.891	.801	.846
	.95	.75	1.133	2.468	1.960	.777	12.995	1.025	.895	.800	.826
.05	.8	.75	0.993	2.061	1.645	.668	6.183	1.042	.887	.818	.868
	.9	.75	0.918	2.119	1.645	.725	8.564	1.032	.891	.816	.845
	.95	.75	0.846	2.176	1.645	.781	10.822	1.024	.895	.815	.824

[a] Divide each value by Δ^2.

= 0.25. Designs optimized under H_0 are more efficient for p = 0.25 and p = 0.5.

Owen developed two-stage designs such that the total sample size was constrained to equal that of a fixed sample design. He fixed α and obtained the power for several alternatives. He showed that one could develop two-stage designs which result in expected savings in sample size with only slight reductions in power.

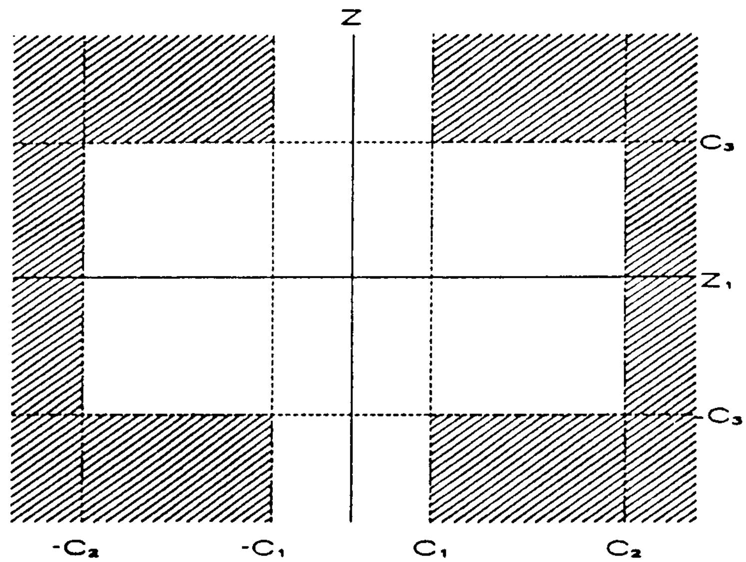
Let $1-\beta$ denote the power of a single stage design with a type I error level of α and a sample size of n_f. Also let Power(n_f) denote the power of an α-level two-stage design with the total sample size constrained to equal the fixed sample size. Then Power(n_f) $\leq 1-\beta$. Additionally, Power(n_f) $\rightarrow 1-\beta$ as $p \rightarrow 1$, and Power(n_f) $\rightarrow 1-\beta$ as $C_1 \rightarrow -\infty$ and $C_2 \rightarrow \infty$. These results are illustrated below for $\alpha = 0.05$ using a sample size such that a single stage design would have a power of 0.80.

	p = .1		p = .5		p = .9	
C_1	C_2	Power	C_2	Power	C_2	Power
1.5	1.67	.232	1.70	.590	1.75	.781
1.0	1.77	.380	1.93	.711	2.17	.799
0.5	1.94	.535	2.24	.773	2.64	.800
0.0	2.16	.662	2.62	.794	3.13	.800
-0.5	2.45	.742	3.03	.799	3.63	.800
-1.0	2.78	.780	3.48	.800	3.90	.800
-1.5	3.14	.795	3.94	.800	3.90	.800

2.6 Two-sided Results

The designs presented thus far in this chapter have all been for one-sided hypotheses. A two-stage design for testing a two-sided hypothesis is illustrated in Figure 2.3. Approximate two-sided α-level plans can be achieved by comparing the absolute value of the test statistic with the absolute values of the cut points obtained for one-sided $\alpha/2$ designs. Such designs give approximately correct α and β levels (α is somewhat larger and β is somewhat smaller than their nominal values; see Figure 2.3). However, the tabled values for the probability of stopping early and the ESS are no longer correct. In addition, these designs will not be optimal, especially under the null hypothesis. They are more nearly optimal for the minimax designs and the designs optimized under the alternative hypothesis. Under H_0, the probability of stopping early decreases by $\Phi(C_2) - \Phi(C_1)$, and the ESS increases by $n[(1-p)(\Phi(C_2) - \Phi(C_1))]$. The value of θ_{max} for a two-sided design is no longer $\theta_0 + (C_1 + C_2)\sigma/2\sqrt{n_1}$; instead, it is $\theta_0 + v\sigma/\sqrt{n_1}$ where v is the value which minimizes $1 - \Phi(C_2-v) + \Phi(C_1-v) - \Phi(-C_1-v) + \Phi(-C_2-v)$ and must be obtained numerically (see section 2.8). For most of the designs considered below, θ_{max} can be approximated by $\theta_0 + (C_1 + C_2)\sigma/2\sqrt{n_1}$, its value for the one-sided designs.

(a)



(b)

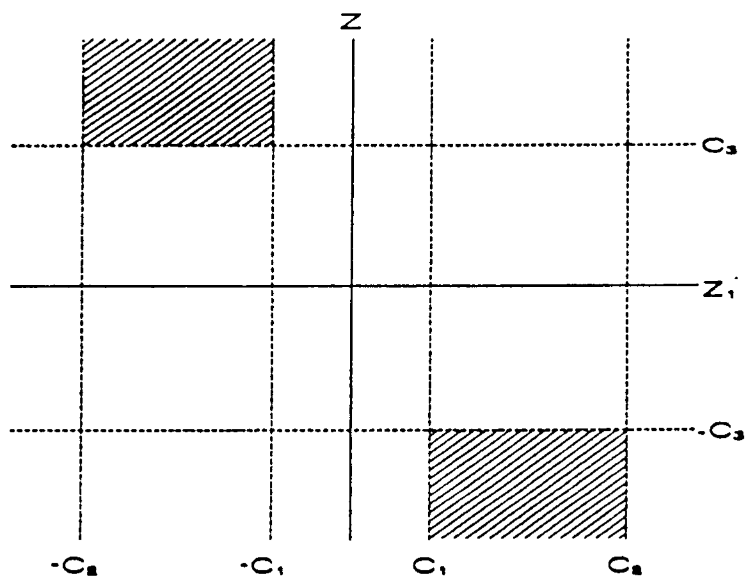


Figure 2.3. The shaded area in (a) represents the rejection region of a two-stage design for testing a two-sided hypothesis. The shaded areas in (b) represent probabilities of rejection which are additional to the rejection region obtained by using a two-stage $\alpha/2$ level design for testing one-sided hypotheses and its reflection as a two-sided α level design.

Design parameters for two-stage tests of two-sided hypotheses obtained using the different criteria are listed in Tables 2.9 - 2.12 for the α and β previously considered. The most striking feature of these tables is the similarity of the two-sided α-level plans and their respective one-sided $\alpha/2$-level plans when $\theta = \theta_{max}$ and $\theta = \theta_1$. The optimal parameter values are similar and the expected sample sizes almost identical for the one and two-sided designs. However, when $\theta = \theta_0$, the ESS for the two-sided plan is invariably greater than its respective one-sided value. This increase is greater than 15% in some cases.

2.7 Determination of θ_{max}

Hald (1975) stated without proof that θ_{max} for the optimal unrestricted two-stage one-sided design was $(h_1 + h_2)/2$, where $h_i = C_i\sigma/\sqrt{n_1} + \theta_0$, $i=1,2$. This is also the value of θ_{max} for the restricted two-stage design since it does not depend on C_3. θ_{max} is found simply by maximizing the expected sample size, or equivalently, minimizing the probability of stopping early. Rather than solving for θ_{max} directly, we solve for v such that $\sqrt{n_1}(\theta_{max} - \theta_0)/\sigma = v$, or $\theta_{max} = \theta_0 + v\sigma/\sqrt{n_1}$. We minimize $P_S(\theta) = 1 - \Phi(C_2-v) + \Phi(C_1-v) = 1 - \int_a^b(1/\sqrt{2\pi})\exp(-x^2/2)dx$, where $a = C_1-v$ and $b = C_2-v$.

By a simple application of Leibnitz' rule,

$dP_S(\theta)/dv = 1/\sqrt{2\pi}\exp(-(C_2-v)^2/2) - 1/\sqrt{2\pi}\exp(-(C_1-v)^2/2)$.

Setting this expression equal to zero and simplifying, we

Table 2.9. Optimal Restricted Two-Stage Two-Sided Designs (Minimax Criterion)

α	$1-\beta$	p	C_1	C_2	C_3	$P_S(\Theta)$	n_f^a	n/n_f	R_m	R_0	R_1
.005	.7	.584	1.868	3.117	2.807	.532	11.098	1.115	.868	.679	.864
	.8	.601	1.862	3.134	2.807	.525	13.313	1.103	.872	.690	.853
	.9	.626	1.858	3.158	2.807	.516	16.717	1.089	.878	.706	.824
	.95	.646	1.856	3.178	2.807	.509	19.819	1.077	.884	.720	.797
	.99	.681	1.856	3.212	2.807	.498	26.352	1.061	.893	.744	.763
.01	.7	.574	1.660	2.901	2.576	.535	9.611	1.121	.866	.688	.861
	.8	.593	1.655	2.918	2.576	.528	11.679	1.108	.870	.699	.849
	.9	.618	1.649	2.943	2.576	.518	14.879	1.092	.877	.715	.820
	.95	.640	1.648	2.962	2.576	.511	17.814	1.081	.882	.729	.793
	.99	.677	1.648	2.996	2.576	.500	24.031	1.063	.892	.753	.760
.025	.7	.561	1.366	2.589	2.241	.541	7.650	1.131	.862	.715	.857
	.8	.580	1.358	2.607	2.241	.532	9.505	1.117	.867	.725	.844
	.9	.609	1.354	2.631	2.241	.523	12.411	1.098	.874	.741	.814
	.95	.632	1.351	2.651	2.241	.515	15.103	1.086	.880	.753	.788
	.99	.670	1.348	2.685	2.241	.503	20.864	1.067	.890	.775	.755
.05	.7	.554	1.133	2.324	1.960	.549	6.172	1.140	.861	.752	.854
	.8	.575	1.126	2.341	1.960	.541	7.849	1.124	.865	.761	.840
	.9	.604	1.117	2.366	1.960	.530	10.507	1.104	.873	.775	.810
	.95	.628	1.113	2.386	1.960	.522	12.995	1.091	.878	.786	.784
	.99	.668	1.108	2.419	1.960	.510	18.372	1.070	.889	.805	.753
.10	.7	.558	0.905	2.014	1.645	.570	4.706	1.152	.862	.806	.852
	.8	.580	0.896	2.031	1.645	.561	6.183	1.134	.867	.814	.837
	.9	.611	0.886	2.055	1.645	.549	8.564	1.112	.874	.824	.808
	.95	.634	0.878	2.074	1.645	.540	10.822	1.097	.880	.832	.784
	.99	.675	0.869	2.106	1.645	.526	15.770	1.074	.890	.847	.758

[a] Divide each value by Δ^2.

Table 2.10. Optimal Restricted Two-Stage Two-Sided Designs (Weighted Average Criterion, W = 0)

α	$1-\beta$	p	C_1	C_2	C_3	$P_S(\theta)$	n_f[a]	n/n_f	R_m	R_0	R_1
.005	.7	.284	1.426	3.057	2.807	.848	11.098	1.401	.985	.550	.981
	.8	.318	1.394	3.097	2.807	.839	13.313	1.330	.973	.570	.969
	.9	.362	1.358	3.152	2.807	.827	16.717	1.259	.962	.595	.915
	.95	.396	1.333	3.197	2.807	.819	19.819	1.215	.957	.614	.851
	.99	.453	1.296	3.278	2.807	.806	26.352	1.157	.954	.647	.729
.01	.7	.314	1.343	2.844	2.576	.825	9.611	1.376	.949	.597	.948
	.8	.348	1.313	2.882	2.576	.815	11.679	1.311	.941	.615	.933
	.9	.393	1.279	2.935	2.576	.803	14.879	1.244	.936	.638	.878
	.95	.427	1.256	2.977	2.576	.794	17.814	1.202	.934	.655	.818
	.99	.483	1.222	3.052	2.576	.780	24.031	1.148	.935	.685	.713
.025	.7	.367	1.208	2.535	2.241	.784	7.650	1.332	.905	.670	.904
	.8	.402	1.183	2.570	2.241	.773	9.505	1.274	.903	.685	.885
	.9	.445	1.154	2.617	2.241	.760	12.411	1.216	.904	.704	.835
	.95	.478	1.133	2.654	2.241	.751	15.103	1.181	.906	.718	.785
	.99	.533	1.103	2.719	2.241	.736	20.864	1.133	.912	.743	.707
.05	.7	.422	1.082	2.272	1.960	.744	6.172	1.287	.879	.734	.875
	.8	.455	1.062	2.302	1.960	.733	7.849	1.241	.881	.745	.856
	.9	.497	1.037	2.342	1.960	.719	10.507	1.191	.885	.760	.812
	.95	.529	1.019	2.374	1.960	.710	12.995	1.160	.890	.772	.773
	.99	.580	0.993	2.429	1.960	.694	18.372	1.119	.898	.793	.718
.10	.7	.496	0.927	1.970	1.645	.695	4.706	1.235	.866	.803	.856
	.8	.525	0.912	1.993	1.645	.684	6.183	1.201	.870	.810	.839
	.9	.564	0.894	2.024	1.645	.672	8.564	1.162	.877	.821	.805
	.95	.593	0.880	2.048	1.645	.662	10.822	1.136	.882	.830	.777
	.99	.640	0.860	2.090	1.645	.647	15.770	1.102	.892	.845	.745

[a] Divide each value by Δ^2.

Table 2.11. Optimal Restricted Two-Stage Two-Sided Designs (Weighted Average Criterion, W = 1/2)

α	$1-\beta$	p	C_1	C_2	C_3	$P_S(\Theta)$	n_f^a	n/n_f	R_m	R_0	R_1
.005	.7	.404	1.462	3.148	2.807	.629	11.098	1.177	.897	.575	.897
	.8	.442	1.481	3.175	2.807	.638	13.313	1.150	.895	.596	.884
	.9	.474	1.484	3.205	2.807	.673	16.717	1.132	.900	.618	.845
	.95	.485	1.463	3.228	2.807	.709	19.819	1.128	.909	.630	.803
	.99	.490	1.367	3.283	2.807	.769	26.352	1.127	.933	.650	.719
.01	.7	.426	1.375	2.920	2.576	.637	9.611	1.184	.885	.617	.885
	.8	.459	1.382	2.945	2.576	.648	11.679	1.159	.886	.634	.871
	.9	.486	1.371	2.975	2.576	.679	14.879	1.140	.893	.653	.832
	.95	.496	1.340	3.001	2.576	.711	17.814	1.135	.902	.664	.792
	.99	.500	1.235	3.062	2.576	.763	24.031	1.129	.926	.686	.712
.025	.7	.458	1.227	2.589	2.241	.644	7.650	1.191	.872	.681	.869
	.8	.486	1.220	2.613	2.241	.654	9.505	1.167	.875	.694	.854
	.9	.506	1.192	2.644	2.241	.680	12.411	1.149	.884	.709	.817
	.95	.514	1.151	2.673	2.241	.705	15.103	1.141	.894	.720	.779
	.99	.519	1.040	2.743	2.241	.744	20.864	1.129	.915	.744	.705
.05	.7	.489	1.084	2.311	1.960	.644	6.172	1.191	.864	.739	.859
	.8	.510	1.068	2.333	1.960	.651	7.849	1.169	.869	.749	.844
	.9	.528	1.031	2.366	1.960	.673	10.507	1.151	.878	.762	.808
	.95	.533	0.985	2.398	1.960	.693	12.995	1.141	.888	.773	.771
	.99	.538	0.869	2.475	1.960	.720	18.372	1.125	.908	.798	.703
.10	.7	.527	0.907	1.997	1.645	.638	4.706	1.186	.863	.804	.852
	.8	.544	0.884	2.019	1.645	.643	6.183	1.166	.868	.811	.837
	.9	.555	0.838	2.055	1.645	.658	8.564	1.148	.877	.822	.802
	.95	.560	0.788	2.090	1.645	.671	10.822	1.136	.886	.833	.768
	.99	.564	0.672	2.173	1.645	.685	15.770	1.116	.906	.859	.706

[a] Divide each value by Δ^2.

Table 2.12. Optimal Restricted Two-Stage Two-Sided Designs (Weighted Average Criterion, W = 1)

α	$1-\beta$	p	C_1	C_2	C_3	$P_S(\theta)$	n_f^a	n/n_f	R_m	R_0	R_1
.005	.7	.596	1.893	3.115	2.807	.550	11.098	1.111	.868	.687	.864
	.8	.607	1.872	3.134	2.807	.573	13.313	1.101	.872	.694	.853
	.9	.596	1.780	3.172	2.807	.615	16.717	1.094	.879	.685	.822
	.95	.575	1.678	3.207	2.807	.660	19.819	1.096	.888	.672	.788
	.99	.525	1.442	3.283	2.807	.742	26.352	1.106	.918	.658	.716
.01	.7	.587	1.686	2.898	2.576	.554	9.611	1.117	.866	.696	.861
	.8	.596	1.660	2.918	2.576	.576	11.679	1.107	.870	.701	.849
	.9	.584	1.569	2.958	2.576	.618	14.879	1.100	.878	.695	.818
	.95	.565	1.470	2.996	2.576	.662	17.814	1.100	.887	.688	.784
	.99	.518	1.242	3.078	2.576	.745	24.031	1.109	.917	.688	.711
.025	.7	.572	1.385	2.587	2.241	.559	7.650	1.126	.863	.720	.857
	.8	.580	1.357	2.608	2.241	.580	9.505	1.116	.867	.725	.844
	.9	.569	1.268	2.651	2.241	.622	12.411	1.109	.875	.725	.812
	.95	.549	1.170	2.692	2.241	.664	15.103	1.109	.886	.726	.777
	.99	.504	0.952	2.783	2.241	.747	20.864	1.117	.919	.749	.703
.05	.7	.561	1.140	2.324	1.960	.565	6.172	1.136	.861	.754	.854
	.8	.569	1.107	2.349	1.960	.585	7.849	1.124	.865	.760	.840
	.9	.556	1.018	2.395	1.960	.626	10.507	1.116	.875	.766	.807
	.95	.537	0.924	2.439	1.960	.668	12.995	1.116	.887	.775	.771
	.99	.495	0.718	2.536	1.960	.752	18.372	1.123	.923	.818	.697
.10	.7	.555	0.885	2.025	1.645	.579	4.706	1.147	.862	.807	.851
	.8	.557	0.840	2.055	1.645	.594	6.183	1.135	.867	.814	.836
	.9	.543	0.748	2.107	1.645	.633	8.564	1.127	.879	.828	.801
	.95	.525	0.661	2.155	1.645	.676	10.822	1.125	.894	.846	.764
	.99	.489	0.481	2.262	1.645	.762	15.770	1.131	.938	.904	.691

[a] Divide each value by Δ^2.

get:

$$(C_2-v)^2 = (C_1-v)^2,$$
$$C_2^2 - 2C_2v = C_1^2 - 2C_1v,$$
$$v = (C_1^2 - C_2^2)/2(C_1 - C_2) = (C_1 + C_2)/2,$$
$$\text{thus } \theta_{max} = \theta_0 + (C_1 + C_2)\sigma/2\sqrt{n_1}.$$

The value of θ_{max} for the two-stage two-sided design can be obtained similarly. Here, $\theta_{max} = \theta_0 + v\sigma/\sqrt{n_1}$ where v is obtained by minimizing

$$P_S(\theta) = 1 - \Phi(C_2-v) + \Phi(-C_2-v) + \Phi(C_1-v) - \Phi(-C_1-v)$$
$$= 1 - \int_a^b(1/\sqrt{2\pi})\exp(-x^2/2)dx - \int_c^d(1/\sqrt{2\pi})\exp(-x^2/2)dx,$$

where $a = -C_2-v$, $b = -C_1-v$, $c = C_1-v$, and $d = C_2-v$.

Again using Leibnitz' rule,

$$dP_S(\theta)/dv = -1/\sqrt{2\pi}[\exp(-(-C_1-v)^2/2) - \exp(-(-C_2-v)^2/2) + \exp(-(C_2-v)^2/2) - \exp(-(C_1-v)^2/2)].$$

Setting this expression equal to zero and simplifying, we get:

$$\exp(-(C_1-v)^2/2) - \exp(-(-C_1-v)^2/2) = \exp(-(C_2-v)^2/2) - \exp(-(-C_2-v)^2/2)$$

$$\exp(-C_1^2/2)[\exp(C_1v) - \exp(-C_1v)] = \exp(-C_2^2/2)[\exp(C_2v) - \exp(-C_2v)]$$

$$\exp(-C_1^2/2)/\exp(-C_2^2/2) = [\exp(C_2v) - \exp(-C_2v)]/[\exp(C_1v) - \exp(-C_1v)]$$

$$\exp(C_2^2/2 - C_1^2/2) = [(\exp(2C_2v) - 1)\exp(C_1v)]/[(\exp(2C_1v) - 1)\exp(C_2v)]$$

$$(C_2^2 - C_1^2)/2 = (C_1 - C_2)v + \ln[(\exp(2C_2v) - 1)/(\exp(2C_1v) - 1)]$$

$$-(C_1 + C_2)/2 =$$

$$v + \ln[(\exp(2C_2v) - 1)/(\exp(2C_1v) - 1)]/(C_1 - C_2)$$

This expression has no explicit analytic solution and must be solved numerically.

The value of θ_{max} can be given approximately whenever $\exp(2C_2v)$ and $\exp(2C_1v) >> 1$. In this case,

$$-(C_1 + C_2)/2 \simeq v + \ln[(\exp(2C_2v))/(\exp(2C_1v))]/(C_1 - C_2)$$

$$-(C_1 + C_2)/2 \simeq v - 2v(C_1 - C_2)/(C_1 - C_2)$$

$$(C_1 + C_2)/2 \simeq v.$$

Thus $\theta_{max} \simeq \theta_0 + (C_1 + C_2)\sigma/2\sqrt{n_1}$ whenever the condition mentioned above is satisfied. This condition holds for most of the designs presented in the tables. However, the iterative solutions were used to obtain the parameter values listed in the tables.

2.8 Discussion

The two-stage design represents the simplest form of a group sequential design. Yet, with only one additional look at the data, the expected sample size can be reduced to as little as 50% of the fixed sample size. In general, the expected savings are not so great, and savings of 15% to 30% are more typical. This expected decrease in sample size would seem to warrant the additional look. Other two-stage designs have been presented in the literature. Many of these were mentioned in Chapter 1. A comparison of some of those designs to the designs developed in this chapter will be given in Chapter 4. Three-stage designs will be

III. OPTIMAL RESTRICTED THREE-STAGE DESIGNS

3.1 Introduction

The two-stage designs examined in the previous chapter represent the simplest form of group sequential plans. In this chapter, an extension of the previous designs to three stages is considered. The addition of a third stage adds three more parameters to the problem, making a complete numerical solution prohibitive. The majority of the results in this chapter will be obtained for designs with equal sample sizes at each stage. Some calculations will demonstrate the loss resulting from this restriction.

3.2 Method

The three-stage design proceeds as follows:
Stage I - Accrue n_1 patients and calculate $Z_1 = (\hat{\theta} - \theta_0)/\sigma_{\hat{\theta}}$, where $\hat{\theta}$ is computed from data on the first n_1 patients. "Accept" H_0 if $Z_1 < C_1$; reject H_0 if $Z_1 > C_2$; otherwise, continue to the next stage.
Stage II - Accrue an additional n_2 patients and calculate $Z_2 = (\hat{\theta} - \theta_0)/\sigma_{\hat{\theta}}$, where $\hat{\theta}$ is computed from data on $n_1 + n_2$ patients. "Accept" H_0 if $Z_2 < C_3$; reject H_0 if $Z_2 > C_4$;

otherwise, continue to the next stage.

Stage III - Accrue an additional n_3 patients and calculate $Z_3 = (\hat{\theta} - \theta_0)/\sigma_{\hat{\theta}}$, where $\hat{\theta}$ is computed from data on all $n_1 + n_2 + n_3$ patients. "Accept" H_0 if $Z_3 < C_5$; otherwise reject H_0.

The joint distribution of Z_1, Z_2, and Z_3 is trivariate normal with zero mean vector and correlation matrix

$$\Sigma = \begin{vmatrix} 1.0 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1.0 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1.0 \end{vmatrix},$$

where $\rho_{12} = [n_1/(n_1+n_2)]^{1/2}$, $\rho_{13} = [n_1/(n_1+n_2+n_3)]^{1/2}$, and $\rho_{23} = [(n_1+n_2)/(n_1+n_2+n_3)]^{1/2}$. The maximum sample size for the three-stage design is $n = n_1 + n_2 + n_3$ and is realized whenever all stages are necessary. Let $P_1(S)$, $P_2(S)$, and $P_3(S) = 1 - P_1(S) - P_2(S)$ denote the probabilities that the trial will be stopped at the first, second, or third stage, respectively. Then the expected sample size (ESS) of the three-stage design is given by:

$$ESS = n_1 + n_2[1 - P_1(S)] + n_3[1 - P_1(S) - P_2(S)]. \qquad (3.1)$$

When the sample sizes are the same at each stage, the ESS can be written in the simple form:

$$ESS = n/3[3 - 2P_1(S) - P_2(S)]. \qquad (3.2)$$

For any given trial, the sample size will be n_1 with probability $P_1(S)$, $n_1 + n_2$ with probability $P_2(S)$, and n with probability $P_3(S)$.

There are eight unknown parameters in the three-stage design, namely: n_1, n_2, n_3, C_1, C_2, C_3, C_4, and C_5. The critical value at the third and final stage, C_5, will be set equal to that of the fixed sample test. Additionally, some results will be given for the case where both C_4 and C_5 are set equal to $\Phi^{-1}(1-\alpha)$ (ie, the critical values for rejection at the second and third stages are both equal to the value of a fixed sample test). Most results will be presented for the case of equal sample sizes at each stage. Some calculations will be given which illustrate the loss in efficiency incurred by this restriction. The other parameters of interest are chosen to satisfy the two equations:

$$\alpha = 1 - \Phi(C_2) + B(C_1, C_2;\ C_4, \infty;\ \rho_{12}^2) + T(C_1, C_2;\ C_3, C_4;\ C_5, \infty;\ \Sigma) \tag{3.3}$$

$$1-\beta = 1 - \Phi(C_2 - u\rho_{13}) + B(C_1 - u\rho_{13}, C_2 - u\rho_{13}; C_4 - u\rho_{23}, \infty; \rho_{12}^2) + T(C_1 - u\rho_{13}, C_2 - u\rho_{13}; C_3 - u\rho_{23}, C_4 - u\rho_{23}; C_5 - u\rho_{12}, \infty; \Sigma), \tag{3.4}$$

where $B(a,b;\ c,d;\ \rho)$ was defined in Chapter 2, $T(a,b;\ c,d;\ e,f;\ \Sigma) = (2\pi\Sigma)^{-1/2}\int_a^b\int_c^d\int_e^f \exp[-(1/2)(X'\Sigma^{-1}X)]dX$, $u = \sqrt{n}(\theta_1 - \theta_0)/\sigma$, and $\Phi(x)$ represents the standard normal distribution function. Equation (3.3) is the probability of rejecting H_0 at the first stage plus the probability of continuing the trial and rejecting H_0 at the second stage plus the probability of continuing the trial and rejecting H_0 at the third stage, assuming H_0 is true. Equation (3.4) is this same probability under the alternative hypothesis.

Equations (3.3) and (3.4) are solved iteratively by numerical integration of the multivariate normal distribution using a subroutine written by Schervish (1984). With more parameters than constraints, optimality criteria are used to determine the parameter values. We consider the same two optimality criteria used in Chapter 2, namely:

(1) Minimax criterion - minimize the maximum ESS, i.e. minimize max ESS(θ) with respect to the eight design parameters. For the three-stage design presented above, θ_{max} must be solved by numerical methods adding yet another parameter to be determined and further increasing the computation time (see section 3.8).

(2) Bayes weighted average criteria - (a) minimize the ESS under H_0 and (b) minimize the ESS under H_1. Hereafter, we call these designs 'null Bayes' and 'alternate Bayes', respectively.

The probability of stopping early varies depending on the design parameters and the true value of θ. In general, $P_1(S) = \Phi(C_1-v_1) + 1 - \Phi(C_2-v_1)$, $P_2(S) = B(C_1-v_1, C_2-v_1; C_4-v_2, \infty; \rho_{12}^2) + B(C_1-v_1, C_2-v_1; -\infty, C_3-v_2; \rho_{12}^2)$ and $P_3(S) = 1 - P_1(S) - P_2(S)$, where $v_1 = \rho_{13}\sqrt{n}(\theta - \theta_0)/\sigma$ and $v_2 = \rho_{23}\sqrt{n}(\theta - \theta_0)/\sigma$. The algorithm used to obtain the parameter values is outlined briefly in section 3.3.

3.3 Algorithm

The algorithm used for the 3-stage problem is almost identical to that described in the previous chapter (and shown in the appendix) with additional loops added to iterate over the extra parameters. However, due to the vastly increased computation time required to numerically integrate the trivariate normal distribution (an average trivariate calculation accurate to the 5^{th} decimal place takes approximately 0.430 seconds of CPU time on a VAX 11-750 computer compared to 0.019 seconds for a bivariate calculation accurate to machine precision), less precise results for three-stage designs are presented in this chapter. The type I and II errors are determined only to within $\pm\ 1 \times 10^{-5}$ of their desired values. Also, tolerance levels of 1×10^{-3} were used for the minimization routines (compared to 1×10^{-5} used previously), resulting in fewer iterations per parameter and less accuracy. These considerations probably indicate that the ESS results presented in this chapter are accurate only to the 2^{nd} decimal place. Several runs using different starting values gave the same ESS to this accuracy.

3.4 Results

Results are shown in Tables 3.1 - 3.3 for three-stage designs with equal samples at each stage. The designs optimized under H_0 (Table 3.1) result in expected sample sizes which are 0.50 to 0.63 times the fixed sample size when $\theta = \theta_0$. In this case, the savings become greater as α becomes smaller and β becomes larger. The ratio of the ESS to n_f increases to 0.70 - 0.83 when $\theta = \theta_1$ and to 0.84 - 0.87 when $\theta = \theta_{max}$. Thus, these designs have inverse efficiencies ranging from 0.50 to 0.87, depending on the true value of θ.

The expected sample sizes for the alternate Bayes designs (Table 3.2) and the minimax designs (Table 3.3) are similar to those of the null Bayes designs. For the alternate Bayes designs, the inverse efficiency ranges from 0.52 to 0.73 when $\theta = \theta_0$, from 0.67 to 0.80 when $\theta = \theta_1$, and from 0.82 to 0.88 when $\theta = \theta_{max}$. This represents increases of up to 10% when $\theta = \theta_0$ and decreases of up to 8% when $\theta = \theta_1$. The inverse efficiencies are relatively unchanged when $\theta = \theta_{max}$. When $\theta = \theta_1$, the savings are greater for larger α and smaller β. For the minimax designs, the efficiency ranges from 0.51 to 0.65 when $\theta = \theta_0$, from 0.69 to 0.80 when $\theta = \theta_1$, and from 0.81 to 0.85 when $\theta = \theta_{max}$. This represents increases of no more than 2% when $\theta = \theta_0$ or $\theta = \theta_1$ and savings of no more than 3% when $\theta = \theta_{max}$.

Table 3.1. Restricted Three-Stage One-Sided Designs – (Optimized under H_0)

α	$1-\beta$	c_1	c_2	c_3	c_4	c_5	n_f^a	n^a	n/n_f	$P_1(S)$	$P_2(S)$	$P_3(S)$	$\bar{n}^a$	$\bar{n}/n_f$
.01	.8	0.734	3.704	1.334	2.602	2.326	10.036	11.675	1.16	.769	.167	.064	5.044	0.50
										.149	.557	.294	8.347	0.83
										.185	.462	.353	8.442	0.84
	.9	0.649	3.704	1.337	2.629	2.326	13.017	15.157	1.16	.742	.190	.068	6.698	0.51
										.128	.647	.225	10.600	0.81
										.170	.450	.380	11.171	0.86
	.95	0.561	3.589	1.330	2.663	2.326	15.770	18.325	1.16	.713	.216	.071	8.302	0.53
										.160	.671	.169	12.275	0.78
										.163	.425	.412	13.738	0.87
.025	.8	0.534	3.112	1.094	2.228	1.960	7.849	9.355	1.19	.704	.202	.094	4.335	0.55
										.198	.544	.258	6.423	0.82
										.225	.452	.323	6.541	0.83
	.9	0.435	2.970	1.093	2.280	1.960	10.507	12.475	1.19	.670	.230	.100	5.949	0.57
										.230	.566	.204	8.206	0.78
										.224	.415	.361	8.888	0.85
	.95	0.342	2.923	1.084	2.325	1.960	12.995	15.359	1.18	.636	.257	.107	7.533	0.58
										.282	.563	.155	9.588	0.74
										.211	.396	.393	11.173	0.86
.05	.8	0.341	2.568	0.875	1.938	1.645	6.183	7.507	1.21	.639	.235	.126	3.722	0.60
										.269	.490	.241	4.931	0.80
										.279	.410	.311	5.083	0.82
	.9	0.235	2.483	0.874	2.011	1.645	8.564	10.318	1.20	.600	.264	.136	5.285	0.62
										.318	.487	.195	6.458	0.75
										.269	.380	.351	7.161	0.84
	.95	0.139	2.423	0.874	2.085	1.645	10.822	12.970	1.20	.563	.293	.144	6.836	0.63
										.392	.454	.154	7.619	0.70
										.257	.356	.387	9.204	0.85

[a] Divide each value by Δ^2.

Note: The cut points are chosen to mimimize the expected sample size under the null hypothesis. The probality of stopping at each stage and the expected sample size for this design are obtained assuming $\Theta = \Theta_0$ (top line), $\Theta = \Theta_1$ (middle line), and $\Theta = \Theta_{max}$ (bottom line).

Table 3.2. Restricted Three-Stage One-Sided Designs - (Optimized under H_1)

α	$1-\beta$	c_1	c_2	c_3	c_4	c_5	n_f^a	n^a	n/n_f	$P_1(S)$	$P_2(S)$	$P_3(S)$	$\bar{n}^a$	$\bar{n}/n_f$
.01	.8	0.816	2.786	1.736	2.664	2.326	10.036	12.691	1.26	.795	.176	.029	5.219	0.52
										.340	.432	.228	7.984	0.80
										.328	.390	.282	8.264	0.82
	.9	0.535	2.728	1.895	2.691	2.326	13.017	16.592	1.27	.707	.271	.022	7.274	0.56
										.388	.450	.162	9.813	0.75
										.273	.458	.269	11.039	0.85
	.95	0.162	2.664	2.006	2.726	2.326	15.770	20.501	1.30	.568	.414	.018	9.902	0.63
										.487	.406	.107	11.069	0.70
										.215	.537	.248	13.894	0.88
.025	.8	0.552	2.453	1.431	2.313	1.960	7.849	10.007	1.27	.717	.232	.051	4.452	0.57
										.367	.418	.215	6.165	0.79
										.344	.387	.269	6.422	0.82
	.9	0.282	2.380	1.584	2.357	1.960	10.507	13.544	1.29	.620	.340	.040	6.414	0.61
										.432	.414	.154	7.773	0.74
										.294	.450	.256	8.856	0.84
	.95	-.190	2.343	1.654	2.397	1.960	12.995	16.759	1.29	.434	.529	.037	8.951	0.69
										.514	.376	.110	8.921	0.69
										.213	.538	.249	11.375	0.88
.05	.8	0.321	2.152	1.177	2.022	1.645	6.183	7.965	1.29	.642	.282	.076	3.809	0.62
										.396	.400	.204	4.801	0.78
										.361	.382	.257	5.034	0.81
	.9	0.004	2.089	1.318	2.067	1.645	8.564	11.081	1.29	.520	.417	.063	5.701	0.67
										.461	.391	.148	6.232	0.73
										.298	.457	.245	7.190	0.84
	.95	-.387	2.055	1.371	2.108	1.645	10.822	14.041	1.30	.369	.572	.059	7.910	0.73
										.548	.345	.107	7.292	0.67
										.231	.531	.238	9.398	0.87

[a] Divide each value by Δ^2.

Note: The cut points are chosen to mimimize the expected sample size under the alternative hypothesis. The probality of stopping at each stage and the expected sample size for this design are obtained assuming $\theta = \theta_0$ (top line), $\theta = \theta_1$ (middle line), and $\theta = \theta_{max}$ (bottom line).

Table 3.3. Optimal Restricted Three-Stage One-Sided Minimax Designs

α	$1-\beta$	c_1	c_2	c_3	c_4	c_5	n_f^a	n^a	n/n_f	$P_1(S)$	$P_2(S)$	$P_3(S)$	$\bar{n}^a$	$\bar{n}/n_f$
.01	.8	0.789	2.900	1.640	2.631	2.326	10.036	12.351	1.23	.787	.177	.036	5.142	0.51
										.299	.459	.242	7.998	0.80
										.298	.401	.301	8.242	0.82
	.9	0.544	2.979	1.647	2.655	2.326	13.017	15.498	1.19	.700	.260	.040	6.878	0.53
										.282	.513	.205	9.935	0.76
										.227	.434	.339	10.904	0.84
	.95	0.342	3.030	1.626	2.685	2.326	15.770	18.366	1.16	.635	.321	.044	8.625	0.55
										.306	.529	.165	11.381	0.72
										.181	.447	.372	13.411	0.85
.025	.8	0.526	2.548	1.359	2.286	1.960	7.849	9.765	1.24	.706	.234	.060	4.407	0.56
										.329	.444	.227	6.178	0.79
										.316	.399	.285	6.410	0.82
	.9	0.288	2.615	1.339	2.330	1.960	10.507	12.581	1.20	.618	.312	.070	6.088	0.58
										.325	.476	.199	7.864	0.75
										.246	.421	.333	8.747	0.83
	.95	0.125	2.658	1.329	2.349	1.960	12.995	15.328	1.18	.554	.372	.074	7.771	0.60
										.362	.485	.153	9.153	0.70
										.206	.439	.355	10.982	0.85
.05	.8	0.305	2.259	1.094	1.987	1.645	6.183	7.750	1.25	.632	.278	.090	3.768	0.61
										.354	.430	.216	4.814	0.78
										.331	.394	.275	5.024	0.81
	.9	0.096	2.287	1.097	2.040	1.645	8.564	10.431	1.22	.549	.351	.100	5.389	0.63
										.375	.440	.185	6.293	0.73
										.274	.414	.312	7.089	0.83
	.95	−.084	2.353	1.075	2.056	1.645	10.822	12.891	1.19	.476	.415	.109	7.018	0.65
										.405	.451	.144	7.469	0.69
										.223	.437	.340	9.094	0.84

[a] Divide each value by Δ^2.

Note: The cut points are chosen to mimimize the maximum expected sample size. The pıobality of stopping at each stage and the expected sample size for this design are obtained assuming $\Theta = \Theta_0$ (top line), $\Theta = \Theta_1$ (middle line), and $\Theta = \Theta_{max}$ (bottom line).

The critical value for rejection at the first stage is always greatest for the null Bayes designs and always smallest for the alternate Bayes designs. Additionally, the critical value for acceptance at the second stage is always smallest for the null Bayes designs and always largest for the alternate Bayes designs. The critical values for rejection at the second stage are similar for all three designs.

3.5 Effect of Restricting Sample Sizes

Due to the prohibitive time involved in numerically evaluating the three-stage problem, some restrictions are necessary. One obvious choice is to require equal sample sizes at each stage. This results in designs which are simpler to explain and administer. However, simplicity should not be the prime consideration if more complex designs result in substantial savings. Unfortunately, it is difficult to get a handle on the loss incurred with this restriction since the optimal solution to the completely unrestricted three-stage design requires more computation time than is currently available. One could restrict other parameters (say C_3 and C_4) and determine the effect of differing sample sizes on this much restricted design. One would then need to look at several restriction sets to determine what effect is due to restricting the cut points and what is due to restricting the sample sizes.

I have opted for a simpler evaluation of this problem by considering in detail the restriction for two-stage designs and for two examples of three-stage designs. Table 3.4 shows the efficiency of two-stage designs with equal samples at each stage relative to the optimal designs. It is seen that equal sample minimax and alternate Bayes designs are more than 98% efficient for the α and β considered here. Additionally, the equal sample null Bayes designs are usually more than 90% efficient.

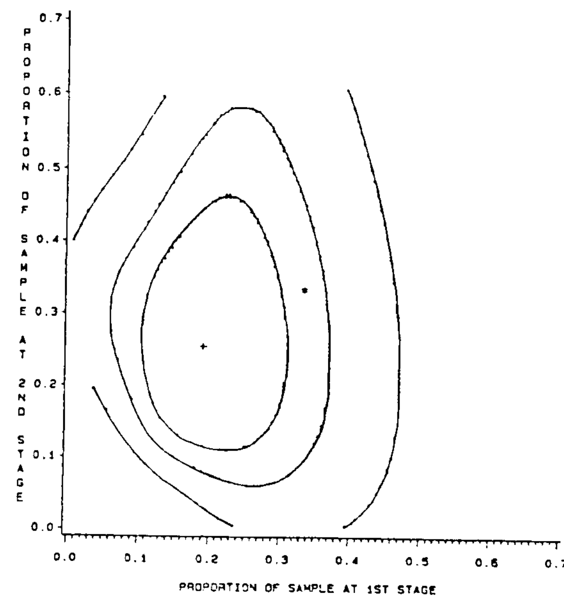
Two three-stage designs were considered. These were designs optimized under the null and alternative hypotheses for $\alpha = 0.01$ and $1-\beta = 0.90$. Minimax designs were not considered due to their similarity to the alternate Bayes designs. The expected sample size was generated for several combinations of p_1 and p_2, where p_1 and p_2 denote the proportion of the sample at the first and second stage, respectively. A bivariate spline algorithm (SAS/GRAPH Users Guide, G3GRID Procedure, 1985) was then used to approximate the overall minimum ESS and the values of p_1 and p_2 associated with this minimum. Approximate contours of equal efficiency were determined from the interpolated values. These are shown in Figure 3.1. The approximate minimum ESS for the null Bayes design is 6.188 and compares to 6.698 for the design with equal sample sizes at each stage. The respective values for the alternate Bayes design are 9.714 and 9.813. The ratio of the minimum ESS to the ESS for the design with equal sample sizes at each stage is 0.92 for the

Table 3.4. Efficiency of Optimal Restricted Two-Stage Design with Equal Sample Sizes at each Stage *

α	$1-\beta$	θ_0	θ_1	θ_{max}
.01	.80	0.88	0.99	0.99
	.90	0.91	0.99	0.99
	.95	0.94	0.99	0.98
.025	.80	0.92	0.99	1.00
	.90	0.95	0.99	0.99
	.95	0.97	1.00	0.98
.05	.80	0.96	1.00	1.00
	.90	0.97	1.00	0.99
	.95	0.99	1.00	0.99

* Efficiency equals the ratio of the expected sample sizes of two-stage designs optimized for different values of θ with unequal and equal sample sizes at each stage.

(a)



(b)

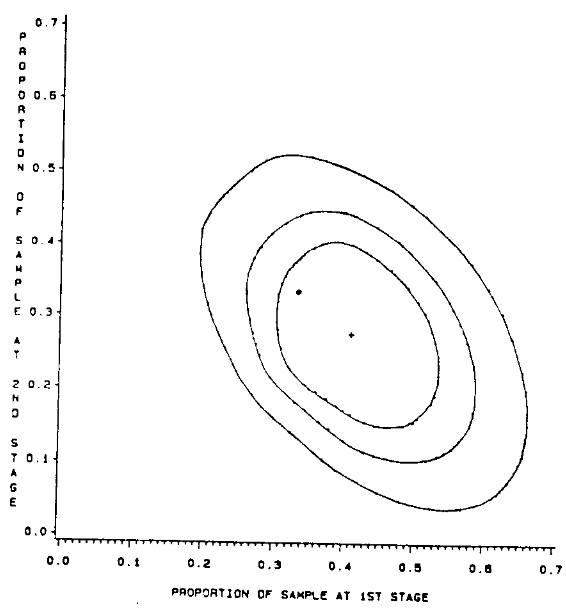


Figure 3.1. Contours of equal relative efficiency for two three-stage designs with $\alpha = 0.01$ and $1-\beta = 0.90$. Figures (a) and (b) give efficiencies for designs optimized under H_0 and H_1, respectively. The inner, middle, and outer contours represents 95%, 90%, and 80% efficiency, respectively. The plus sign denotes the optimal design, and the star denotes the design with equal samples at each stage.

null Bayes design and 0.99 for the alternate Bayes designs. These values are greater than or equal to their respective values for the two-stage designs.

The contours shown in Figure 3.1 actually depict relative efficiency [$E_R = (n_f - ESS)/(n_f - ESS_{opt})$], where ESS_{opt} denotes the minimum expected sample size. This is done so that an expected sample size = n_f would have 0% efficiency while an expected sample size = ESS_{opt} would have 100% efficiency. Thus, the relative efficiency is 0.93 under H_0 and 0.97 under H_1. Note that the null Bayes designs are more efficient when the proportion of patients accrued during the first stage is small (0.1 - 0.3). The efficiency of the alternate Bayes design is greater for larger sample proportions at the first stage (0.3 - 0.5). Additionally, the null Bayes designs are more efficient when the first two stages constitute less than 60% of the total sample size whereas the alternate Bayes designs are more efficient when the proportion of the total sample accrued during the first two stages is greater than 50%.

3.6 Doubly Restricted Designs

In addition to restricting C_5 equal to $\Phi^{-1}(1-\alpha)$, one might restrict C_4 likewise so that the rejection value at both the second and third stage equals that of a fixed sample design. Results obtained with this added restriction are shown in Tables 3.5 - 3.7. The results are interesting.

Table 3.5. Doubly Restricted Three-Stage One-Sided Designs - (Optimized Under H_0)

α	$1-\beta$	c_1	c_2	c_3	c_4	c_5	n_f^a	n^a	n/n_f	$P_1(S)$	$P_2(S)$	$P_3(S)$	$\bar{n}^a$	$\bar{n}/n_f$
.01	.8	1.197	4.663	1.899	2.326	2.326	10.036	14.948	1.49	.884	.103	.013	5.620	0.56
										.158	.776	.066	9.508	0.95
										.089	.886	.025	9.646	0.96
	.9	1.095	4.663	2.017	2.326	2.326	13.017	19.771	1.52	.863	.128	.009	7.548	0.58
										.089	.880	.031	12.804	0.98
										.080	.896	.024	12.815	0.98
	.95	1.026	4.663	2.075	2.326	2.326	15.770	24.301	1.54	.848	.145	.007	9.391	0.60
										.069	.916	.015	15.762	1.00
										.074	.904	.022	15.790	1.00
.025	.8	0.950	3.980	1.523	1.960	1.960	7.849	11.788	1.50	.829	.146	.025	4.700	0.60
										.174	.761	.065	7.433	0.95
										.139	.820	.041	7.476	0.95
	.9	0.853	3.980	1.645	1.960	1.960	10.507	16.102	1.53	.803	.179	.018	6.520	0.62
										.120	.849	.031	10.259	0.98
										.125	.838	.037	10.265	0.98
	.95	0.783	3.980	1.706	1.960	1.960	12.995	20.211	1.56	.783	.202	.015	8.294	0.64
										.118	.867	.015	12.776	0.98
										.115	.851	.034	12.925	0.99
.05	.8	0.703	3.414	1.244	1.645	1.645	6.183	9.308	1.51	.759	.203	.038	3.966	0.64
										.194	.743	.063	5.797	0.94
										.185	.761	.054	5.802	0.94
	.9	0.626	3.414	1.338	1.645	1.645	8.564	13.192	1.54	.735	.236	.029	5.692	0.66
										.165	.805	.030	8.204	0.96
										.170	.782	.048	8.258	0.96
	.95	0.544	3.486	1.409	1.645	1.645	10.822	16.872	1.56	.707	.270	.023	7.399	0.68
										.166	.820	.014	10.390	0.96
										.146	.814	.040	10.650	0.98

[a] Divide each value by Δ^2.

Note: The cut points are chosen to mimimize the expected sample size under the null hypothesis. The probality of stopping at each stage and the expected sample size for this design are obtained assuming $\theta = \theta_0$ (top line), $\theta = \theta_1$ (middle line), and $\theta = \theta_{max}$ (bottom line).

Table 3.6. Doubly Restricted Three-Stage One-Sided Designs - (Optimized Under H_1)

α	$1-\beta$	c_1	c_2	c_3	c_4	c_5	n_f^a	n^a	n/n_f	$P_1(S)$	$P_2(S)$	$P_3(S)$	$\bar{n}^a$	$\bar{n}/n_f$
.01	.8	1.330	2.746	2.195	2.326	2.326	10.036	16.928	1.69	.911	.086	.003	6.157	0.61
										.503	.480	.017	8.542	0.85
										.479	.494	.027	8.732	0.87
	.9	1.266	2.714	2.321	2.326	2.326	13.017	22.591	1.74	.901	.099	.000	8.279	0.64
										.582	.418	.000	10.683	0.82
										.469	.530	.001	11.537	0.89
	.95	1.177	2.765	2.320	2.326	2.326	15.770	26.978	1.71	.883	.117	.000	10.044	0.64
										.627	.373	.000	12.351	0.78
										.427	.571	.002	14.157	0.90
.025	.8	1.045	2.403	1.853	1.960	1.960	7.849	13.362	1.70	.860	.136	.005	5.096	0.65
										.528	.457	.015	6.619	0.84
										.497	.480	.023	6.794	0.87
	.9	0.948	2.396	1.959	1.960	1.960	10.507	18.124	1.72	.837	.163	.000	7.027	0.67
										.590	.409	.001	8.516	0.81
										.469	.530	.001	9.249	0.88
	.95	0.892	2.432	1.955	1.960	1.960	12.995	22.297	1.72	.821	.178	.001	8.761	0.67
										.649	.351	.000	10.044	0.77
										.441	.558	.001	11.594	0.89
.05	.8	0.795	2.102	1.565	1.645	1.645	6.183	10.631	1.72	.804	.190	.006	4.255	0.69
										.551	.438	.011	5.172	0.84
										.513	.469	.018	5.330	0.86
	.9	0.694	2.114	1.644	1.645	1.645	8.564	14.781	1.73	.773	.227	.000	6.044	0.71
										.606	.394	.000	6.870	0.80
										.478	.522	.000	7.501	0.88
	.95	0.624	2.160	1.643	1.645	1.645	10.822	18.490	1.71	.749	.251	.000	7.711	0.71
										.658	.342	.000	8.272	0.76
										.442	.557	.001	9.603	0.89

[a] Divide each value by Δ^2.

Note: The cut points are chosen to mimimize the expected sample size under the alternative hypothesis. The probality of stopping at each stage and the expected sample size for this design are obtained assuming $\theta = \theta_0$ (top line), $\theta = \theta_1$ (middle line), and $\theta = \theta_{max}$ (bottom line).

Table 3.7. Doubly Restricted Three-Stage One-Sided Minimax Designs

α	$1-\beta$	c_1	c_2	c_3	c_4	c_5	n_f^a	n^a	n/n_f	$P_1(S)$	$P_2(S)$	$P_3(S)$	$\bar{n}^a$	$\bar{n}/n_f$
.01	.8	1.246	2.838	2.161	2.326	2.326	10.036	16.255	1.62	.896	.101	.003	6.001	0.60
										.445	.531	.024	8.560	0.85
										.427	.536	.037	8.721	0.87
	.9	1.177	2.769	2.314	2.326	2.326	13.017	21.922	1.68	.883	.117	.000	8.163	0.63
										.537	.462	.001	10.697	0.82
										.426	.571	.003	11.524	0.89
	.95	1.066	2.838	2.312	2.326	2.326	15.770	26.059	1.65	.859	.141	.000	9.914	0.63
										.574	.426	.000	12.397	0.79
										.376	.620	.004	14.144	0.90
.025	.8	0.999	2.469	1.814	1.960	1.960	7.849	12.975	1.65	.848	.146	.006	5.010	0.64
										.488	.491	.021	6.629	0.84
										.463	.505	.032	6.788	0.86
	.9	0.907	2.425	1.953	1.960	1.960	10.507	17.852	1.70	.825	.174	.001	6.992	0.67
										.568	.431	.001	8.523	0.81
										.448	.551	.001	9.247	0.88
	.95	0.813	2.479	1.959	1.960	1.960	12.995	21.773	1.68	.798	.202	.000	8.722	0.67
										.615	.385	.000	10.052	0.77
										.405	.595	.000	11.581	0.89
.05	.8	0.758	2.161	1.530	1.645	1.645	6.183	10.354	1.67	.791	.201	.008	4.200	0.68
										.516	.467	.017	5.178	0.84
										.483	.491	.026	5.325	0.86
	.9	0.662	2.134	1.644	1.645	1.645	8.564	14.627	1.71	.763	.237	.000	6.033	0.70
										.591	.409	.000	6.871	0.80
										.462	.538	.000	7.500	0.88
	.95	0.592	2.182	1.643	1.645	1.645	10.822	18.300	1.69	.738	.262	.000	7.701	0.71
										.644	.356	.000	8.274	0.76
										.427	.573	.000	9.599	0.89

[a] Divide each value by Δ^2.

Note: The cut points are chosen to mimimize the maximum expected sample size. The probality of stopping at each stage and the expected sample size for this design are obtained assuming $\Theta = \Theta_0$ (top line), $\Theta = \Theta_1$ (middle line), and $\Theta = \Theta_{max}$ (bottom line).

For the null Bayes designs, huge differences are required to declare an early significant difference. The minimax and alternate Bayes designs are similar to one another and give results almost identical to the two-stage results (i.e., $C_3 \approx C_4 = C_5$), indicating that this added restriction is not feasible for the latter two designs in terms of expected sample size. This restriction results in two-stage designs for all practical purposes, eliminating any expected savings in sample size which might have been realized with a third look.

3.7 Two-Sided Results

Compared to the one-sided problem which requires one evaluation of the trivariate normal distribution per area or power calculation, the two-sided problem requires four evaluations for an area calculation and eight evaluations for a power calculation. Due to these increased computational requirements, two-sided results are limited to $\alpha = 0.05$ and 0.10. These values are chosen for comparison with the one-sided results.

Two-sided results are shown in Tables 3.8 - 3.10. As with the two-stage designs, the two-sided three-stage minimax and alternate Bayes α-level designs are very similar to their equivalent one-sided $\alpha/2$-level designs. The expected sample sizes never differ by more than 3.5% for the designs considered here, and the cut points are similar.

Table 3.8. Restricted Three-Stage Two-Sided Designs - (Optimized under H_0)

α	$1-\beta$	c_1	c_2	c_3	c_4	c_5	n_f^a	n^a	n/n_f	$P_1(S)$	$P_2(S)$	$P_3(S)$	$\bar{n}^a$	$\bar{n}/n_f$
.05	.8	0.826	2.556	1.292	2.176	1.960	7.849	10.699	1.36	.602	.290	.108	5.372	0.68
										.393	.450	.157	6.289	0.80
										.390	.387	.223	6.537	0.83
	.9	0.744	2.467	1.328	2.268	1.960	10.507	14.433	1.37	.557	.332	.111	7.478	0.71
										.464	.415	.121	7.968	0.76
										.383	.361	.256	9.012	0.86
	.95	0.673	2.416	1.363	2.352	1.960	12.995	18.019	1.39	.515	.373	.112	9.593	0.74
										.551	.362	.087	9.230	0.71
										.370	.348	.282	11.484	0.88
.10	.8	0.665	2.107	1.121	1.921	1.645	6.183	8.887	1.44	.529	.340	.131	4.744	0.77
										.487	.377	.136	4.887	0.79
										.449	.349	.202	5.190	0.84
	.9	0.583	2.071	1.163	2.006	1.645	8.564	12.368	1.44	.479	.387	.134	6.828	0.80
										.553	.343	.104	6.391	0.75
										.423	.352	.225	7.431	0.87
	.95	0.496	2.064	1.183	2.072	1.645	10.822	15.567	1.44	.420	.439	.141	8.938	0.83
										.620	.305	.075	7.556	0.70
										.386	.365	.249	9.666	0.89

[a] Divide each value by Δ^2.

Note: The cut points are chosen to mimimize the expected sample size under the null hypothesis. The probality of stopping at each stage and the expected sample size for this design are obtained assuming $\Theta = \Theta_0$ (top line), $\Theta = \Theta_1$ (middle line), and $\Theta = \Theta_{max}$ (bottom line).

Table 3.9. Restricted Three-Stage Two-Sided Designs - (Optimized under H_1)

α	$1-\beta$	c_1	c_2	c_3	c_4	c_5	n_f^a	n^a	n/n_f	$P_1(S)$	$P_2(S)$	$P_3(S)$	$\bar{n}^a$	$\bar{n}/n_f$
.05	.8	0.619	2.429	1.441	2.309	1.960	7.849	10.198	1.30	.479	.423	.098	5.502	0.70
										.383	.415	.202	6.188	0.79
										.348	.393	.259	6.498	0.83
	.9	0.364	2.360	1.606	2.355	1.960	10.507	13.762	1.31	.302	.622	.076	8.136	0.77
										.445	.410	.145	7.797	0.74
										.269	.490	.241	9.048	0.86
	.95	0.008	2.330	1.685	2.392	1.960	12.995	16.863	1.30	.026	.906	.068	11.476	0.88
										.517	.377	.106	8.933	0.69
										.099	.693	.208	11.857	0.91
.10	.8	0.435	2.118	1.198	2.012	1.645	6.183	8.198	1.33	.370	.489	.141	4.838	0.78
										.414	.399	.187	4.844	0.78
										.344	.415	.241	5.184	0.84
	.9	0.115	2.093	1.326	2.072	1.645	8.564	11.008	1.29	.128	.745	.127	7.336	0.86
										.444	.403	.153	6.268	0.73
										.174	.598	.228	7.540	0.88
	.95	0.003	2.066	1.369	2.106	1.645	10.822	13.900	1.28	.041	.838	.121	9.936	0.89
										.535	.355	.110	7.300	0.67
										.099	.690	.211	9.784	0.90

[a] Divide each value by Δ^2.

Note: The cut points are chosen to mimimize the expected sample size under the alternative hypothesis. The probality of stopping at each stage and the expected sample size for this design are obtained assuming $\Theta = \Theta_0$ (top line), $\Theta = \Theta_1$ (middle line), and $\Theta = \Theta_{max}$ (bottom line).

Table 3.10. Restricted Three-Stage Two-Sided Minimax Designs

α	$1-\beta$	c_1	c_2	c_3	c_4	c_5	n_f^a	n^a	n/n_f	$P_1(S)$	$P_2(S)$	$P_3(S)$	$\bar{n}^a$	$\bar{n}/n_f$
.05	.8	0.638	2.511	1.361	2.270	1.960	7.849	10.066	1.28	.488	.399	.113	5.451	0.69
										.358	.436	.206	6.201	0.79
										.336	.395	.269	6.483	0.83
	.9	0.468	2.568	1.377	2.290	1.960	10.507	13.155	1.25	.370	.509	.121	7.677	0.73
										.365	.471	.164	7.892	0.75
										.266	.438	.296	8.899	0.85
	.95	0.310	2.630	1.348	2.325	1.960	12.995	15.782	1.21	.252	.609	.139	9.928	0.76
										.387	.478	.135	9.190	0.71
										.198	.467	.335	11.236	0.86
.10	.8	0.496	2.193	1.116	1.952	1.645	6.183	8.211	1.33	.409	.435	.156	4.782	0.77
										.403	.418	.179	4.859	0.79
										.356	.401	.243	5.165	0.84
	.9	0.328	2.261	1.118	1.989	1.645	8.564	10.941	1.28	.281	.541	.178	6.915	0.81
										.407	.442	.151	6.357	0.74
										.264	.457	.279	7.353	0.86
	.95	0.003	2.523	1.004	2.106	1.645	10.822	12.019	1.11	.014	.712	.274	9.054	0.84
										.301	.505	.194	7.580	0.70
										.054	.544	.402	9.412	0.87

[a] Divide each value by Δ^2.

Note: The cut points are chosen to mimimize the maximum expected sample size.
The probality of stopping at each stage and the expected sample size for this design are obtained assuming $\Theta = \Theta_0$ (top line), $\Theta = \Theta_1$ (middle line), and $\Theta = \Theta_{max}$ (bottom line).

The $\alpha/2$ one-sided minimax and alternate Bayes plans could be used as good approximations to the two-sided α-level designs. Alpha will increase slightly and beta will decrease slightly, but the changes are not dramatic. However, the changes are more substantial for the null Bayes designs and use of one-sided approximations is not appropriate. The ESS can increase by more than 30%, and the optimal cut points are not close.

3.8 Determination of θ_{max}

Let's consider three-stage one-sided designs with equal sample sizes at each stage. As before, θ_{max} is found by maximizing the expected sample size,

$$ESS = n/3(3 - 2P_1(S) - P_2(S)).$$

Maximizing the ESS is equivalent to minimizing

$$\begin{aligned} P(S) = 2P_1(S) + P_2(S) &= 2[1 - \Phi(C_2-\rho_{13}v) + \Phi(C_1-\rho_{13}v)] \\ &+ B(C_1-\rho_{13}v, C_2-\rho_{13}v;\ C_4-\rho_{23}v, \infty;\ 1/\sqrt{2}) \\ &+ B(C_1-\rho_{13}v, C_2-\rho_{13}v;\ \infty, C_3-\rho_{23}v;\ 1/\sqrt{2}), \end{aligned}$$

where $v = \sqrt{n}(\theta-\theta_0)/\sigma$.

Minimizing P(S) requires differentiating under the normal and bivariate normal integrals. A two-dimensional extension of Leibnitz' rule is used to differentiate under the bivariate normal integral. This extension is given by:

$$\begin{aligned} d/dv(\textstyle\int_a^b\int_c^d f(x,y)dxdy) &= \textstyle\int_a^b(d/dv\int_c^d f(x,y)dy)dx \\ &+ \textstyle\int_c^d f(b,y)(db/dv)dy - \int_c^d f(a,y)(da/dv)dy \\ &= \textstyle\int_a^b[f(x,d)dd/dv - f(x,c)dc/dv]dx \end{aligned}$$

$+ \int_c^d [f(b,y)db/dv - f(a,y)da/dv]dy.$

Let $f(x,y) = (1/\sqrt{2\pi})\exp(-(x^2 - \sqrt{2}xy + y^2))$ denote the bivariate normal density with correlation(x,y) = $1/\sqrt{2}$. Also let $a = C_1-\rho_{13}v$, $b = C_2-\rho_{13}v$, $c = C_3-\rho_{23}v$, and $d = C_4-\rho_{23}v$. Then applying Leibnitz' rule,

$$dP(S)/dv = (2/\sqrt{2\pi})\exp(-b^2/2) - (2/\sqrt{2\pi})\exp(-a^2/2)$$
$$+ 2\rho_{13}\int_c^d(f(a,y) - f(b,y))dy + \rho_{23}\int_a^b(f(x,c) - f(x,d))dx,$$

The equation $dP(S)/dv = 0$ has no explicit solution in general and must be solved numerically. In practice, it is probably easier, though somewhat more time consuming, to input P(S) into a derivative-free minimization routine to find θ_{max}.

3.9 Discussion

The additional savings realized in going from 2 to 3 stages are not as substantial as the savings realized in going from 1 to 2 stages. The additional decrease in sample size ranges from 5% to 10% depending on α, β, and the true value of θ. The additional savings are greatest for the null Bayes designs and smallest for the minimax designs. Although the decrease in expected sample size seems small, it actually represents up to 40% of the remaining savings possible (as compared to a Wald SPRT design). In Chapter 4, the three-stage designs will be compared to the two-stage designs presented earlier and to other group sequential plans presently in use.

IV. EVALUATION OF TWO AND THREE-STAGE DESIGNS

4.1 Introduction

It was shown in Chapters 2 and 3 that optimal restricted two and three-stage designs result in sample size savings relative to fixed sample designs. As discussed in Chapter 1, other two-stage and multi-stage designs have been presented in the literature, many of which also result in sample size savings. In this chapter, the efficiency of the restricted designs relative to several other designs is explored. The optimal restricted two-stage designs will be compared with the optimal unrestriced two-stage designs and the optimal two-stage designs which do not allow acceptance of H_0 at the first stage. These comparisons will illustrate the loss in efficiency incurred by restricting the final stage critical value and by not allowing acceptance of H_0 at the first stage. The two and three-stage designs will also be compared with each other and with Wald's SPRT designs, which have been shown to minimize the ESS under both the null and the specified alternative hypotheses. The percentage of possible savings in sample size afforded by the proposed two-stage designs will be determined, and the additional gains realized with three looks will be examined.

Finally, the efficiency of the three-stage designs relative to a few of the more popular group sequential designs currently in use will also be examined.

4.2 Optimal Unrestricted Two-stage Designs

As mentioned in Chapter 1, Hald (1975) derived optimal unrestricted two-stage designs. The designs presented in Chapter 2 were optimal given the restriction that $C_3 = \Phi^{-1}(1-\alpha)$. They do not give quite the expected savings in sample size achieved by the unrestricted designs, but the difference is usually small. The efficiency of the restricted design relative to Hald's unrestricted design is given by the ratio of the ESS of the unrestricted and restricted plans and is listed in Table 4.1 for the plans previously presented. These values range from 82% to 100%. For the minimax plans, the efficiency is greater than 98% for the α and β considered while the efficiency is greater than 96% for the designs optimized under H_0. The restricted plans optimized under H_1 generally perform most poorly, particularly for small beta.

Note that Hald derived the optimal symmetric plans when $\alpha = \beta$. For this comparison, we have used the optimal unrestricted nonsymmetric plans since they are slightly more efficient when optimizing under H_0 or H_1. For example, when $\alpha = \beta = 0.05$, $ESS\times\Delta^2$ for the unrestricted symmetric design optimized under H_0 is 7.80 compared to 7.56 for the respective nonsymmetric design. The optimal unrestricted

Table 4.1. Efficiency of Restricted vs Unrestricted Design

α	$1-\beta$	θ_0	θ_1	θ_{max}
.005	.70	96.44	99.55	99.83
	.80	97.51	98.66	99.66
	.90	98.47	96.57	99.39
	.95	98.96	94.15	99.17
	.99	99.49	89.21	98.79
.01	.70	96.34	99.49	99.81
	.80	97.46	98.51	99.63
	.90	98.45	96.20	99.34
	.95	98.96	93.54	99.10
	.99	99.50	88.18	98.69
.025	.70	96.56	99.39	99.79
	.80	97.68	98.24	99.58
	.90	98.65	95.56	99.25
	.95	99.14	92.51	98.98
	.99	99.63	86.44	98.53
.05	.70	97.24	99.28	99.77
	.80	98.27	97.97	99.53
	.90	99.11	94.90	99.16
	.95	99.50	91.46	98.87
	.99	99.86	84.70	98.39
.10	.70	98.48	99.14	99.74
	.80	99.21	97.59	99.47
	.90	99.73	93.98	99.06
	.95	99.92	90.03	98.73
	.99	100.00	82.38	98.21

minimax designs are symmetric when $\alpha = \beta$; this is not true of the restricted designs.

4.3 Optimal Design Not Allowing Early Acceptance

The purpose of this section is to stress two important points. First, designs which restrict the final stage critical value to that of a fixed sample design must allow for early acceptance of the null hypothesis to maintain the overall significance level. Second, from the standpoint of efficiency, there is little reason not to allow early acceptance.

A two-stage design which does not allow early acceptance of H_0 is shown in Figure 4.1. For this design, restricting the final stage critical value to equal $\Phi^{-1}(1-\alpha)$ makes the overall significance level of the design greater than α. In fact, the overall significance level would be $\alpha + B(C_2,\infty; -\infty,C_3; p)$, and since the latter term is greater than zero, the overall significance level is greater than α.

Colton (1963) developed optimal two-stage designs which allowed acceptance of H_0 at the final stage only, arguing that there were no valid reasons to allow early acceptance. He stated that development of designs allowing for early acceptance were not numerically feasible. In addition, he thought drugs would not be readily accepted on the basis of one test. Later, Colton and McPherson (1976) argued that there was no real ethical advantage in stopping a trial

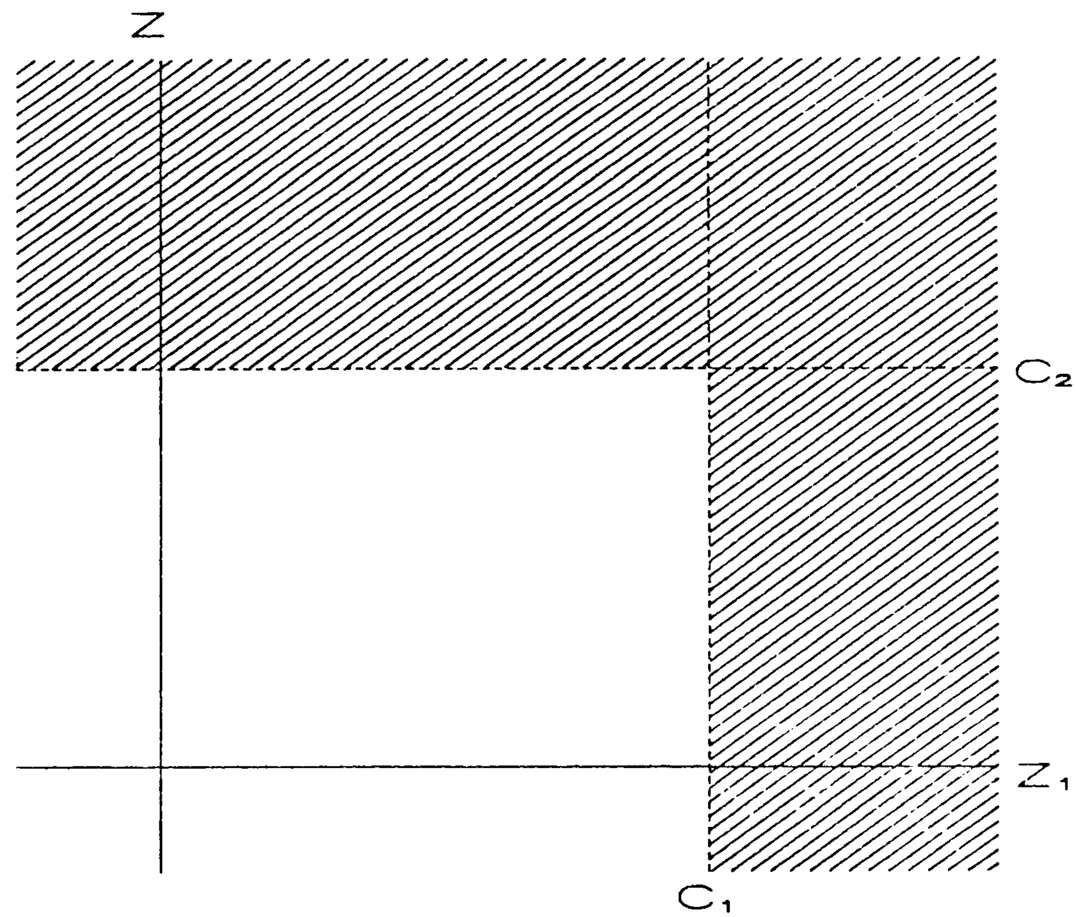


Figure 4.1. A two-stage one-sided design which does not allow acceptance of H_0 at the first stage. The shaded area represents the rejection region of the design.

early and accepting the null hypothesis. The real advantage occurs when trials are stopped due to early significant differences. With advancements in computers, numerical feasibility is no longer a problem. Additionally, if drug acceptance is to require multiple tests, then it seems reasonable to expedite matters as quickly as possible. The first acceptance should occur as soon as possible so the second test can begin. A two-stage design should not count as two tests simply because the investigator chose not to allow early acceptance. Also, it appears unethical to me to continue trials which have little chance of showing statistically significant results. This is a waste of patient resources. New patients could be treated with different and potentially beneficial therapies or they could be participating in other more valuable studies.

It is of interest to compare designs which do not allow acceptance of H_0 at the first stage to those which do. Since it is not possible to develop designs which restrict the final stage critical value as in Chapter 2 without allowing early acceptance, it would be more appropriate to compare Hald's designs, which allow early acceptance, to Colton's designs, which do not. It has been shown above, however, that the restricted designs are highly efficient when compared to the unrestricted designs. With this in mind, the restricted designs are compared to Colton's designs in Table 4.2.

Table 4.2. Efficiency of Restricted vs Unrestricted Design Which Does Not Allow Early Acceptance of H_0

		Optimized under H_1	
α	$1-\beta$	θ_0	θ_1
.005	.70	153.73	104.84
	.80	154.48	101.44
	.90	160.58	97.29
	.95	167.05	94.26
	.99	177.76	89.21
.01	.70	153.24	104.81
	.80	154.67	101.27
	.90	160.06	96.89
	.95	166.01	93.64
	.99	174.50	88.18
.025	.70	150.44	104.72
	.80	152.14	100.96
	.90	157.31	96.20
	.95	161.82	92.59
	.99	167.01	86.44
.05	.70	146.16	104.60
	.80	148.02	100.63
	.90	152.39	95.48
	.95	155.84	91.53
	.99	158.37	84.70
.10	.70	139.04	104.39
	.80	140.96	100.14
	.90	144.40	94.50
	.95	146.69	90.08
	.99	146.97	82.38

It is clear that the restricted designs perform much better under H_0 than do Colton's designs. In some cases, Colton's designs require up to 78% more patients than the restricted designs. The restricted designs also do well under the specified alternative hypothesis. For typical plans, say $\alpha = 0.01$ or 0.05 and $1-\beta = 0.80$ or 0.90, the restricted plans never require more than 1.05 times the number of patients needed in Colton's plans.

4.4 SPRT Designs

As stated before, Wald and Wolfowitz (1948) showed that the fully sequential SPRT designs result in the lowest expected sample sizes under both the null and the specified alternative hypotheses. The expected sample sizes under the competing hypotheses for the SPRT design are given approximately by:

$$\bar{n}_0\Delta^2 = -2[\alpha\ln((1-\beta)/\alpha) + (1-\alpha)\ln(\beta/(1-\alpha))] \text{ and}$$
$$\bar{n}_1\Delta^2 = 2[\beta\ln(\beta/(1-\alpha)) + (1-\beta)\ln((1-\beta)/\alpha)].$$

A comparison of these expected sample sizes and the two-stage expected sample sizes is shown in Table 4.3. The numbers listed in the table give the percentage of the difference between the fixed sample size and the SPRT sample size which is attributable to the two-stage designs. The values are expressed as $S = [(n_f - \bar{n}_{2w})/(n_f - \bar{n}_{sw})]100$, where n_f denotes the fixed sample size, $\bar{n}_{2w}$ denotes the expected sample size for a two-stage design when $\theta = \theta_w$, and $\bar{n}_{sw}$ denotes the expected sample size for a sequential design

Table 4.3. Relative Efficiency of Restricted Two-Stage Design to the Fully Sequential Design *

α	$1-\beta$	Optimized under H_0		Optimized under H_1	
		θ_0	θ_1	θ_0	θ_1
.005	.70	68.8	-5.3	42.9	39.1
	.80	68.3	-4.4	43.6	41.9
	.90	67.6	7.5	47.5	45.2
	.95	67.1	20.5	51.6	47.3
	.99	66.2	41.1	59.5	50.9
.01	.70	66.4	3.1	44.6	39.8
	.80	66.0	5.5	45.7	42.3
	.90	65.5	17.0	49.3	45.2
	.95	65.0	28.2	53.3	47.1
	.99	64.2	45.1	60.2	50.5
.025	.70	62.6	15.9	46.8	40.9
	.80	62.5	19.1	48.0	42.8
	.90	62.1	28.8	51.6	45.1
	.95	61.7	37.3	55.1	46.7
	.99	61.0	48.6	59.8	49.9
.05	.70	59.1	25.8	48.3	41.7
	.80	59.1	28.9	49.6	43.1
	.90	58.8	36.4	52.7	44.9
	.95	58.4	42.3	55.3	46.3
	.99	57.6	49.3	57.5	49.3
.10	.70	54.7	34.8	49.0	42.4
	.80	54.8	37.2	50.2	43.4
	.90	54.5	41.9	52.3	44.6
	.95	54.0	45.2	53.4	45.9
	.99	52.7	47.9	51.7	48.6

* Relative Efficiency = $[(n_f-\bar{n}_{2w})/(n_f-\bar{n}_{sw})]*100$, where n_f denotes the fixed sample size, $\bar{n}_{2w}$ denotes the ESS of the two-stage design when $\theta = \theta_w$, and $\bar{n}_{sw}$ denotes the ESS for a sequential design when $\theta = \theta_w$, $w = 0, 1$.

when $\theta = \theta_w$, $w = 0,1$.

It is somewhat surprising that two looks at the data can realize a substantial portion of the savings expected from a fully sequential design. For designs optimized under H_0, between 52% and 69% of the possible savings in sample size is realized with only two looks when H_0 is true. However, if $\theta = \theta_1$, then designs optimized under H_0 give much smaller gains. It is interesting, though, that in this case almost 50% of the savings can still be realized. The proportion of potential savings gained with a two-stage trial is not as great for designs optimized under H_1 (when $\theta = \theta_1$) as for designs optimized under H_0 (when $\theta = \theta_0$). Between 39% and 51% of the savings is gained with two looks when $\theta = \theta_1$. Note that these designs actually realize a greater proportion of the possible savings when $\theta = \theta_0$ than when $\theta = \theta_1$.

4.5 Two-Stage vs Three-Stage Designs

Table 4.4 gives the efficiency of the two-stage design relative to the three-stage design for several α and β combinations. Two-stage designs with equal sample sizes at each stage were used for this comparison since this restriction was used in obtaining the three-stage results. It can be seen that little is gained by the addition of a third stage for minimax designs when $\theta = \theta_{max}$. The savings are more substantial for these designs when $\theta \neq \theta_{max}$. The

Table 4.4. Efficiency of Two vs Three-Stage Design *

α	$1-\beta$	θ_0	θ_1	θ_{max}
.01	.80	81.2	93.3	94.0
	.90	82.5	91.8	94.6
	.95	83.7	89.6	94.8
.025	.80	84.0	93.0	94.1
	.90	85.2	91.2	94.6
	.95	86.5	88.9	94.8
.05	.80	86.6	92.7	94.1
	.90	87.6	90.7	94.5
	.95	88.8	88.2	94.7

* Efficiency = $(ESS_3(\theta)/ESS_2(\theta))*100$, where $ESS_i(\theta)$ is the expected sample size of an i-stage design for a given value of θ. Both designs are restricted to have equal sample sizes at each stage.

greatest benefits usually occur for the null and alternate Bayes designs when $\theta = \theta_0$ or $\theta = \theta_1$.

The efficiencies of several three-stage designs relative to their respective SPRT designs are listed in Table 4.5. As mentioned previously, SPRT designs give the maximum savings possible. As much as 75% of the potential savings can be realized with three looks. The greatest savings occur for designs optimized under H_0 when $\theta = \theta_0$. For these designs, approximately 50% of the savings is realized even when $\theta = \theta_1$. Between 53% and 69% of the savings is realized by three-stage designs optimized under H_1, the savings slightly greater when $\theta = \theta_0$ for larger β and slightly greater when $\theta = \theta_1$ for smaller β.

Let $P_s = n_f - \bar{n}_{sw}$, $P_2 = n_f - \bar{n}_{2w}$, and $P_3 = n_f - \bar{n}_{3w}$ denote the savings in sample size obtained using a SPRT design, a restricted two-stage design, and a restricted three-stage design, respectively. Then the proportion of maximum savings possible with two-stage designs is given by P_2/P_s; this proportion is given by P_3/P_s for three-stage designs. These results are listed in Tables 4.3 and 4.4 and have been described above. The additional savings in sample size gained by use of a three-stage design above that accounted for by a two-stage design is given by $(P_3-P_2)/P_s$ and is listed in Table 4.6. Whereas the two-stage design typically accounted for 50% or more of the possible savings, three-stage designs account for 40% or less of the remaining savings possible. This indicates that the proportionate

Table 4.5. Relative Efficiency of Restricted Three-Stage Design to the Fully Sequential Design *

α	$1-\beta$	Optimized under H_0		Optimized under H_1	
		θ_0	θ_1	θ_0	θ_1
.01	.80	71.8	46.1	69.2	56.0
	.90	73.8	45.0	67.0	59.6
	.95	75.1	47.1	59.0	63.4
.025	.80	71.2	48.5	68.9	57.3
	.90	73.0	51.0	65.5	60.6
	.95	74.0	53.4	54.8	63.8
.05	.80	70.3	52.8	67.8	58.3
	.90	71.7	55.3	62.6	61.2
	.95	72.2	58.0	52.7	63.9

* Relative Efficiency = $[(n_f-\bar{n}_{3w})/(n_f-\bar{n}_{sw})]*100$, where n_f denotes the fixed sample size, $\bar{n}_{3w}$ denotes the ESS of the three-stage design when $\theta = \theta_w$, and $\bar{n}_{sw}$ denotes the ESS for a sequential design when $\theta = \theta_w$, $w = 0, 1$.

Table 4.6. Additional Savings Realized With Three-Stage Design

α	$1-\beta$	Optimized under H_0 θ_0	θ_1	Optimized under H_1 θ_0	θ_1
.01	.80	37.3	12.9	34.3	26.3
	.90	38.8	5.5	26.0	28.7
	.95	39.6	4.0	4.0	32.1
.025	.80	36.7	13.4	33.4	27.0
	.90	37.9	13.4	22.5	29.6
	.95	38.1	14.2	-5.6	32.6
.05	.80	35.7	18.8	31.6	27.7
	.90	36.5	19.9	17.1	30.2
	.95	36.1	22.2	-8.0	32.9

* Additional Savings = $[(\bar{n}_{2w}-\bar{n}_{3w})/(\bar{n}_{2w}-\bar{n}_{sw})]*100$, where $\bar{n}_{2w}$ and $\bar{n}_{3w}$ denote the ESS of the two and three-stage designs when $\theta = \theta_w$, and $\bar{n}_{sw}$ denotes the ESS for a sequential design when $\theta = \theta_w$, $w = 0, 1$.

gain in efficiency is less at the third stage than at the second stage. The addition of an extra stage can result in a loss of efficiency when θ is not equal to the θ for which the design was optimized, but this is not common.

4.6 Other Group Sequential Designs

Among the more popular group sequential plans currently in use are those of Pocock (1977), O'Brien and Fleming (1979), and Haybittle (1971). Pocock suggests using a constant critical value at each stage. O'Brien and Fleming suggest using a sloping critical value which is stringent during the early interim analyses and is close to the fixed sample critical value at the final stage. Haybittle suggests using a standard normal critical value of 3.0 for each interim analysis followed by the fixed sample critical value at the final stage.

The expected sample sizes of these designs relative to the fixed sample size are shown in Table 4.7 for several α and β combinations. Since the three designs mentioned above do not allow acceptance of H_0 until the final stage, their performance under the null hypothesis is poor. They perform much better as one moves away from the null hypothesis. Thus, for comparison, I have chosen the restricted designs optimized under H_1. The relative expected sample sizes of these designs are also shown in Table 4.7. As expected, the differences in designs are particularly dramatic under the

Table 4.7. ESS / n_f for Various Designs[a]

α	$1-\beta$	Haybittle[b]	Pocock	O'Brien Fleming	Proposed[c]
.05	.80	0.99 0.88	1.14 0.82	1.00 0.89	0.70 0.79
	.90	0.99 0.83	1.13 0.74	1.00 0.83	0.77 0.74
	.95	1.00 0.77	1.12 0.65	1.00 0.78	0.88 0.69
.10	.80	1.00 0.92	1.13 0.81	1.00 0.86	0.78 0.78
	.90	1.00 0.87	1.12 0.71	1.00 0.80	0.86 0.73
	.95	1.00 0.82	1.11 0.63	1.00 0.74	0.89 0.67

[a] Top number is ESS/n_f when $\theta = \theta_0$.
Bottom number is ESS/n_f when $\theta = \theta_1$.

[b] Exact α-levels for these designs are .05171 and .10113.

[c] Designs optimized under H_1.

null hypothesis. In each case, the ESS is smallest for the restricted designs under the null hypothesis. Of course, the other designs could be improved in this respect by allowing acceptance of H_0 at any stage. Additionally, the ESS under the alternative hypothesis for the restricted designs is usually smaller than that of the other designs. It is curious that Pocock's designs outperform the restricted designs under the alternative hypothesis for large powers. The difference in these designs, other than allowing or not allowing early acceptance of H_0, is the restrictions placed on the cutpoints. Pocock sets $C_2 = C_4 = C_5$ while the restricted designs have $C_5 = \Phi^{-1}(1-\alpha)$. This seems to indicate that the unrestricted three-stage designs optimized under H_1 would be more similar to Pocock's designs for large powers. He did show that five-stage designs with large powers optimized under H_1 had approximately equal cutpoints at each stage (1982).

The ratio of the expected sample size to the fixed sample size is plotted against different alternatives in Figure 4.2 for $\alpha = 0.05$ and $1 - \beta = 0.80$ and 0.95. Haybittle's plans are not shown since their significance levels are greater than α. The horizontal axis of the plots shows the ratio of the true difference (Δ) to the difference specified by the alternative hypothesis (Δ_1). The values at 0 and 1 correspond to the values found in Table 4.7. Clearly the proposed plans are more efficient for most realistic values of Δ. Pocock's designs are somewhat more

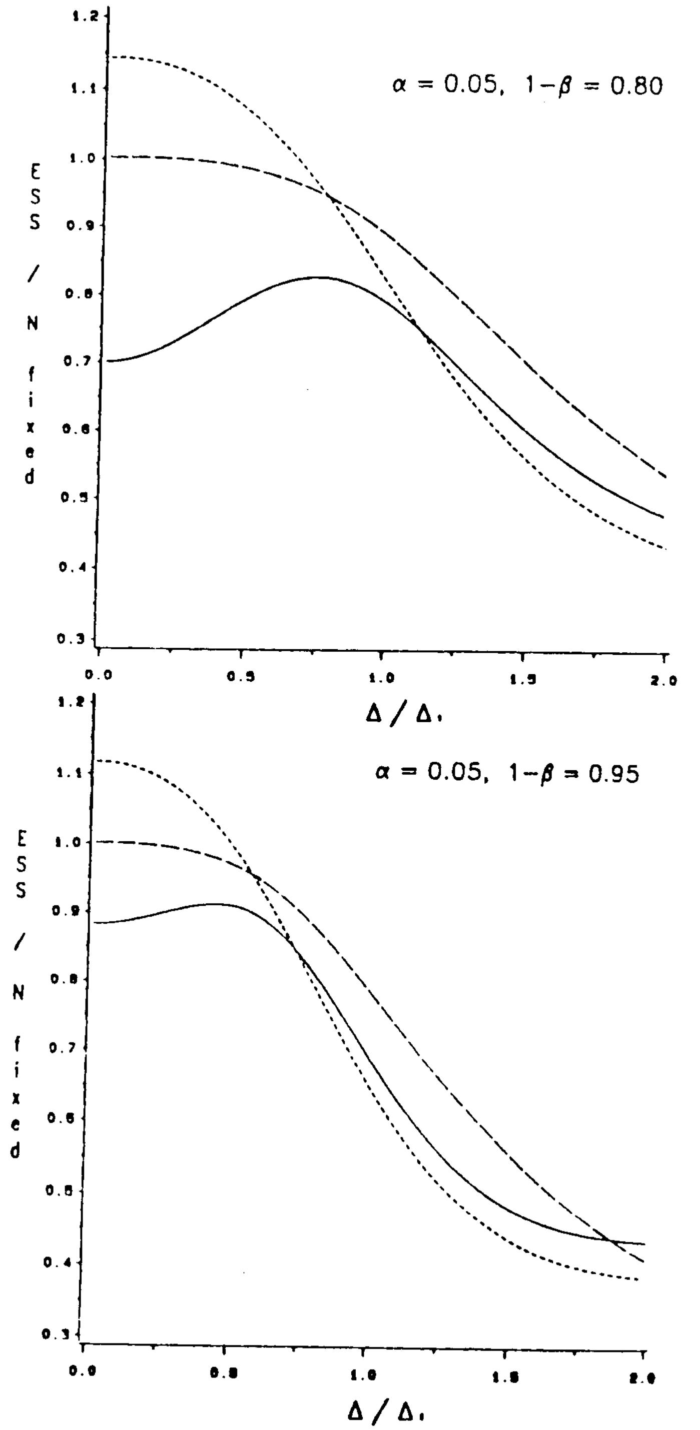


Figure 4.2. Expected sample size relative to the fixed sample size as a function of the true difference (Δ) relative to the difference specified by the alternative hypothesis (Δ_1) for various two-sided group sequential designs;
—— Proposed restricted designs,
---- Pocock's designs,
—–— O'Brien-Fleming designs.

efficient for larger Δ, especially for the plans with large powers. The designs of O'Brien and Fleming are only more efficient for extremely large Δ.

4.7 Deterministic Curtailment

Suppose a clinical trial is designed in which all N patients enter simultaneously and are followed for some endpoint, say survival. Alling (1963) showed that when a fixed number of observations are observed sequentially, the outcome of the Wilcoxon test based on the entire sample can usually be determined before all events are known. This method, sometimes called deterministic curtailment, has the desirable property that the early conclusion reached is the same conclusion which would have been reached had all events been observed.

The efficiency of deterministic curtailment depends on the sample sizes in the two groups as well as the true population difference. The efficiency is greater for smaller sample sizes and alternatives close to the null (Alling, 1963). Since the efficiency of deterministic curtailment depends on the sample size, it is difficult to compare it to the approach taken in this research. However, for illustration, the two approaches are compared for a specific trial which requires 20 patients per group to detect a difference of 0.786 standard deviations with 80% power at the 5% one-sided significance level (Figure 4.3).

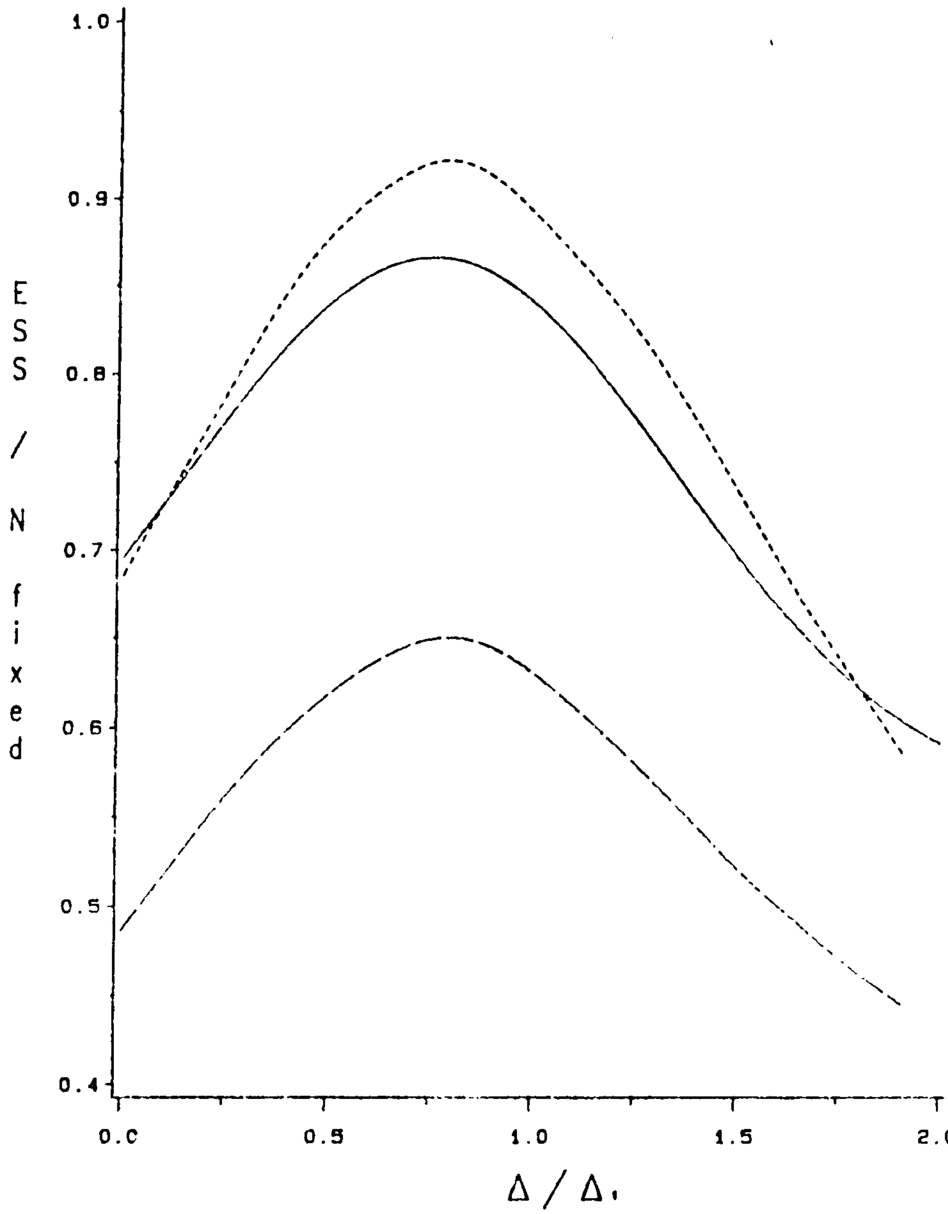


Figure 4.3. Expected sample size relative to the fixed sample size as a function of the true difference (Δ) relative to the difference specified by the alternative hypothesis (Δ_1) for various one-sided designs;
—— Proposed two-stage restricted design,
---- Two-stage deterministic curtailment,
—— Sequential deterministic curtailment.

Two deterministic plans are shown, a sequential plan in which test statistics are calculated after each event and a two-stage plan in which test statistics are calculated after every 10 events. Since deterministic curtailment is based on the entire sample, analysis after every event, or every other event, or every 10 events will all give identical α and β levels and the same results. However, it is clear from the figure that the expected sample size decreases rather dramatically with increasing reviews. In terms of the ESS, the two-stage deterministic plan is similar to the two-stage plan proposed in Chapter 2 (Table 2.7, W = 0), with the proposed plan slightly more efficient over the typical range of Δ. The sequential deterministic plan clearly gives the lowest ESS's. (Note that the ESS's for the deterministic curtailment plans were obtained from 1000 simulations at several values of Δ. Smoothed spline curves were fit to these points to generate the plots in Figure 4.3).

Since all patients must be accrued initially, the savings resulting from deterministic curtailment are really savings in observation time. Surviving patients could be treated with new therapies, but all patients will have received the original therapy. On the other hand, the proposed plans do not require that all patients enter simultaneously so savings can be realized in terms of patient numbers as well as time. However, the sample size required in a survival problem is usually the number of events needed rather than the number of patients. Thus, one must decide on appropriate durations of accrual and follow-up to obtain the necessary number of events. This may necessitate accrual of additional patients, partially or completely eliminating the 'savings in patients.' This problem is discussed in more detail in Chapter 5.

4.8 Discussion

We have seen that the restricted designs presented in Chapter 2 are almost fully efficient when compared to the unrestricted designs. Little is lost with the added restriction, and simplicity and ease of interpretation are gained. Additionally, the restricted designs do well relative to designs which do not allow acceptance of H_0 at the first stage. This comparison is somewhat moot since one must allow early acceptance to restrict the final stage critical value while maintaining the overall significance level. However, one should keep in mind that efficiency is always lost by not allowing acceptance of H_0 at the first stage. The restricted designs also provide a substantial proportion of the savings possible using a fully sequential design. It is exciting that, in some cases, more than 50% of the possible savings can be realized with only one interim look at the data and 75% of the possible savings can be realized with just two interim looks. The greatest savings per stage occur in going from a fixed sample to a two-stage design. Approximately 40% to 50% of the potential savings can be realized with this one additional look. The proportionate reduction in sample size is not as great in going from two to three stages. Here, approximately 15% to 25% of the remaining savings can be recouped. Additionally, the restricted designs do well compared to other group sequential designs presented in the literature. The efficiency of the restricted designs is consistently greater than that of the O'Brien-Fleming designs over a wide range of Δ. The efficiency of Pocock's designs is slightly greater for larger Δ, particularly for designs with large powers. However, Pocock's plans do very poorly for small Δ, whereas the restricted plans do well.

V. APPLICATIONS

5.1 Introduction

Examples are presented in this chapter which illustrate the use of the restricted two and three-stage plans developed previously. Hypothetical designs are given which cover problems typically encountered in clinical trials. These include the comparison of means with known or unknown variance, the comparison of binomial responses, and the comparison of survival times. Additionally, these designs are applied to several completed single stage trials of the Piedmont Oncology Association (POA) to illustrate the savings in sample size which would have occurred had these designs been used. Finally, a current trial designed using a nonoptimal three-stage stopping rule is presented.

5.2 Normal means - known or unknown variance

Suppose we are designing a clinical trial with equal allocation to test an experimental drug versus a control drug in reducing the cholesterol level of adult males. The parameter of interest is the difference in mean cholesterol reduction between the two populations and is given by $\theta = \mu_e - \mu_c$. The hypothesis to be tested is H_0: $\theta = 0$ versus the alternative H_1: $\theta > 0$. The estimator of θ, denoted by $\hat{\theta}$, will be the difference in the sample mean reduction for each group. Let s denote the common standard deviation of cholesterol reduction in the two groups. The trial is designed to detect a decrease in the mean reduction in the experimental group compared to the control group of one-half a standard deviation with 80% power at the 5% one-sided level of significance. The standard deviation of $\hat{\theta}$ is $(\sqrt{2}s)/\sqrt{n}$, where n is the sample size in each group. For this example, $\Delta^2 = (1/(2\sqrt{2}))^2 = 0.125$. The fixed sample size required for this problem is $6.18/(0.125) = 49.4$ or 50 patients per group. A two-stage study can be designed using the information in Tables 2.2 - 2.5. Suppose we desire to mimimize the maximum expected sample size. The cut points for this design can be obtained from Table 2.2. We see that the maximum sample size in each group is $1.13(49.4) \simeq 56$ patients. Of these, $56(0.557) \simeq 32$ are accrued during the first stage and the remaining 24 during the second stage. For this particular trial, the sample size will be 32 with probability 0.538 and 56 with probability 0.462, assuming $\theta = \theta_{max}$. The expected sample size is $.86(50) \simeq 43$ if $\theta = \theta_{max}$. The ESS will be less if $\theta \neq \theta_{max}$. For example, the ESS is approximately 36 if $\theta = \theta_0$ and 41.5 if $\theta = \theta_1$. Note that it is not necessary to know the variance explicitly. One need only specify the Δ^2 under consideration to determine the required sample size.

5.3 Binomial Responses

Suppose a clinical trial is designed to compare the response rates of patients receiving standard and experimental chemotherapy regimens. Patients are to be randomized to the two treatments with equal allocation. The hypothesis of interest is H_0: $\theta = 0$ versus H_1: $\theta \neq 0$, where θ denotes the difference in the true response rates to the two regimens. The estimator of θ, $\hat{\theta}$, will be the difference in the observed response rates. A two-stage two-sided plan optimized under H_1 can be designed using information in Table 2.12. Suppose the response rate to the standard regimen is 40% and the investigator is interested in detecting an increase to 60% with 90% power at the 1% two-sided level of significance. The total sample size required for the fixed sample case is approximately $14.88 \times 2(.4 \times .6 + .6 \times .4)/.04$ or 358. For the two-stage design, the maximum sample size required is approximately $358(1.1) = 394$. Of these, $394(.584) \simeq 230$ are accrued during the first stage and 164 during the second stage (if required). The probability of stopping at the first stage is .618 if $\theta = \theta_1$. The sample size under this alternative will be 230 with probability .618 and 394 with probability .382, and the ESS is 292.6. The ESS is 248.6 if $\theta = \theta_0$ and 314.0 if $\theta = \theta_{max}$.

One can calculate the normal statistic at each stage for comparison with the cut points shown in the table. Alternatively, one can write the cut points in terms of differences for ease of interpretation, where the critical differences = $C/\sqrt{N}$. Since $C_1 = 1.569$ and $C_2 = 2.958$, a difference less than $1.569/\sqrt{230} = 0.10346$ would lead to acceptance of H_0 and a difference greater than $2.958/\sqrt{230} = 0.19504$ would lead to rejection of H_0. The critical difference at the final stage would be $2.576/\sqrt{393} = 0.12994$. The level of significance and power of this design as determined by 10,000 simulations are 0.008 and 0.896, respectively. The probability of stopping at the first stage and the ESS are .590 and 297.2 if $\theta = \theta_1$ and .885 and 248.9 if $\theta = \theta_0$. These numbers for the binomial distribution are close to their theoretical values based on the normal approximation.

5.4 Survival Times

Now consider designing a clinical trial to compare experimental and control drugs in which survival is the endpoint of interest. A proportional hazards model (PHM) will be assumed where θ, the parameter for treatment effect, is the logarithm of the hazard ratio between the two groups. The Cox model (1972) will be used to estimate the parameter. Tsiatis et al (1984) have shown that the asymptotic joint distribution of the PHM score statistic calculated at fixed time points is multivariate normal. Their simulation studies indicate that this statistic is appropriate in designs which require analyses after fixed numbers of deaths. The Wald statistic which is used in this example is asymptotically equivalent to the score statistic. Schoenfeld (1983) has previously derived the fixed sample size formula for the PHM problem, where the sample size is actually the number of deaths required. Assuming equal accrual to both regimens, $\Delta^2 = 0.25(\theta^2)$. Here we test the null hypothesis H_0: $\theta = 0$ versus the alternative hypothesis H_1: $\theta \neq 0$. Assume that we are interested in detecting a hazard ratio of 2.0 with 80% power at the 5% two-sided level of significance. Assume also that we are interested in accruing equal samples at each stage. The variance of $\hat{\theta}$ under H_0 is given by $0.25/f(n)$, where $f(n)$ is the expected number of deaths in a sample of size n (Schoenfeld, 1983). For this problem, $\Delta^2 = 0.25(\ln(2))^2 = 0.120$. The necessary number of deaths for the fixed sample case is $7.85/(0.120) = 65.4$ or 66. Parameters for a three-stage design optimized under H_0 can be obtained from Table 3.8. For this design, the maximum number of deaths required is $1.363(65.4) \simeq 90$. For this particular trial, the number of deaths required is 30 with probability 0.602, 60 with probability 0.290, and 90 with probability 0.108, assuming $\theta = \theta_0$. The expected number of deaths required is approximately 45.2 if $\theta = \theta_0$, 52.9 if $\theta = \theta_1$, and 55.0 if $\theta = \theta_{max}$.

The actual number of patients required will depend on the rate of patient accrual and the death rate for the particular disease under study. For studies in which death usually occurs shortly after accrual, the total sample size will be close to the number of deaths required. For studies in which patients enjoy long survival times, the total sample size might be several times the required number of deaths. One approach is to accrue a sample size equal to the number of deaths required for first stage evaluation and follow each of these patients until death. Problems with this approach include the possibility of a long follow-up period, uncertainity of how to treat new patients during this time, and the difficulties involved in resuming the study once the interim review is complete. Alternatively, patients could be accrued until the required number of deaths is observed. A problem with this approach is that many more patients than necessary may be accrued to the study. Additional work needs to be done to determine cost effective durations of accrual and follow-up for multi-stage designs.

5.5 Savings in Retrospect

In this section, the two and three-stage designs of the previous chapters are applied to several completed single stage studies of the POA to illustrate the results which would have occurred had those designs been used. These results will only be approximate since it is not possible to exactly recreate the data files as they would have appeared at the time. Some patients become eligible over time and others become ineligible. Response data are continually reevaluated, changing the status of many patients throughout the study.

All of the examples which follow are oncology clinical trials designed to compare response rates to various chemotherapy regimens. The design specifications were not always clearly spelled out in the earlier studies. For example, one study read that a sample size of 100 patients was 'statistically acceptable.' I have taken the liberty to furnish specifications when necessary. Most fixed sample sizes were calculated using the normal approximation; some were calculated using an arc sine transformation. I have used the normal approximation for all of the multi-stage sample sizes. The two and three-stage designs optimized under H_0 are arbitrarily selected to illustrate the results which might have occurred. These are not necessarily the designs which would have been chosen at the time.

5.5.1 Protocol 74181

This randomized phase III trial was designed to compare the response rates of receptor positive patients with advanced breast cancer undergoing hormonal therapy with either megestrol acetate or tamoxifen (Muss et al, 1985). Literature on breast cancer suggested that the response rate to hormonal therapy was approximately 50% in patients with receptor positive tumors. This study was designed to detect a difference in response rates of 20% with 80% power at the 5% two-sided level of significance. Approximately 180 total patients were to be accrued.

Had a two-stage study been used, a maximum of 224 patients would be required, 102 of whom would be accrued during the first stage. At the first stage, 15/50 (30%) megace patients responded compared to 16/52 (31%) tamoxifen patients. The normal statistic for this comparison is 0.084 and would have led to acceptance of H_0 at the first stage and discontinuation of the trial after 102 patients.

Had a three-stage study been designed, approximately 82 patients would be necessary at each stage. At the first stage, 14/40 (35%) megace and 13/42 (31%) tamoxifen patients responded, resulting in a normal statistic of 0.390. This would have led to acceptance of H_0 and discontinuation of the trial after only 82 patients.

This trial was discontinued early, but not because of the similarity in response rates. It was stopped after 136 patients were accrued because of a large difference in survival between the two regimens. Since the study was stopped early, it is not possible to say what the results would have been had all 180 patients been accrued. However, the difference in response rates was still not statistically significant at closure. It is interesting that the comparison of survival curves was not even a major objective. But the difference was so striking that the investigators felt it unethical to continue accruing patients to the trial. This points out a potential problem with group sequential designs. It is doubtful that the difference in survival would have been significant after the first 82 patients. At that point very few deaths would have occurred. Had the results been published after 82 patients, it is not clear that the survival difference would ever have been detected. The problem is that these plans are designed to compare one endpoint. Comparison of other endpoints must be made with caution.

5.5.2 Protocol 62279

This randomized phase III trial was designed to compare the response rates of advanced lung cancer patients treated with either CAV or VCAV (Jackson et al, 1984). The study was to accrue 100 evaluable patients, allowing detection of a 25% difference in response rates (0.6 to 0.85) with 80% power at the 5% two-sided level of significance.

A two-stage study would have required a maximum of 114 patients, 52 of whom would be accrued during the first stage. Of the first 52 evaluable patients, 16/27 (59%) on CAV and 20/25 (80%) on VCAV responded, resulting in a normal statistic of 1.619. No decision would have been reached at the first stage and accrual would have continued. After 114 patients, 32/54 (59%) on CAV and 50/60 (83%) on VCAV had responded. The normal statistic for this comparison is

2.856 and results in rejection of H_0.

A three-stage study would have required 126 patients, 42 of whom would be accrued at each stage. After the first 42 evaluable patients, 14/23 (61%) on CAV and 16/19 (84%) on VCAV had responded, resulting in a normal statistic of 1.667. No decision could have been made at this stage. After the next 42 patients, 25/42 (60%) on CAV and 33/42 (79%) on VCAV had responded. The normal statistic for this comparison is 1.888, and again, no decision could have been made. After 126 patients, 35/59 (59%) on CAV and 57/67 (85%) on VCAV had responded. The normal statistic for this comparison is 3.250 and leads to rejection of H_0.

Of the 100 patients in the fixed sample design, 29/48 (60%) responded to CAV and 43/52 (83%) responded to VCAV ($p = 0.013$). Thus the fixed sample design and the two and three-stage designs would have led to the same conclusion. In this example, the fixed sample design actually required fewer patients. (Note that although only 100 patients were required, many more were accrued while other protocols were developed. Thus we could determine what would have happened at the last stage of the group sequential designs. When the protocol was finally closed, the difference in response rates was still statistically significant.)

5.5.3 <u>Protocol 59177</u>

This randomized phase III trial was designed to compare the response rates of advanced colorectal cancer patients treated with either 5-FU or 5-FU plus mitocycin-C and methyl-CCNU (MMF) (Richards et al, 1986). Approximately 140 total patients were to be randomly accrued, giving 80% power to detect an increase in response rates from 15% (5-FU) to 35% (MMF) at the 5% two-sided level of significance.

A two-stage study would have required a maximum of 173 patients, 79 of whom would be accrued during the first stage. Of the first 79 patients, 8/40 (20%) on 5-FU and 4/39 (10%) on MMF responded. This comparison results in a normal statistic of 1.206; no decision could have been reached, and a second stage would have been required. It is not known what the second stage results would have been since accrual stopped after 140 patients.

A maximum of 190 patients would have been required for a three-stage study, approximately 63 to be accrued at each stage. After first stage accrual, 7/32 (22%) on 5-FU and 3/31 (10%) on MMF had responded, resulting in a normal statistic of 1.324 and continuation of the trial. After the second stage accrual, 10/64 (16%) on 5-FU and 4/62 (6%) on MMF had responded, resulting in a normal statistic of 1.636 and continuation of the trial. Since accrual stopped, we do not know what the results would have been after the third stage accrual.

This study actually accrued 140 patients. However, only 126 were eligible and evaluable for response so the final results of this study were the same as they would have been at the second stage of the three-stage study. In this example, both multi-stage designs would have required more patients than the fixed sample design, and it is not known if the ultimate conclusions would have been the same.

5.5.4 Protocol 62179

This nonrandomized single arm phase II study was designed to estimate the response rate of advanced non-small cell lung cancer patients treated with CMAP (Richards et al, 1984). One hundred patients were to be accrued to the study, giving 90% power to detect a increase of 0.155 over the typical response rate of 0.2 at the 5% two-sided level of significance.

A maximum of 120 patients would have been required for a two-stage trial, 60 of whom would be accrued during the first stage. Nine (15%) of the first 60 patients responded, resulting in a normal statistic of 0.968 and acceptance of H_0 at the first stage. A three-stage design would have required a maximum of 138 patients or 46 per stage. Eight (17%) of the first 46 evaluable patients responded, resulting in a normal statistic of 0.442 and acceptance of H_0 at the first stage.

Eighteen (18%) of the 100 patients required in the fixed sample design responded. The normal statistic for testing the null hypothesis that the true response rate is 20% versus the alternative that the true response rate is not 20% is 0.5 and leads to acceptance of H_0. Both group sequential designs led to the same conclusion with sample sizes much less than that of the single stage design. (Note, 171 evaluable patients were actually accrued to this study to increase the numbers in certain subgroups. A response rate of 19% (32/171) was observed overall.)

5.6 Current Example

Agressive chemotherapy can often lower a patient's neutrophil count, leaving him highly susceptible to infection. This infection often arises in the area of the catheter supplying the chemotherapy. A clinical trial was designed at Bowman Gray School of Medicine (BGSM) to compare the incidence of infection in patients receiving therapy via a teflon plastic intravenous device and those receiving therapy via a vialon plastic device. It was hoped that the vialon plastic device would reduce the incidence of infection, but it was possible that it could actually increase infections.

Patients were randomly assigned with equal allocation to either the teflon plastic or the vialon plastic device. It was anticipated that approximately 300 patients could be accrued in the six months during which the study was to run. This would allow detection of a 16% difference in the proportion of patients infected in the two groups with 80% power at the 5% two-sided level of significance. Interim reviews were to be undertaken twice during the study (after 1/3 and 2/3 of the patients were accrued) to determine if: (a) either device resulted in a substantially lower incidence of infection, or (b) the two devices are not different enough to warrant continuation of the study.

A three-stage clinical trial was designed which allowed detection of a 16% difference (50% vs 34%) in infection rates with 80% power at the 5% two-sided level of significance. Additionally, the critical value at the final stage was set equal to that of a fixed sample design to allow for easy interpretation should a final analysis be necessary. The three-stage work presented in Chapter 3 had only just begun when this design was required. A bit of trial and error showed that values for C_3 and C_4 of 1.4 and 2.4 were close to optimal for a one-sided 0.025 α-level design. Since one-sided $\alpha/2$-level designs can be used as approximations for two-sided α-level designs, these values were fixed along with $C_5 = 1.96$, and C_1 and C_2 were chosen to minimize the ESS under the alternative hypothesis. This resulted in $C_1 = 0.28$ and $C_2 = 2.49$. The maximum sample size was chosen to satisfy the power requirement and was 1.19 times the fixed sample size or approximately 360 patients. The cut points at each stage were translated to differences in proportions for ease of understanding. The final stopping rule is shown below.

Stopping Rule: The following criteria will be used for possible early closure after accrual of 120 eligible patients (60 on each device) and 240 eligible patients (120 on each device). Note that d is the difference in the proportion of patients who become infected using the two devices.

Stage 1 - (after 120 pts) -	$\lvert d\rvert < .02$, stop and accept H_0
	$\lvert d\rvert > .22$, stop and reject H_0
	$.02 < \lvert d\rvert < .22$, continue.
Stage 2 - (after 240 pts) -	$\lvert d\rvert < .09$, stop and accept H_0
	$\lvert d\rvert > .15$, stop and reject H_0
	$.0904 < \lvert d\rvert < .1549$, continue.
Stage 3 - (after 360 pts) -	$\lvert d\rvert \leq .10$, accept H_0
	$\lvert d\rvert > .10$, reject H_0.

The level of significance and power of this design as determined by 10,000 simulations were 0.052 and 0.809, respectively. The probabilities of stopping at each stage and the expected sample sizes are shown below. These values were also determined by simulations. The theoretical values are shown in parentheses. For this trial, the sample size will be 120, 240, or 360 with probabilities .23, .63, and .14 under H_0 and .28, .42, and .30 under H_1. It is likely that the trial will be closed early (assuming only the stopping rule is considered). However, it is possible that the study will need to be extended an extra month to accrue the 60 additional patients required by this design (above that needed for a fixed sample design) if it is not stopped early.

True difference in proportions			
H_0 (d = 0)		H_1 (d = .16)	
Sample Size	Probability *	Sample Size	Probability *
120	.231 (.236)	120	.283 (.285)
240	.628 (.636)	240	.419 (.430)
360	.141 (.128)	360	.298 (.285)
ESS = 229.2 (227.0)		ESS = 241.8 (240.0)	

* The probabilities were determined by 10,000 simulations. The numbers in parentheses are the theoretical probabilities.

The optimal parameters for this design can be found in Table 3.9. The expected sample size under H_1 is 0.79(300) = 237, which is very close to 242 obtained using the nonoptimal design. The ESS under H_0 would be 210, which is somewhat less than 230 expected with the nonoptimal design. For the optimal design, C_1 = 0.62 compared to 0.28 for the nonoptimal design. This greater value for C_1 results in a greater probability of stopping during the first stage, leading to the lower ESS under H_0.

5.7 Discussion

The multi-stage designs developed in Chapters 2 and 3 are easily applied, and several examples have been given which illustrate their use in typical situations. We have seen that sample size savings could have occurred had these designs been used for some of the clinical trials previously carried out in the POA. However, other clinical trials would have required more patients. For any particular trial, both the maximum sample size and the expected sample size should be considered. A design with a lower maximum sample size and a slightly higher expected sample size might be preferable to one with a higher maximum sample size and a lower expected sample size. These designs have also been used in actual clinical trials. One such example was given in section 5.5. That example, albeit an example of a nonoptimal design, illustrates how a multi-stage study might be planned.

In the following chapter, we review the advantages and disadvantages of these designs and give some discussion on the choice of appropriate designs.

VI. DISCUSSION

6.1 Introduction

In the previous chapters, two and three-stage designs were constructed which satisfied the ethical demand for interim reviews while retaining the simplicity of a fixed sample design at the final stage. These designs not only satisfy ethical needs, but also result in sample size savings in expectation. The expected sample size can be as little as one-half that of a fixed sample design, and up to 75% of the savings expected with a Wald SPRT design can be realized. In this chapter, the motivation for these designs is reexamined, and some of the issues to consider in choosing appropriate designs are discussed. Some possible extensions are also considered.

6.2 Why These Designs?

This question really has two parts. First, why do a group sequential design in general? Second, why choose these designs in particular? Both of these questions were addressed in Chapter 1. I'll briefly review the highlights here. In most clinical trials involving human subjects, there is an ethical concern for each patient's well-being.

Investigators want to ensure that no patient is receiving an inferior treatment, suffering any toxic side-effects, or experiencing any adverse reactions to their therapy. To this end, data from a clinical trial are often reviewed one or more times before the study is complete. During these reviews, the investigators usually test their major hypotheses. They do not want to continue a study whose conclusion is almost certain partway through the trial. This is costly in terms of time, money, and patient resources. The problem from a statistical point of view is that multiple interim hypothesis tests carried out at a fixed significance level α result in an overall significance level for the trial which is greater than α. Thus, some form of statistical design is needed which allows for interim reviews while maintaining a specified overall significance level. Sequential and group sequential designs were developed for this purpose.

The problem then becomes one of complexity and interpretation. Multi-stage designs <u>are</u> more complex than fixed sample designs. They are more difficult to plan, to implement, and to execute. They require more investigator interaction, more timely record keeping and data processing, and more analyses. However, if the investigator is already planning on interim reviews then a formal design is a necessity. Problems of interpretation occur partly because of the increased design complexity and partly because of the investigators lack of statistical sophistication. Most group sequential designs use critical values for rejection at each stage which are greater than the fixed sample value. Investigators, whose training usually includes only one or two introductory statistics courses, are confused when their test statistic is greater than the customary 1.645 (or 1.96), and yet they cannot reject the null hypothesis. This confusion is exacerbated during a final review when 'positive' results are perceived as more publishable than 'negative' results. Investigators might feel that they were deprived of their publishable p-value much as Tantulus is deprived of his thirst-quenching water and that their sure paper was mysteriously vanished by some sort of statistical magic. Such seeming contradictions do not inspire good working relations between investigators and statisticians.

As mentioned above, most group sequential designs use critical values for rejection at each stage which are larger than that of a fixed sample design. However, it is interesting that designs can have a critical value at the final stage which is smaller than that of a fixed sample design. For example, optimal unrestricted two-stage two-sided $\alpha = 0.05$ designs optimized under H_0 have second stage critical values ranging from 1.82 to 1.88 as the power ranges from 80% to 95%. It might even confuse statisticians to report that a test statistic of 1.85 led to rejection of H_0 at the 0.05 two-sided level of significance!

The preceeding remarks indicate that confusion can arise because of critical values for rejection which are too high or too low. Much of this confusion is due to the investigators lack of statistical sophistication and much is due to our overemphasis on p-values and fixed sample designs. This confusion will continue until investigators understand the principles of sequential design (which is not likely to happen soon). To alleviate the confusion somewhat, designs were needed which allowed for multiple interim reviews while maintaining the simplicity of a fixed sample design at the final stage. The designs in this research were developed with that goal in mind.

6.3 Which of these Designs?

The designs presented in this research minimize the maximum expected sample size or the expected sample size for particular values of θ, given certain restrictions. An appropriate question might be which of these designs should be used in planning a particular trial. It is clear that no one design is 'best' for every alternative. That is, the plan that is 'best' when $\theta = \theta_0$ is not 'best' when $\theta = \theta_1$ and vice versa. The choice of a particular design is not clear-cut. If the choice is based solely on the ESS, then one must make some qualitative decision as to the likelihood of a particular alternative. If one is almost positive that H_0 (H_1) is true, then the design of choice would be the one which minimizes the ESS under H_0 (H_1). Of course, one must question why such a trial is undertaken in the first place. If one feels that H_0 and H_1 are equally likely, he might choose the design which minimizes the average of $ESS(\theta_0)$ and $ESS(\theta_1)$. Although results were presented only for $W = 0$, 1/2, and 1, designs can be optimized for any specific weights. However, the choice of weights is subjective.

Rather than averaging risk as a Bayesian would, one might choose to minimize the maximum risk. That is, he would use the minimax criterion to choose the design which minimizes the maximum ESS. This criterion seems appropriate if one is pessimistic and believes that θ will be that value which maximizes the ESS. It is conservative in the sense that it does not lead to as small an ESS as the Bayes criterion for particular θ. However, it is safe in the sense that it ensures one of having the lowest maximum ESS.

The ratio of the expected sample size to the fixed sample size is plotted against Δ/Δ_1 in Figures 6.1 - 6.4 for several α and β combinations. These figures show one and two-sided, two and three-stage designs obtained using the various criteria. These plots illustrate how the designs perform relative to one another over a range of Δ. One interesting feature of these plots is the similarity of the minimax designs and the designs optimized for $\theta = \theta_1$. They are practically indistinguishable for the two-stage plans with 80% power. The two-stage designs optimized for $\theta = \theta_1$ are slightly more efficient over a wide range of Δ, indicating that the minimax two-stage designs are probably

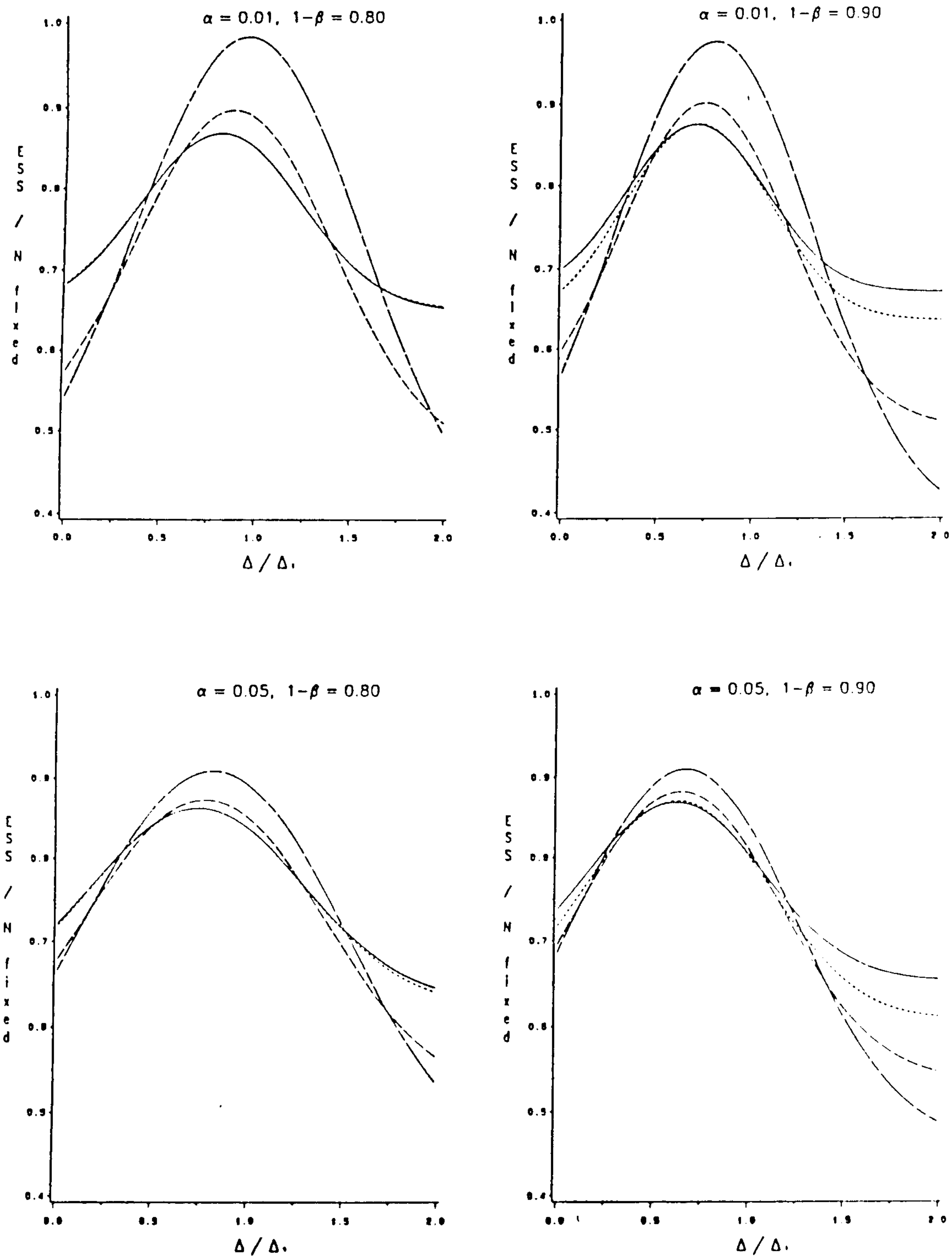


Figure 6.1. Expected sample size relative to the fixed sample size as a function of the true difference (Δ) relative to the difference specified by the alternative hypothesis (Δ_1) for various two-stage one-sided designs;
——— Minimax criterion,
- —— Bayes criterion, W=0,
— — Bayes criterion, W=0.5,
----- Bayes criterion, W=1.

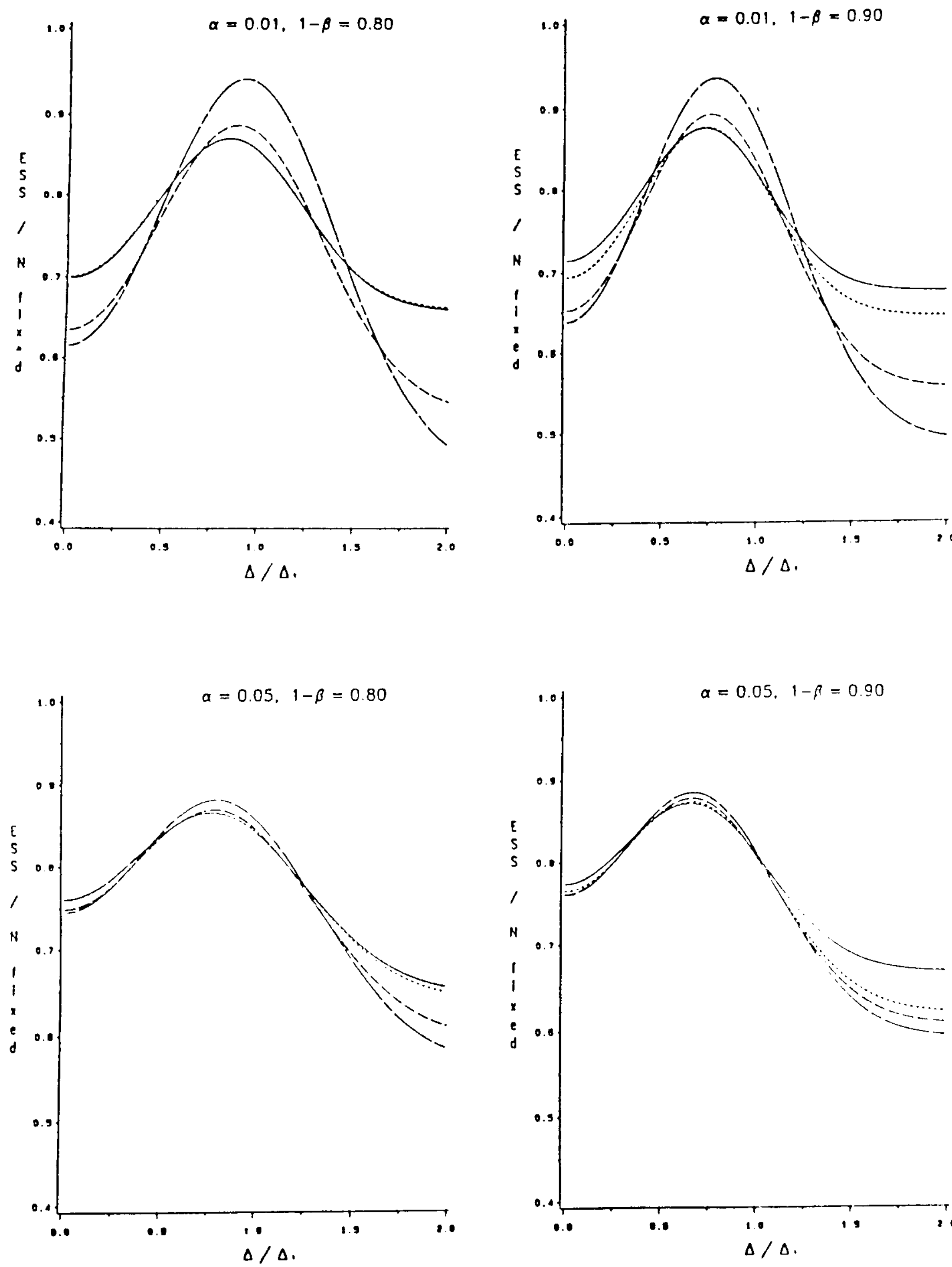


Figure 6.2. Expected sample size relative to the fixed sample size as a function of the true difference (Δ) relative to the difference specified by the alternative hypothesis (Δ_1) for various two-stage two-sided designs;
——— Minimax criterion,
- —— Bayes criterion, W=0,
— — Bayes criterion, W=0.5,
----- Bayes criterion, W=1.

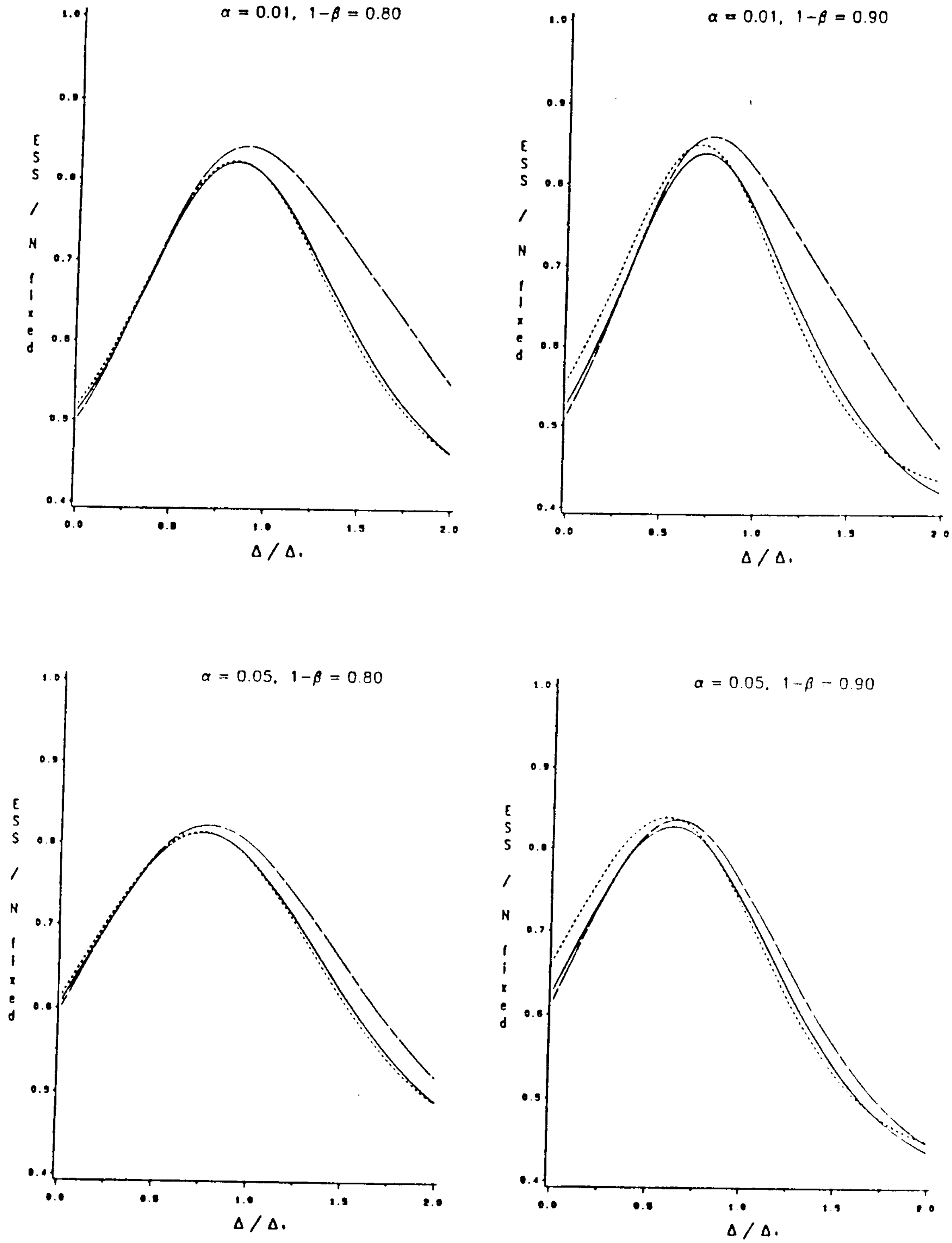


Figure 6.3. Expected sample size relative to the fixed sample size as a function of the true difference (Δ) relative to the difference specified by the alternative hypothesis (Δ_1) for various three-stage one-sided designs;
——— Minimax criterion,
- —— Bayes criterion, W=0,
----- Bayes criterion, W=1.

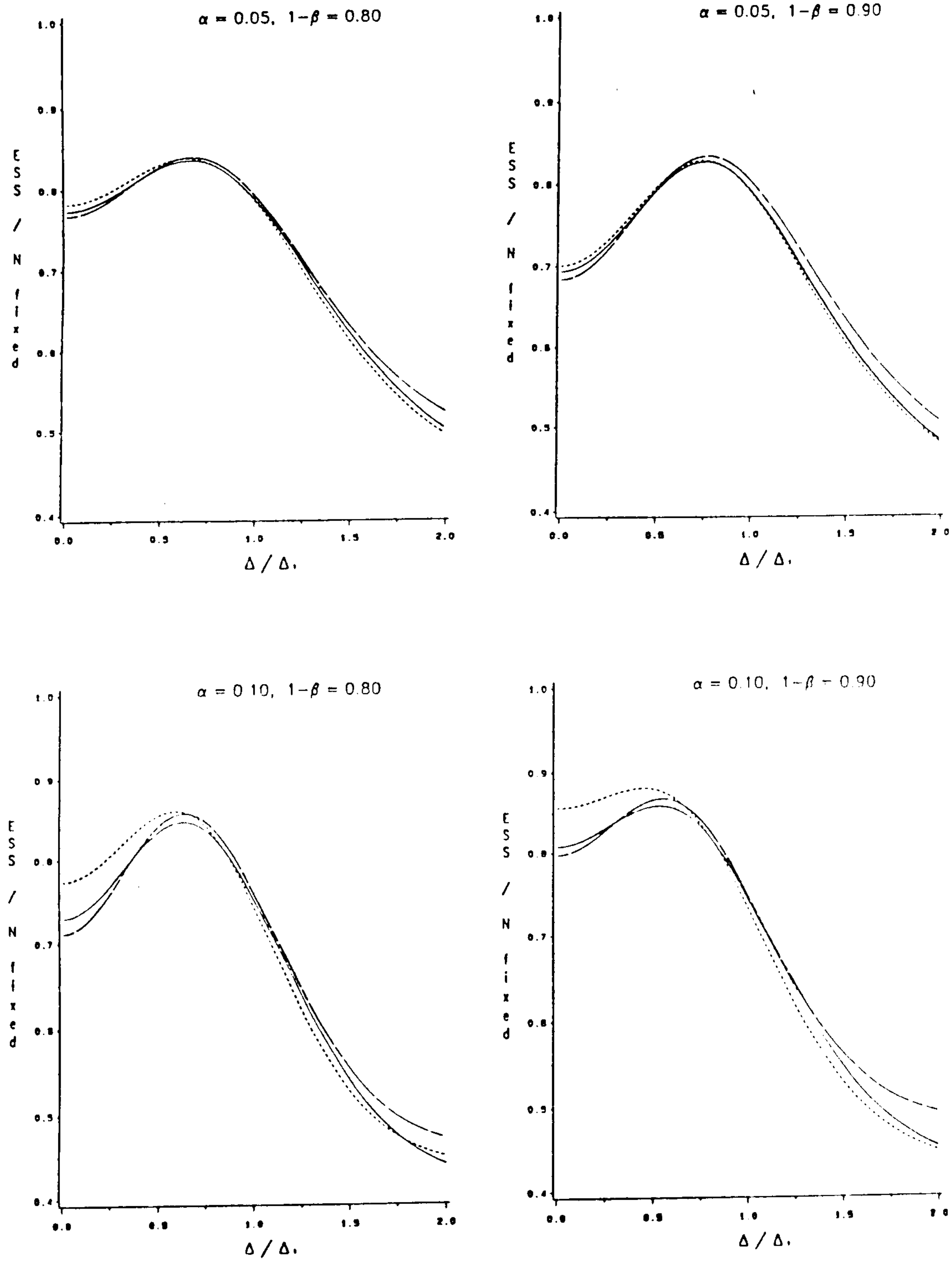


Figure 6.4. Expected sample size relative to the fixed sample size as a function of the true difference (Δ) relative to the difference specified by the alternative hypothesis (Δ_1) for various three-stage two-sided designs;
——— Minimax criterion,
- —— Bayes criterion, W=0,
----- Bayes criterion, W=1.

too conservative. The designs optimized for $\theta = \theta_0$ have the lowest ESS for very small and very large differences; however, these designs are the least efficient for differences between 0.5 and 1.5 times that specified by the alternative hypothesis. The Bayes weighted average criterion with W = 0.5 seems to produce the 'best' designs over the range of Δ pictured. The ESS of these designs is never much greater than the lowest ESS of the other designs, and it is sometimes smaller than the others.

The three-stage designs are all similar over most of the range plotted. The plans optimized for $\theta = \theta_0$ perform worst for large Δ while the plans optimized for $\theta = \theta_1$ are worst for small Δ. The minimax plans seem to do fairly well over the entire range shown. Another consideration in choosing a design for a particular trial is the maximum sample size. Of the two-stage designs, those optimized for $\theta = \theta_0$ have the largest maximum sample sizes; the other plans have comparable values. The maximum samples sizes are also comparable for the different three-stage designs.

6.4 Possible Extensions

Of course, the obvious extension to this research is to develop K-stage plans which allow acceptance and rejection at each stage and use the fixed sample critical value at the final stage. However, this is not computationally feasible. The three-stage design was already computationally burdensome, even when restricting the sample sizes to be equal at each stage. Numerical integration of the multivariate normal distribution becomes prohibitively expensive as the number of dimensions becomes large. (For example, one extreme five dimensional integration took over three hours of CPU time on a VAX 11-750 computer.) Glaz and McK. Johnson (1986) developed equations for approximating multivariate normal areas which are fairly accurate and can be calculated in approximately one second. Their methods could be used in developing approximately correct multi-stage designs. However, each additional stage also adds three additional parameters, and the problem soon becomes unwieldy even when the areas can be calculated quickly. To reduce the number of parameters, the critical values for acceptance and rejection might be selected so they are a linear or quadratic function of the stage number. This is similar to the idea used by Pocock (1977), O'Brien and Fleming (1979), and Canner (1977) in developing their plans. The regression coefficients could be chosen to minimize the expected sample size using the criteria of Chapters 2 and 3. Also, one could modify the plans mentioned above to use the fixed sample critical value at the final stage by allowing early acceptance. For example, a simple modification of Pocock's (1977) plans might be to have constant critical values for rejection and acceptance at the first K-1 stages and the fixed sample critical value at the K^{th} and final stage.

6.5 Summary

In addition to α, β, and Δ, an investigator planning a multi-stage study must also choose the number of interim analyses and a design which maintains the overall significance level. Several designs are available for this purpose; these were summarized in Chapter 1. In choosing the design, the investigator must decide whether or not to allow acceptance or rejection at each stage and whether or not it is important for the critical value at the final stage to match that of a fixed sample design.

The multi-stage designs presented in this research allow acceptance or rejection at any stage. Early rejection of H_0 allows for termination of trials in which a difference is detected prior to the final review. In this case it may be unethical to expose additional patients to an inferior treatment. Early acceptance of H_0 allows for the termination of trials which show no treatment benefit. This allows future patients to be treated with new and more promising regimens. Additionally, these designs use a final stage critical value equal to that of the fixed sample test so investigators are not troubled by seemingly inconsistent results during a final review. If it is desirable to allow early acceptance and for the fixed sample critical value to be used at the final stage, then the designs presented in this research provide a practical and efficient method for interpreting the results of hypothesis tests during interim and final reviews.

APPENDIX

The program shown on the following pages determines the optimal two-stage parameters for minimizing the maximum ESS or a weighted average of ESS(θ_0) and ESS(θ_1) given the restriction that $C_3 = \Phi^{-1}(1 - \alpha)$. Although the one-sided two-stage design is the simplest of multi-stage designs, the program used in its evaluation does illustrate the basic algorithm used to obtain all of the results presented in Chapters 2 and 3. For two-sided designs, the area and power calculations are somewhat more complicated and the evaluation of θ_{max} requires a separate iterative calculation. The determination of area and power for three-stage designs is much more complicated, requiring several evaluations of the trivariate normal and bivariate normal distributions. In addition, the evaluation of θ_{max} requires separate iterative calculations for both one and two-sided designs.

This program requires several subroutines. These can be found in most computer algorithm books or in packages such as IMSL. The following subroutines are needed for the one-sided two-stage case:

(1) BMIN and FMIN - two one-dimensional subroutines for iterating over p and C_1. These routines are identical; the

names have been changed to reduce confusion. Both are golden-section search and parabolic interpolation algorithms as described by Brent (1973). The subroutine ZXMIN in IMSL could be used instead.

(2) BIVNOR - A subroutine written by Donnelly (1973) to evaluate the bivariate normal upper right area defined by X, Y, and the correlation between X and Y.

(3) GAUSS - A function which calculates the lower tail area of the standard normal distribution using the central error function in IMSL.

(4) ZVAL - A subroutine which calculates the standard normal deviate corresponding to a given area using successive evaluations of GAUSS.

Free computer time for this research was provided by the ORC Statistical Center headed by Dr. H. B. Wells and the Center for Prevention Research and Biometry headed by Dr. Curt Furberg. Since computer time was free, I did not spend much time optimizing the code and the program can be made more efficient in several places. Single precision rather than double precision calculations can be used throughout if the reduced accuracy is tolerable. The bisection bracketing algorithm used in ZVAL and in obtaining $\alpha \pm \varepsilon$ and $\beta \pm \varepsilon$ can be replaced by a more efficient Newton-Raphson algorithm. The two one-dimensional minimization subroutines could also be replaced by a single two-dimensional minimization subroutine. These changes would probably cut the execution time in half.

```
      IMPLICIT REAL*8(A-Z)
      COMMON ZALPHA,ZPOWER,ALPHA,POWER,CALPHA,CPOWER,P,
     *  EARLYS1,ESS1,EARLYS2,ESS2,ESS3,EARLYS4,ESS4,
     *  U,C1,C2,CHOICE,A2,B2,A3,B3
C
C     VARIABLE  CREATED     DESCRIPTION
C     --------  -------     -----------
C
C     ALPHA     INPUT       DESIRED SIGNIFICANCE LEVEL OF TRIAL
C     POWER     INPUT       DESIRED POWER OF TRIAL
C     ZALPHA    DERIVED     NORMAL DEVIATE CORRESPONDING TO ALPHA
C     ZPOWER    DERIVED     NORMAL DEVIATE CORRESPONDING TO POWER
C     CALPHA    DERIVED     ACTUAL SIGNIFICANCE LEVEL OF TRIAL
C     CPOWER    DERIVED     ACTUAL POWER OF TRIAL
C     P         DERIVED     PROPORTION OF SAMPLE ACCRUED IN 1ST STAGE
C     EARLYS1   DERIVED     PROB OF STOPPING AT 1ST STAGE UNDER Ho
C     ESS1      DERIVED     EXPECTED SAMPLE SIZE UNDER Ho
C     EARLYS2   DERIVED     PROB OF STOPPING AT 1ST STAGE UNDER Ha
C     ESS2      DERIVED     EXPECTED SAMPLE SIZE UNDER Ha
C     ESS3      DERIVED     AVG OF ESS1 AND ESS2
C     EARLYS4   DERIVED     PROB OF STOPPING AT 1ST STAGE UNDER Omax
C     ESS4      DERIVED     EXPECTED SAMPLE SIZE UNDER Omax
C     U         DERIVED     SQRT(N)*DELTA
C     A2        DERIVED     LOWER BOUND FOR C1
C     B2        DERIVED     UPPER BOUND FOR C1
C     A3        INPUT       LOWER BOUND FOR P
C     B3        INPUT       UPPER BOUND FOR P
C     CHOICE    INPUT       DENOTES MINIMIZATION CHOICE
C     C1        DERIVED     CUT POINT FOR ACCEPTANCE OF Ho AT 1ST STAGE
C     C2        DERIVED     CUT POINT FOR REJECTION OF Ho AT 1ST STAGE
C
      EXTERNAL B
      EXTERNAL BMIN
      EXTERNAL ZVAL
C
C     BMIN IS AN EXTERNAL ONE-DIMENSIONAL MINIMIZATION SUBROUTINE
C     WHICH MINIMIZES THE FUNCTION B, WHERE B IS THE ESS AS A
C     FUNCTION OF P
C
C     ZVAL IS AN EXTERNAL SUBROUTINE WHICH CALCULATES THE NORMAL
C     DEVIATE CORRESPONDING TO A GIVEN PROBABILITY
C
C     INPUT FORMAT STATEMENTS
C
   01 FORMAT(////' THIS PROGRAM CALCULATES THE TWO STAGE PARAMETERS')
   02 FORMAT(' FOR 1-SIDED HYPOTHESES ABOUT NORMAL MEANS WHICH')
   03 FORMAT(' MINIMIZE EITHER (1) EXPECTED SAMPLE SIZE UNDER Ho')
   04 FORMAT('                 (2) EXPECTED SAMPLE SIZE UNDER Ha')
   05 FORMAT('                 (3) AVERAGE OF (1) AND (2)')
   06 FORMAT('                 (4) MAXIMUM EXPECTED SAMPLE SIZE')
   07 FORMAT(' UNDER THE RESTRICTION (C3 = GAUSS(1-ALPHA))'//)
   08 FORMAT('$PLEASE ENTER YOUR MINIMIZATION CHOICE (1-4) ')
   09 FORMAT(//'$PLEASE ENTER ALPHA AND  POWER DESIRED  ')
   10 FORMAT(//'$PLEASE ENTER THE BOUNDS FOR P ')
      WRITE(*,01)
```

```
      WRITE(*,02)
      WRITE(*,03)
      WRITE(*,04)
      WRITE(*,05)
      WRITE(*,06)
      WRITE(*,07)
      WRITE(*,08)
      READ(*,*)CHOICE
      WRITE(*,09)
      READ(*,*)ALPHA,POWER
      ZALPHA = ZVAL(1.0D0-ALPHA)
      ZPOWER = ZVAL(POWER)
      WRITE(*,10)
      READ(*,*)A3,B3
      TOL = 0.0000001D0
C
C     TOL IS THE DESIRED LENGTH OF THE UNCERTAINTY INTERVAL OF B
C
      Z1 = BMIN(A3,B3,B,TOL)
C
C     OUTPUT FORMAT STATEMENTS
C
   12 FORMAT(/' ALPHA = ',F9.6,' POWER = ',F9.6,' P = ',F9.6)
   14 FORMAT(/' C3 = ',F9.6,' C1 = ',F9.6,' C2 = ',F9.6)
   16 FORMAT(' NMAX*DELTA^2 = ',F12.6)
   18 FORMAT(' P(STOP EARLY | Ho) =',F8.6,'   ESS*DELTA^2 | Ho =',F12.6)
   20 FORMAT(' P(STOP EARLY | Ha) =',F8.6,'   ESS*DELTA^2 | Ha =',F12.6)
   22 FORMAT('                               ESS*DELTA^2 | Havg =',F12.6)
   24 FORMAT(' P(STOP EARLY | Hm) =',F8.6,'   ESS*DELTA^2 | Hm =',F12.6)
   26 FORMAT(' FIXED N*DELTA^2 = ',F10.6)
   28 FORMAT(' MAX N / FIXED N = ',F12.6)
   30 FORMAT(' ESSo / FIXED N = ',F12.6)
   32 FORMAT(' ESSa / FIXED N = ',F12.6)
   34 FORMAT(' ESSavg / FIXED N = ',F12.6)
   36 FORMAT(' ESSm / FIXED N = ',F12.6)
   38 FORMAT(//' MINIMIZATION UNDER Ho WAS CHOSEN')
   40 FORMAT(//' MINIMIZATION UNDER Ha WAS CHOSEN')
   42 FORMAT(//' MINIMIZATION OF (ESSo + ESSa)/2 WAS CHOSEN')
   44 FORMAT(//' MINIMIZATION OF THE MAXIMUM ESS WAS CHOSEN')
C
      USQ = U**2
      FIXSS = (ZALPHA + ZPOWER)**2
      PINC = USQ/FIXSS
      P1DEC = ESS1/FIXSS
      P2DEC = ESS2/FIXSS
      P3DEC = ESS3/FIXSS
      P4DEC = ESS4/FIXSS
C
C     USQ IS N*(DELTA SQUARED)
C     FIXSS IS THE FIXED SAMPLE SIZE
C     PINC IS N/Nfix
C     P1DEC IS ESSo/Nfix
C     P2DEC IS ESSa/Nfix
C     P3DEC IS ESSavg/Nfix
C     P4DEC IS ESSm/Nfix
```

```
C
      IF (CHOICE .EQ. 1.0D0) WRITE(89,38)
      IF (CHOICE .EQ. 2.0D0) WRITE(89,40)
      IF (CHOICE .EQ. 3.0D0) WRITE(89,42)
      IF (CHOICE .EQ. 4.0D0) WRITE(89,44)
C
C     OUTPUT RESULTS TO FILE FOR089.DAT
C
      WRITE(89,12)CALPHA,POWER,P
      WRITE(89,14)ZALPHA,C1,C2
      WRITE(89,16)USQ
      WRITE(89,18)EARLYS1,ESS1
      WRITE(89,20)EARLYS2,ESS2
      WRITE(89,22)ESS3
      WRITE(89,24)EARLYS4,ESS4
      WRITE(89,26)FIXSS
      WRITE(89,28)PINC
      WRITE(89,30)P1DEC
      WRITE(89,32)P2DEC
      WRITE(89,34)P3DEC
      WRITE(89,36)P4DEC
      STOP
      END

      REAL*8 FUNCTION B(Z2)
      IMPLICIT REAL*8(A-Z)
      COMMON ZALPHA,ZPOWER,ALPHA,POWER,CALPHA,CPOWER,P,
     *  EARLYS1,ESS1,EARLYS2,ESS2,ESS3,EARLYS4,ESS4,
     *  U,C1,C2,CHOICE,A2,B2,A3,B3
C
      EXTERNAL F
      EXTERNAL FMIN
      EXTERNAL BIVNOR
C
C     FMIN IS A ONE-DIMENSIONAL MINIMIZATION SUBROUTINE WHICH MINIMIZES
C     THE FUNCTION F, WHERE F IS THE ESS AS A FUNCTION OF C1
C
C     BIVNOR IS AN EXTERNAL SUBROUTINE WHICH CALCULATES THE UPPER RIGHT
C     AREA DEFINED BY X, Y, AND THE CORRELATION BETWEEN X AND Y.
C
      P = Z2
C
C     P IS THE PARAMETER BEING ITERATED OVER
C
C     DETERMINE LOWER LIMIT FOR C1; UPPER LIMIT IS ZALPHA
C
      TOL = 0.0000001D0
      B2 = ZALPHA
      LOWER = -6.0D0
      UPPER = 6.0D0
      DO 120 K=1,50
       A2 = (LOWER + UPPER)/2.0D0
       AREA = BIVNOR(A2,ZALPHA,DSQRT(P))
       IF (DABS(AREA - ALPHA) .LE. TOL) GO TO 130
       IF (AREA .GT. ALPHA) THEN
```

```
        LOWER = A2
        ELSE
        UPPER = A2
        END IF
  120  CONTINUE
       IF (AREA .LT. ALPHA) THEN
        A1 = -6.0D0
        GO TO 130
        END IF
       WRITE(89,*)ZALPHA
       WRITE(89,*)P
       WRITE(89,*)A2
       WRITE(89,*)AREA
       WRITE(89,2)
       STOP
    1 FORMAT(' LOWER LIMIT OF C1 GREATER THAN UPPER LIMIT')
    2 FORMAT(' NO C1 WORKS FOR THIS C3, P, AND ALPHA')
    3 FORMAT(' LOWER LIMIT FOR C1 SET TO -6')
  130 IF (A2 .GT. B2) THEN
        WRITE(89,*)
        WRITE(89,*)A2
        WRITE(89,*)B2
        WRITE(89,1)
        STOP
        END IF

       Z = FMIN(A2,B2,F,TOL)

       IF (CHOICE .EQ. 1.0D0) B = ESS1
       IF (CHOICE .EQ. 2.0D0) B = ESS2
       IF (CHOICE .EQ. 3.0D0) B = ESS3
       IF (CHOICE .EQ. 4.0D0) B = ESS4

       RETURN
       END



       REAL*8 FUNCTION F(Z1)
       IMPLICIT REAL*8(A-Z)
       COMMON ZALPHA,ZPOWER,ALPHA,POWER,CALPHA,CPOWER,P,
     *  EARLYS1,ESS1,EARLYS2,ESS2,ESS3,EARLYS4,ESS4,
     *  U,C1,C2,CHOICE,A2,B2,A3,B3

       EXTERNAL BIVNOR
       EXTERNAL DERF
       GAUSS(T) = (1.0D0+DERF(T/DSQRT(2.0D0)))/2.0D0
C
C      BIVNOR IS AN EXTERNAL SUBROUTINE WHICH CALCULATES THE UPPER RIGHT
C      AREA DEFINED BY X, Y, AND THE CORRELATION BETWEEN X AND Y.
C
C      GAUSS IS A FUNCTION WHICH CALCULATES THE LOWER TAIL AREA
C      CORRESPONDING TO THE NORMAL DEVIATE T USING THE CENTRAL ERROR
C      FUNCTION DERF AVAILABLE IN IMSL
C
```

```
      C1 = Z1
C
C     C1 IS THE PARAMETER BEING ITERATED OVER
C
      PSQRT = DSQRT(P)
      PTOL = 0.0000001D0
      ATOL = 0.0000001D0
C
C     PTOL IS THE TOLERANCE ON THE ACCURACY OF THE POWER
C     ATOL IS THE TOLERANCE ON THE ACCURACY OF ALPHA
C
C     FIND C2, GIVEN C1, P, AND C3, SUCH THAT THE SIGNIFICANCE
C     LEVEL FOR THE TRIAL IS ALPHA PLUS OR MINUS ATOL
C
       UPPER = 6.0D0
       LOWER = C1
       DO 100 I=1,50
       C2 = (LOWER + UPPER)/2.0D0
       CALPHA = 1.0D0 - GAUSS(C2)
     *        + BIVNOR(C1,ZALPHA,PSQRT)
     *        - BIVNOR(C2,ZALPHA,PSQRT)
       IF (DABS(CALPHA - ALPHA) .LE. ATOL) GO TO 110
        IF (CALPHA .GT. ALPHA) THEN
         LOWER = C2
         ELSE
         UPPER = C2
         END IF
  100   CONTINUE
      WRITE(89,*)ZALPHA
      WRITE(89,*)C1
      WRITE(89,*)C2
      WRITE(89,*)P
      WRITE(89,*)CALPHA
      STOP
C
C      FIND THE N*(DELTA SQUARED) NECESSARY TO HAVE A POWER
C      FOR THE TRIAL OF POWER PLUS OR MINUS PTOL
C
  110 LOWER = 0.0D0
      UPPER = 10.0D0
C
C  U = SQRT(N)(MU - MU0)/SIGMA
C
      DO 120 K=1,50
       U = (LOWER + UPPER)/2.0D0
       U2 = C2 - U*PSQRT
       CPOWER = 1.0D0 - GAUSS(U2)
     *        + BIVNOR(C1-U*PSQRT,ZALPHA-U,PSQRT)
     *        - BIVNOR(C2-U*PSQRT,ZALPHA-U,PSQRT)
       IF (DABS(POWER - CPOWER) .LE. PTOL) GO TO 130
       IF (CPOWER .LT. POWER) THEN
        LOWER = U
        ELSE
        UPPER = U
        END IF
```

```
  120  CONTINUE
       WRITE(89,*)ZALPHA
       WRITE(89,*)C1
       WRITE(89,*)C2
       WRITE(89,*)P
       WRITE(89,*)CALPHA
       WRITE(89,*)POWER
       WRITE(89,*)U
       STOP
C
C      CALCULATE THE PROB OF STOPPING EARLY AND THE ESS UNDER Ho
C
  130  Q3 = GAUSS(C1)
       Q4 = 1.0D0 - GAUSS(C2)
       EARLYS1 = Q3 + Q4
       ESS1 = (1.0D0 -  (1.0D0 - P)*EARLYS1)*U**2.0D0
C
C      CALCULATE THE PROB OF STOPPING EARLY AND THE ESS UNDER Ha
C
       U2 = C1 - PSQRT*U
       U3 = C2 - PSQRT*U
       Q1 = GAUSS(U2)
       Q2 = 1.0D0 - GAUSS(U3)
       EARLYS2 = Q1 + Q2
       ESS2 = (1.0D0 - (1.0D0 - P)*EARLYS2)*U**2.0D0
C
C      CALCULATE (ESSo + ESSa)/2
C
       ESS3 = (ESS1 + ESS2)/2.0D0
C
C      CALCULATE PROB OF STOPPING EARLY AND THE EXPECTED
C      SAMPLE SIZE IF THETA EQUALS (THETA)max
C
       U5 = C1 - (C1+C2)/2.0D0
       U6 = C2 - (C1+C2)/2.0D0
       Q5 = GAUSS(U5)
       Q6 = 1.0D0 - GAUSS(U6)
       EARLYS4 = Q5 + Q6
       ESS4 = (1.0D0 - (1.0D0 - P)*EARLYS4)*U**2.0D0


       IF (CHOICE .EQ. 1.0D0) F = ESS1
       IF (CHOICE .EQ. 2.0D0) F = ESS2
       IF (CHOICE .EQ. 3.0D0) F = ESS3
       IF (CHOICE .EQ. 4.0D0) F = ESS4

       RETURN
       END
```

REFERENCES

Alling D.W. (1963). Early decision in the Wilcoxon two-sample test. Journal of the American Statistical Association, 58, 713-720.

Anscombe, F.J. (1954). Fixed-sample-size analysis of sequential observations. Biometrics, 10, 89-100.

Anscombe, F.J. (1963). Sequential medical trials. Journal of the American Statistical Association, 58, 365-383.

Armitage, P. (1975). Sequential Medical Trials, 2nd edition, Blackwell, Oxford.

Armitage, P. (1978). Sequential medical trials. Biomedicine Special Issue, 28, 40-41.

Armitage, P., McPherson, C.K., and Rowe, B.C. (1969). Repeated significance tests on accumulating data. Journal of the Royal Statistical Society, Series A, 132, 235-244.

Armitage, P., Stratton, I.M., and Worthington, H.V. (1985). Repeated significance tests for clinical trials with a fixed number of patients and variable follow-up. Biometrics, 41, 353-359.

Bangdiwala, S.I. and Sen, P.K. (1985). Statistical monitoring of clinical trials with staggered entry. In Sen, P.K., ed., Biostatistics: Statistics in Public Health and Environmental Sciences, North-Holland, New York, 13-30.

Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). Journal of the American Statistical Association, 57, 269-326.

Brent, R. (1973). Algorithms for minimization without derivatives. Prentice-Hall, Inc.

Canner, P.L. (1977). Monitoring treatment differences in long-term clinical trials. Biometrics, 33, 603-615.

Chatterjee, S.K. and Sen, P.K. (1973). Nonparametric testing under progressive censoring, Calcutta Statistical Association Bulletin, 22, 13-50.

Colton, T. (1963). Optimal drug screening plans. Biometrika, 50, 31-45.

Colton, T. and McPherson, K. (1976). Two-stage plans compared with fixed-sample-size and Wald SPRT plans. Journal of the American Statistical Association, 71, 80-86.

Cornfield, J. (1966a). Sequential trials, sequential analysis and the likelihood principle. The American Statistician, 20, 18-23.

Cornfield, J. (1966b). A Bayesian test of some classical hypotheses with applications to sequential clinical trials. Journal of the American Statistical Association, 61, 577-594.

Cornfield, J. (1976). Recent methodological contributions to clinical trials. American Journal of Epidemiology, 104, 408-421.

Cox, D.R. (1972). Regression models and life tables (with discussion). Journal of the Royal Statistical Society, Series B, 34, 187-220.

DeMets, D.L. (1984). Stopping guidelines vs stopping rules: a practitioner's point of view. Communications in Statistics - Theory and Methods, 13(19), 2395-2417.

DeMets, D.L. and Lan, K.K.G. (1984). An overview of sequential methods and their application in clinical trials. Communications in Statistics - Theory and Methods, 13(19), 2315-2338.

DeMets, D.L. and Ware, J.H. (1980). Group sequential methods for clinical trials with a one-sided hypothesis. Biometrika, 67, 651-660.

DeMets, D.L. and Ware, J.H. (1982). Asymmetric group sequential boundaries for monitoring clinical trials. Biometrika, 69, 661-663.

De With, C. (1983). Two-stage plans for the testing of binomial parameters. Controlled Clinical Trials, 4, 215-226.

Dodge, H.F. and Romig, H.G. (1929). A method of sampling inspection. The Bell System Technical Journal, 8, 613-631.

Dodge, H.F. and Romig, H.G. (1941). Single sampling and double sampling inspection tables. The Bell System Technical Journal, 20, 1-61.

Donnely, T.G. (1973). Algorithm 462 - Bivariate normal distribution [S15]. Communications of the Association of Computing Machines, 16, 638.

Dupont, W.D. (1983). Sequential stopping rules and sequentially adjusted p values: does one require the other? (with discussion). Controlled Clinical Trials, 4, 3-35.

Edwards, A.W.F. (1972). Likelihood, Cambridge University Press, London.

Elashoff, J.D. and Reedy, T.J. (1984). Two-stage clinical trial stopping rules. Biometrics, 40, 791-795.

Elashoff, R.M. and Beal, S. (1976). Two-stage screening designs applied to chemical-screening problems with binary data. Annual Review of Biophysics and Bioengineering, 5, 561-587.

Elder, R.S. (1974). Double sampling for lot average. Technometrics, 16, 435-439.

Elfring, G.L. and Schultz, J.R. (1973). Group sequential designs for clinical trials. Biometrics, 29, 471-477.

Fairbanks, K. and Madsen, R. (1982). P values for tests using a repeated significance test design. Biometrika, 69, 69-74.

Feller, W. (1940). Statistical aspects of ESP. Journal of Parapsychology, 4, 271-298.

Fleming, T.R. (1982). One-sample multiple testing procedure for phase II clinical trials. Biometrics, 38, 143-151.

Fleming, T.R., Harrington, D.P., and O'Brien, P.C. (1984). Designs for group sequential tests. Controlled Clinical Trials, 5, 348-361.

Freedman, L.S., Lowe, D., and Macaskill, P. (1984). Stopping rules for clinical trials incorporating clinical opinion. Biometrics, 40, 575-586.

Friedman, L.M., Furberg, C.D., and DeMets, D.L. (1981), Fundamentals of Clinical Trials, Wright, Boston.

Gail, M. (1982). Monitoring and stopping clinical trials. In Mike, V. and Stanley, K., eds., Statistics in Medical Research: Methods and Issues, With Applications in Cancer Research, Wiley, New York, 455-484.

Gail, M.H., DeMets, D.L., and Slud, E.V. (1982). Simulation studies on increments of the two-sample log rank score test for survival data, with application to group sequential boundaries. In Crowley, J. and Johnson, R.A., eds., Survival Analysis, IMS Monograph Series, Hayward, California, 287-301.

Glaz, J. and McK. Johnson, B. (1986). Approximating boundary crossing probabilities with applications to sequential tests. Sequential Analysis, 5(1), 37-72.

Gould, A.L. and Pecore, V.J. (1982). Group sequential methods for clinical trials allowing early acceptance of Ho and incorporating costs. Biometrika, 69, 75-80.

Hald, A. (1975). Optimum double sampling tests of given strength. Journal of the American Statistical Association, 70, 451-456.

Halperin, M., Lan, K.K.G., Ware, J.H., Johnson, N.J., and DeMets, D.L. (1982). An aid to data monitoring in long-term clinical trials. Controlled Clinical Trials, 3, 311-323.

Hamaker, H.C. and Van Strik, R. (1955). The efficiency of double sampling for attributes. Journal of the American Statistical Association, 50, 830-849.

Haybittle, J.L. (1971). Repeated assessment of results in clinical trials of cancer treatment. British Joural of Radiology, 44, 793-797.

Jackson, D.V., Zekan, P.J., Caldwell, R.D., Slatkoff, M.L., Harding, R.W., Case, L.D., Hopkins, J., Muss, H.B., Richards, F. II, White, D.R., Cooper, M.R., Stuart, J.J., Capizzi, R.L., Spurr, C.L. (1984). VP-16-214 in combination chemotherapy with chest irradiation for small-cell lung cancer: A randomized trial of the Piedmont Oncology Association. Journal of Clinical Oncology, 2, 1343-1351.

Johnson, L.W. and Riess, R.D. (1977). Numerical Analysis, Addison-Wesley, Reading, Massachusetts.

Jones, D. and Whitehead, J. (1979a). Sequential forms of the log rank and modified Wilcoxon tests for censored data. Biometrika, 66, 105-113.

Jones, D.R. and Whitehead, J. (1979b). Group sequential methods. British Journal of Cancer, 40, 171.

Lan, K.K.G. and DeMets, D.L. (1983). Discrete sequential boundaries for clinical trials. Biometrika, 70, 659-663.

Lan, K.K.G., DeMets, D.L., and Halperin, M. (1984). More flexible sequential and non-sequential designs in long-term clinical trials. Communications in Statistics - Theory and Methods, 13(19), 2339-2353.

Lan, K.K.G., Simon, R., and Halperin, M. (1982). Early stopping in the two-sample problem for bounded random variables. Communications in Statistics, Sequential Analysis, 1, 207-219.

Majumdar, H. and Sen, P.K. (1978). Nonparametric testing for single regression under progressive censoring with staggering entry and random withdrawal. Communications in Statistics - Theory and Methods, 7, 349-371.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemotherapy Reports, 50, 163-170.

McPherson, K. (1974). Statistics: the problem of examining accumulating data more than once. The New England Journal of Medicine, 290, 501-502.

McPherson, K. (1982). On choosing the number of interim analyses in clinical trials. Statistics in Medicine, 1, 25-36.

McPherson, C.K. and Armitage, P. (1971). Repeated significance tests on accumulating data when the null hypothesis is not true. Journal of the Royal Statistical Society, Series A, 134, 15-25.

Muss, H.B., Paschold, E.H., Black, W.R., Cooper, M.R., Capizzi, R.L., Christian, R., Cruz, J.M., Jackson, D.V., Stuart, J.J., Richards, F. II, White, D.R., Zekan, P.J., Spurr, C.L., Pope, E., Case, D., Morgan, T., Wells, H.B. (1985). Megestrol acetate vs. tamoxifen in advanced breast cancer: A phase III trial of the Piedmont Oncology Association (POA). Seminars in Oncology, 12, 55-61.

O'Brien, P.C. and Fleming, T.R. (1979). A multiple testing procedure for clinical trials. Biometrics, 35, 549-556.

Owen, D.B. (1953). A double sample procedure. Annals of Mathematical Statistics, 24, 449-457.

Peace, K.E. and Schriver, R.C. (1987). P-value and power computations in multiple look trials. Journal of Chronic Diseases, 40, 23-30.

Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S.V., Mantel, N., McPherson, K., Peto, J., and Smith, P.G. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. British Journal of

Cancer, 34, 585-612.

Pocock, S.J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64, 191-199.

Pocock, S.J. (1979). Group sequential methods. *British Journal of Cancer*, 40, 171-172.

Pocock, S.J. (1980). Interim analyses and stopping rules for clinical trials. In Bithell, J.F. and Coppi, R., eds., *Perspectives in Medical Statistics*, Academic Press, London, 191-223.

Pocock, S.J. (1982). Interim analyses for randomized clinical trials: the group sequential approach. *Biometrics*, 38, 153-162.

Pocock, S.J. (1983). *Clinical Trials - A Practical Approach*, Wiley, New York.

Richards, F. II, Case, L.D., White, D.R., Muss, H.B., Spurr, C.L., Jackson, D.V., Cooper, M.R., Zekan, P., Cruz, J.M., Stuart, J.J., Capizzi, R.L., McCulloch, J.H., McFarland, J.A., Kincaid, W.R., Harding, R.W., Pope, E., McMahan, R., and Wells, H.B. (1986). Combination chemotherapy (5-Fluorouricil, methyl-CCNU, and mitomycin C) vs. 5-Fluorouricil alone for advanced previously untreated colorectal carcinoma. A phase III study of the Piedmont Oncology Association. *Journal of Clinical Oncology*, 4, 565-570.

Richards, F., Case, D., Jackson, D., Muss, H., Cooper, M.R., White, D., Stuart, J., Hopkins, J., Slatkoff, M., Brodkin, R., Sartiano, G., McCulloch, J., Bearden, J., Harding, R., Fenning, B., McMahan, R., Wells, H.B., Spurr, C.L., and Capizzi, R. (1984). Prognostic factors as determinants of response and survival in advanced non-small cell carcinoma of the lung (NSCLC) treated with CMAP (CCNU, methotrexate, adriamycin, cisplatinum). *Proceedings of ASCO*, 3, 230.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58, 527-535.

Roseberry, T.D. and Gehan, E.A. (1964). Operating characteristic curves and accept-reject rules for two and three stage screening procedures. *Biometrics*, 20, 73-84.

Rubinstein, L.V. and Gail, M.H. (1982). Monitoring rules for stopping accrual in comparative survival studies. *Controlled Clinical Trials*, 3, 325-343.

SAS Institute Inc. (1985). *SAS/GRAPH User's Guide - G3GRID Procedure*, Cary, NC.

Schervish, M.J. (1984). Algorithm AS-195, Multivariate normal probabilities with error bound. Applied Statistics, 33, 81-94.

Sellke, T. and Siegmund, D. (1983). Sequential analysis of the proportional hazards model. Biometrika, 70, 315-326.

Sen, P.K. (1976). A two dimensional functional permutational central limit theorem for linear rank statistics. Annals of Probability, 4, 13-26.

Spurrier, J.D. (1978). Two-stage sign tests that allow first stage acceptance and rejection. Communications in Statistics - Theory and Methods, 7(4), 399-408.

Spurrier, J.D. and Hewett, J.E. (1975). Double sample tests for the mean of a normal population. Journal of the American Statistical Association, 70, 448-450.

Spurrier, J.D. and Hewett, J.E. (1976). Two-stage Wilcoxon tests of hypotheses. Journal of the American Statistical Association, 71, 982-987.

Tsiatis, A.A. (1981). The asymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time. Biometrika, 68, 311-315.

Tsiatis, A.A. (1982). Repeated significance testing for a general class of statistics used in censored survival analysis. Journal of the American Statistical Association, 77, 855-861.

Tsiatis, A.A., Rosner, G.L., and Tritchler, D.L. (1985). Group sequential tests with censored survival data adjusting for covariates. Biometrika, 72, 365-373.

Wald, A. (1947), Sequential Analysis, Wiley, New York.

Wald, A. and Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. Annals of Mathematical Statistics, 19, 326-339.

Ware, J.H., Muller, J.E., and Braundwald, E. (1985). The futility index - an approach to the cost-effective termination of randomized clinical trials. The American Journal of Medicine, 78, 635-643.

Whitehead, J. (1983). The Design and Analysis of Sequential Clinical Trials, Ellis Horwood Limited, Chichester.