UC Riverside

UC Riverside Electronic Theses and Dissertations

Title

Designing One-Sided Group Sequential Clinical Trials to Detect a Mixture Alternative

Permalink

<https://escholarship.org/uc/item/7qx0d0xr>

Author

Peng, Hua

Publication Date

2017

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
RIVERSIDE

Designing One-Sided Group Sequential Clinical Trials to Detect a Mixture Alternative

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Applied Statistics

by

Hua Peng

December 2017

Dissertation Committee:

Dr. Daniel R. Jeske, Chairperson

Dr. Weixin Yao

Dr. Gloria Gonzalez-Rivera

Copyright by
Hua Peng
2017

The Dissertation of Hua Peng is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

My most sincere thanks to my advisor, Dr. Daniel R. Jeske, for his enthusiasm, dedication and guidance. Without his help and support, it would not be possible for me to complete the doctoral dissertation.

I would also like to thank my co-advisor Dr. Ashis SenGupta for his valuable suggestions for L-optimal similar test in the paper. I would also like to thank my committee members Dr. Weixin Yao and Dr. Gloria Gonzalez-Rivera. I am grateful to Prof. Jun Li, Prof. Subir Ghosh, Prof. James Flegal, Prof. Barry C. Arnold, Prof. Shujie Ma, and Prof. Keh-Shin Lii for their instruction during my UCR study. My appreciation goes to Jill Smith and Linda Penas for the wonderful teaching experience. I would like to thank all my friends at UCR for the wonderful life. Especially, I wish to thank my best friends Chen Lin and Lijie Li for their friendship.

Finally I would like to thank my husband Chenyin Xiao, my parents and parents in law for their unconditional love and support.

To my beloved husband and parents.

ABSTRACT OF THE DISSERTATION

Designing One-Sided Group Sequential Clinical Trials to Detect a Mixture Alternative

by

Hua Peng

Doctor of Philosophy, Graduate Program in Applied Statistics

University of California, Riverside, December 2017

Dr. Daniel R. Jeske, Chairperson

Group sequential designs play an important role in monitoring clinical trials. In this dissertation, we consider the construction of one-sided group sequential designs where the stopping rule includes boundaries for early stopping to accept for futility and to reject for efficacy. The traditional assumption that all patients have the same likelihood of benefiting from the treatment is sometimes unrealistic and it can underestimate the required sample size. This motivates us to power the design for an alternative where the treatment group observations come from a mixture of normal distributions. For the proposed setting, we use standardized test statistics based on sample means. Stopping boundaries and arm size for the design are determined by Type I and Type II error spending equations. We demonstrate the need for larger arm sizes when trying to detect a mixture alternative compared to trying to detect a pure shift alternative. The unknown variance case is discussed, and we introduce the group sequential t-test for the mixture setting. With the

mixture model, we discuss a more general definition of treatment effect. The maximum likelihood estimator and method of moment estimator for the treatment effect are discussed.

Contents

List of Figures	x
List of Tables.....	xii
Chapter 1 Introduction.....	1
1. 1 Group Sequential Methods Overview	1
1. 2 Motivation for the Mixture Alternative	2
1. 3 Group Sequential Designs to Detect a Mixture Alternative	4
1. 4 Dissertation Structure	5
Chapter 2 Preliminaries on Group Sequential Methods	6
2. 1 A Fixed Sample Test	6
2. 2 Group Sequential Setting and Notation	8
2. 3 Group Sequential Tests Fixed Boundary Shape Methods	13
2. 3. 1 Pocock	14
2. 3. 2 O’Brien and Fleming	16
2. 3. 3 Wang and Tsatis	19
2. 4 Group Sequential Tests Error Spending Methods.....	21
2. 4. 1 Efficacy Stopping Rules	21
2. 4. 2 Efficacy and Futility Stopping Rules	25
2. 5 Estimation Following Group Sequential Tests	27
2. 5. 1 Point Estimation	27
2. 5. 2 Bias Reduction Methods for Point Estimates	27
2. 5. 2. 1 Whitehead Method	27
2. 5. 2. 2 Bootstrap Bias Reduction Method	28

2. 5. 2. 3	Numerical Results	29
2. 5. 3	Comparison of MLE and Bias-Corrected MLE	34
Chapter 3	One-Sided Group Sequential Designs to Detect a Mixture Alternative	40
3. 1	A Fixed Sample Test	40
3. 2	Group Sequential Tests	42
3. 3	Properties of the Proposed Group Sequential Designs	46
3. 4	Examples Illustration	48
Chapter 4	Implementation Details and Practical Concerns	50
4. 1	Unknown σ^2	50
4. 2	Treatment Effect Estimation	53
4. 2. 1	MLE and Bias-Corrected MLE	53
4. 2. 2	MOM and Bias-Corrected MOM	60
Chapter 5	Conclusions	69
5. 1	Conclusion	69
5. 2	Future Work	70
Bibliography		72
Appendix		75
Appendix A	Sub-density	75
Appendix B	Ratio and Stopping Boundaries	77
Appendix C	The R Code for the Stopping Boundaries	83

List of Figures

Figure 2.1: General Form for Stopping Boundaries	9
Figure 2.2: Two-sided Pocock Test with Efficacy Boundaries for $K=5$	14
Figure 2.3: Two-sided O'Brien and Fleming Test with Efficacy Boundaries for $K=5$	17
Figure 2.4: Type I and Type II error spending functions with $\alpha = 0.05, \beta = 0.2$	23
Figure 2.5: Bias of MLE and bias-corrected MLE for Two-Stage One-Sided Group Sequential Design: Example 1	31
Figure 2.6: Bias of MLE and bias-corrected MLE for Two-Stage One-Sided Group Sequential Design: Example 2	32
Figure 2.7: Bias of MLE and bias-corrected MLE for Two-Stage One-Sided Group Sequential Design: Example 3	33
Figure 2.8: Bias, variance and MSE of MLE and bias-corrected MLE of (θ, δ) with $\delta =$ 0.3	35
Figure 2.9: Bias, variance and MSE of MLE and bias-corrected MLE of (θ, δ) with $\delta =$ 0.5	36
Figure 2.10: Bias, variance and MSE of MLE and bias-corrected MLE of (θ, δ) with $\delta =$ 0.7	37
Figure 2.11: Bias, variance and MSE of MLE and bias-corrected MLE of (θ, δ) with $\delta =$ 0.9	38
Figure 2.12: Bias, variance and MSE of MLE and bias-corrected MLE of (θ, δ) with $\delta =$ 1.1	39
Figure 3.1: General picture for one-sided group sequential design stopping boundaries with $K=5$	44
Figure 3.2: Stopping boundaries for $K = 3, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \delta = 0.5, \theta =$ $0.7, \rho = 2$	49
Figure 4.1: Bias, variance and MSE of MLE and bootstrap bias-corrected MLE of (θ, δ) . $K = 3, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 2, \mu_0 = 0$, specified values of $(\theta, \delta) = (0.3, 0.8)$, $B = 1,000$, and 1,000 sample paths.	55

Figure 4.2: Bias, variance and MSE of MLE and bootstrap bias-corrected MLE of (θ, δ) . $K = 3, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 2, \mu_0 = 0$, specified values of $(\theta, \delta) = (0.5, 0.8)$, $B = 1,000$, and 1,000 sample paths.	56
Figure 4.3: Bias, variance and MSE of MLE and bootstrap bias-corrected MLE of (θ, δ) . $K = 3, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 2, \mu_0 = 0$, specified values of $(\theta, \delta) = (0.8, 0.8)$, $B = 1,000$, and 1,000 sample paths.	57
Figure 4.4: Bias, variance and MSE of MLE and bootstrap bias-corrected MLE of (θ, δ) . $K = 3, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 2, \mu_0 = 0$, specified values of $(\theta, \delta) = (0.3, 1)$, $B = 1,000$, and 1,000 sample paths.	58
Figure 4.5: Bias, variance and MSE of MLE and bootstrap bias-corrected MLE of (θ, δ) . $K = 3, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 2, \mu_0 = 0$, specified values of $(\theta, \delta) = (0.5, 1)$, $B = 1,000$, and 1,000 sample paths.	59
Figure 4.6: Bias, variance and MSE of MLE and bootstrap bias-corrected MLE of (θ, δ) . $K = 3, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 2, \mu_0 = 0$, specified values of $(\theta, \delta) = (0.8, 1)$, $B = 1,000$, and 1,000 sample paths.	60
Figure 4.7: Bias, variance and MSE of MOM and bootstrap bias-corrected MOM of (θ, δ) . $K = 3, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 2, \mu_0 = 0$, specified values of $(\theta, \delta) =$ $(0.3, 0.8)$, $B = 1,000$, and 1,000 sample paths.	62
Figure 4.8: Bias, variance and MSE of MOM and bootstrap bias-corrected MOM of (θ, δ) . $K = 3, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 2, \mu_0 = 0$, specified values of $(\theta, \delta) =$ $(0.5, 0.8)$, $B = 1,000$, and 1,000 sample paths.	63
Figure 4.9: Bias, variance and MSE of MOM and bootstrap bias-corrected MOM of (θ, δ) . $K = 3, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 2, \mu_0 = 0$, specified values of $(\theta, \delta) =$ $(0.8, 0.8)$, $B = 1,000$, and 1,000 sample paths.	64
Figure 4.10: Bias, variance and MSE of MOM and bootstrap bias-corrected MOM of (θ, δ) . $K = 3, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 2, \mu_0 = 0$, specified values of $(\theta, \delta) =$ $(0.3, 1)$, $B = 1,000$, and 1,000 sample paths.	65
Figure 4.11: Bias, variance and MSE of MOM and bootstrap bias-corrected MOM of (θ, δ) . $K = 3, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 2, \mu_0 = 0$, specified values of $(\theta, \delta) =$ $(0.5, 1)$, $B = 1,000$, and 1,000 sample paths.	66
Figure 4.12: Bias, variance and MSE of MOM and bootstrap bias-corrected MOM of (θ, δ) . $K = 3, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 2, \mu_0 = 0$, specified values of $(\theta, \delta) =$ $(0.8, 1)$, $B = 1,000$, and 1,000 sample paths.	67

List of Tables

Table 2.1: Notation in Group Sequential Designs.....	13
Table 2.2: Pocock tests: ratio of the maximum arm size for the group sequential test to the arm size of the fixed sample test.....	16
Table 2.3: O'Brien and Fleming tests: ratio of the maximum arm size for the group sequential test to the arm size of the fixed sample test.....	18
Table 2.4: Wang and Tsatis tests: ratio of the maximum arm size for the group sequential test to the arm size of the fixed sample test.....	20
Table 2.5: Two-sided group sequential tests using Type I error spending functions: ratio of the maximum arm size for the group sequential test to the arm size of the fixed sample test.....	24
Table 2.6: One-sided group sequential tests using Type I and Type II error spending functions: ratio of the maximum arm size for the group sequential test to the arm size of the fixed sample test.....	26
Table 3.1: Arm size for a 5% level fixed sample one-sided test based on Z that will achieve 80% power for the mixture alternative. Numbers in parentheses are arm sizes for a 5% test based Monte Carlo simulation (simulation size 100,000).....	42
Table 3.2: Runtime for $\alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \delta = 0.5, \theta = 0.7, \rho = 2$, and specified $\theta, \delta) = (0.7, 0.5)$	46
Table 3.3: Proposed one-sided group sequential tests: ratio of the maximum arm size for the group sequential test to the arm size of the fixed sample test with Type I error rate 1%, 5%, and 10%.....	47
Table 3.4: Target and achieved Type I and Type II error at each stage for $K = 3, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 2$, and specified $\theta, \delta = (0.7, 0.5)$	49
Table 4.1: Type I and Type II error rate for two variance estimates. $K = 2, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 1, \mu_0 = 0$, simulation size is 10,000.....	51
Table 4.2: Properties of Group Sequential t-Tests: Type I and Type II error rate.....	53
Table 4.3: Bias, variance and MSE comparisons between MLE and MOM.....	68

Table B.1: Ratio of the maximum arm size for the group sequential test to the arm size for the fixed sample test for 5% level and will achieve 80% power for the mixture alternative with $\rho = 1$	80
Table B.2: Ratio of the maximum arm size for the group sequential test to the arm size for the fixed sample test for 5% level and will achieve 80% power for the mixture alternative with $\rho = 2$	81
Table B.3: Ratio of the maximum arm size for the group sequential test to the arm size for the fixed sample test for 5% level and will achieve 80% power for the mixture alternative with $\rho = 3$	82

Chapter 1

Introduction

1.1 Group Sequential Methods Overview

Group sequential designs play an important role in monitoring clinical trials. The reasons to conduct group sequential designs can be categorized in three classes: economic, ethical, and administrative.

When early termination is desired for either a small or a large treatment difference, a group sequential design with both early stopping to reject the null hypothesis for efficacy and early stopping to accept the null hypothesis for futility can be constructed. Numerous techniques for formulating this design have been developed. Popular methods are fixed boundary shape methods and error spending methods. The fixed boundary shape (i.e.

boundaries are parameterized by one or two constants) methods derive boundaries with specified boundary shapes, and these methods include Haybittle-Peto (Peto, et al., 1976), Pocock (1977), O'Brien and Fleming (1979), Wang and Tsiatis (1987). The error spending method uses an error spending function to specify the errors at each stage, and the stopping boundaries are determined by Type I or Type II error equations. Lan and DeMets (1983) propose a group sequential design for early stopping with rejection boundaries, and the design is based on a Type I error spending function. Chang, Hwang, and Shih (1990) extend the method of Lan and DeMets to early stopping with rejection and acceptance boundaries, and propose a group sequential design using both Type I and Type II error spending functions.

Several software applications are available for designing a group sequential design. Zhu, Ni, and Yao (2011) summarize four popular software applications, which are EAST v5.2 (Cytel Inc., 2010), ADDPLAN v5.0, the gsDesign package v2.3 in R (Anderson, 2009), and the SEQDESIGN and SEQTEST procedures in SAS v9.2 (SAS Institute Inc., 2009).

1.2 Motivation for the Mixture Alternative

For the conventional group sequential design discussed above, it is assumed that both the control and treatment group responses come from normal distributions that potentially have different means. We call this setting a pure shift setting in later sections. As discussed in

‘Paving the Way for Personalized Medicine’ by the Food and Drug Administration (FDA, 2013), the actual safety and effectiveness of a treatment or product may vary from one individual to the next. Reasons for different results include genetic and environmental factors, as well as the interaction of these factors. For some drug therapies we can anticipate only a percentage of treated patients will see an effect. A study conducted by Spear, Heath-Chiozzi, and Huff (2001) shows that the response rates of patients to medications from different therapeutic classes varied from 80% (analgesics) to 25% (oncology). This motivates us to set up the group sequential design where the responses from the treatment group patients potentially follow a mixture of normal distributions. Similar to the conventional group sequential design, we test the null hypothesis that there is no difference between the distribution of responses in the control and treatment group. However, the power is determined through the specification of two parameters. Namely, a mixing proportion parameter and mean shift parameter. We refer to this as a ‘mixture alternative’ in later sections. Our examples will illustrate the need for larger arm sizes when trying to detect a mixture alternative compared to trying to detect a pure shift alternative.

Jeske and Yao (in review) discussed the sample size determination and treatment effect estimation if a sub-population has a higher likelihood of benefiting from a treatment for a fixed sample test nonparametric setting. This dissertation focuses on the mixtures of normal for treatment group patients in group sequential setting.

1.3 Group Sequential Designs to Detect a Mixture Alternative

No literature is available for designing group sequential designs to detect a mixture alternative. A related idea is the evaluation of a treatment effect in randomized clinical trials using a genomic biomarker by Wang, O'Neil, and Hung (2007). Our work, and Wang's paper both consider that a subset of treatment group patients might have a higher likelihood of benefiting from a treatment. However we differ from Wang's work in the following aspects: 1) Wang's test utilizes known subgroup information, while the subgroup information is unknown in our setting; 2) Wang's work considers two test statistics (one using biomarker positive patients and one using all the patients), while our work uses only one test statistic; 3) Wang's work analyses the treatment effect for the biomarker positive subgroup and the overall population simultaneously, while our work only focuses on the overall treatment effect.

The main contributions of this dissertation are:

- 1) Stopping boundaries and arm size for the proposed design are determined by Type I and Type II error spending equations. An R program with user documentation is provided.
- 2) Group sequential t-tests in mixture setting is discussed to cover the unknown variance case.
- 3) With the mixture model, we discuss a more general definition of treatment effect.

The maximum likelihood estimator and method of moment estimator for the treatment effect are proposed. We also apply bootstrap algorithm to reduce bias. It

turns out MLE has the best performance in terms of bias and mean squared error.

1.4 Dissertation Structure

The rest of the dissertation is organized as follows. In Chapter 2, we review fixed sample test and group sequential test when both control and treatment group patients come from normal distributions. In Chapter 3, we introduce the fixed sample test when the treatment group responses potentially follow a mixture of normal distributions. Then we describe the proposed one-sided group sequential designs to detect a mixture alternative. In Chapter 4, we discuss implementation details and practical concerns related to the group sequential designs, which includes the group sequential t-tests, estimator for treatment effect, and overview of the R codes. We conclude in Chapter 5 with a brief summary and future works.

Chapter 2

Preliminaries on Group Sequential

Methods

2.1 A Fixed Sample Test

Let X_{Ci} and X_{Ti} , $i = 1, 2, \dots, m$, denote the responses of subjects allocated to two treatments, C and T . Suppose $X_{Ci} \sim N(\mu_C, \sigma^2)$. Likewise, suppose $X_{Ti} \sim N(\mu_T, \sigma^2)$. All observations are independent with unknown means μ_C and μ_T and common known

variance σ^2 . Testing problem could either be one-sided or two-sided. For a one-sided test, the null hypothesis of no treatment effect against the one-sided alternative that the treatment is superior to the control. For a two-sided test, the null hypothesis of no treatment effect against the alternative of an treatment effect in either direction.

Specifically, for one-sided test, we wish to test,

$$H_0: \mu_C = \mu_T \text{ vs } H_0: \mu_T - \mu_C > 0.$$

Also we wish this test has Type I error rate probability α and power $1 - \beta$ at a specified value of $\delta = \mu_T - \mu_C$.

A fixed sample test based on a standardized test statistic Z attains Type I error probability α by rejecting H_0 if $Z > \Phi^{-1}(1 - \alpha)$, where

$$Z = \frac{1}{\sqrt{2m\sigma^2}} \left(\sum_{i=1}^m X_{Ti} - \sum_{i=1}^m X_{Ci} \right). \quad (2.1)$$

Hence, in order to have power $1 - \beta$ at a specified value of δ , the required arm size is

$$m_f = \frac{\{\Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta)\}^2 2\sigma^2}{\delta^2}. \quad (2.2)$$

The final conclusion is,

If $Z > \Phi^{-1}(1 - \alpha)$ Reject H_0

Otherwise Accept H_0

Since sample size is collected to achieve a target power to detect the specified difference δ (which is the smallest difference that is practically important), the phrase ‘accept H_0 ’ is appropriate. Power is controlled for mean shift greater than δ , and power for mean shift in $(0, \delta)$ is not of particular interest since shifts of that size are not consequential.

2.2 Group Sequential Setting and Notation

For a group sequential design with a maximum of K analyses, the data is analysed after every group of $2m$ patients have been accrued. It should be noted that this m is different from the arm size in the fixed sample test. This m is the arm size per stage for the group sequential design, and it is determined to satisfy the power requirement.

For the analysis at stage k , all the observations collected up to that point are used to calculate the test statistic. Then the test statistic is compared with the stopping boundaries at stage k , and the trial is either stopped or continued. If the trial is in the final stage, the null hypothesis is rejected or accepted.

Stopping rules could be either early stopping with efficacy boundaries, or early stopping with efficacy and futility boundaries. The efficacy boundaries and futility boundaries are also called rejection boundaries and acceptance boundaries correspondingly. Choice of the test (one-sided test or two-sided test) and stopping rules

(efficacy boundaries, or efficacy and futility boundaries) depends on the purpose of the clinical trial.

Figure 2.1 illustrates the general form for a one-sided test with efficacy boundaries only, a one-sided test with efficacy and futility boundaries, a two-sided test with efficacy boundaries only, and a two-sided test with efficacy and futility boundaries.

Figure 2.1: General Form for Stopping Boundaries

For a one-sided group sequential test with efficacy boundaries only, it takes the form:

After stage $k = 1, \dots, K - 1$

If $Z_k \geq c_k$ stop, reject H_0

Otherwise continue to stage $k + 1$

After stage K

If $Z_K \geq c_K$ stop, reject H_0

Otherwise stop, accept H_0 .

For a one-sided group sequential test with efficacy and futility boundaries, it takes the form:

After stage $k = 1, \dots, K - 1$

If $Z_k \geq r_k$ stop, reject H_0

If $Z_k \leq a_k$ stop, accept H_0

Otherwise continue to stage $k + 1$

After stage K

If $Z_K \geq r_K$ stop, reject H_0

If $Z_K \leq a_K$ stop, accept H_0

Here, $a_K = r_K$ ensures that the test terminates at analysis K .

For a two-sided group sequential test with efficacy boundaries only, it takes the form:

After stage $k = 1, \dots, K - 1$

If $|Z_k| \geq c_k$ stop, reject H_0

Otherwise continue to stage $k + 1$

After stage K

If $|Z_K| \geq c_K$ stop, reject H_0

Otherwise stop, accept H_0 .

For a two-sided group sequential test with efficacy and futility boundaries, it takes the form:

After stage $k = 1, \dots, K - 1$

If $|Z_k| \geq r_k$ stop, reject H_0

If $|Z_k| \leq a_k$ stop, accept H_0

Otherwise continue to stage $k + 1$

After stage K

If $|Z_K| \geq r_K$ stop, reject H_0

If $|Z_K| \leq a_K$ stop, accept H_0

Here, $a_K = r_K$ ensures that the test terminates at analysis K .

Similar as the fixed sample test, suppose observations X_{Ci} on control group and X_{Ti} on treatment group, $i = 1, 2, \dots$, are independent distributed with unknown means μ_C and μ_T and common known variance σ^2 . Denote $\mu_C = \mu_0$ and $\mu_T = \mu_0 + \delta$. Under normal assumption, $X_{Ci} \sim N(\mu_0, \sigma^2)$ and $X_{Ti} \sim N(\mu_0 + \delta, \sigma^2)$. Table 2.1 presents the notation in the group sequential setting.

Suppose a group sequential study with up to K analyses yields the sequence of test statistics $\{Z_1, \dots, Z_K\}$. These statistics have joint distribution,

- (i) $(Z_1, \dots, Z_K)$ is multivariate normal
- (ii) $E(Z_k) = \delta \sqrt{I_k}$, $k = 1, \dots, K$
- (iii) $cov(Z_{k1}, Z_{k2}) = \sqrt{I_{k1}/I_{k2}}$, $1 \leq k_1 \leq k_2 \leq K$.

This implies that $\{Z_1, \dots, Z_K\}$ is a Markov sequence, which is important in simplifying calculations for group sequential tests

$$Z_k | Z_{k-1}, \dots, Z_1 \sim Z_k | Z_{k-1} \sim N \left(\frac{\delta \Delta_k + Z_{k-1} \sqrt{I_{k-1}}}{\sqrt{I_k}}, \frac{\Delta_k}{I_k} \right). \quad (2.3)$$

K	Maximum possible number of analyses
α	Type I error rate
β	Type I error rate
m	Arm size per stage
μ_0	Mean for the control group patients
σ^2	Variance for control and treatment group patients
$\delta = \mu_T - \mu_C$	Mean difference between the control and treatment group, and it is the parameter we are interested in
$\widehat{\delta}_k = \overline{X}_T^{(k)} - \overline{X}_C^{(k)}$	Estimator of δ using cumulative data up to stage k
$I_k = (\text{var}(\widehat{\delta}_k))^{-1} = \frac{mk}{2\sigma^2}$	Information level at stage k
$\Delta_k = I_k - I_{k-1}$	Increment of information level for stage k
$Z_k = \widehat{\delta}_k \sqrt{I_k}$	Standardized test statistic at stage k
a_K	Acceptance boundary at stage k
r_K	Rejection boundary at stage k

Table 2.1: Notation in Group Sequential Designs

2.3 Group Sequential Tests Fixed Boundary Shape Methods

The fixed boundary shape methods include Pocock, O'Brien and Fleming, and Wang and Tsiatis. These methods derive boundaries with specified boundary shapes. And these tests require that equal increments of sample size.

2.3.1 Pocock

Pocock (1977) applies repeated significance tests to group sequential trials with equal-size groups for two-sided test with efficacy boundaries. It derives a constant critical value on the standardized normal scale across all stages that maintains the specified Type I error level. Figure 2.2 shows the plot for two-sided Pocock test boundaries with $K = 5$.

Figure 2.2: Two-sided Pocock Test with Efficacy Boundaries for $K=5$

At stage k , the null hypothesis is rejected if $|Z_k| > C_p(K, \alpha), k = 1, \dots, K$. Further, the critical value $C_p(K, \alpha)$ is determined by the overall Type I error α ,

$$\begin{aligned} \alpha = & P_{\delta=0}\{|Z_1| \geq C_p(K, \alpha)\} + \dots \\ & + P_{\delta=0}\{|Z_1| < C_p(K, \alpha), \dots, |Z_K| \geq C_p(K, \alpha)\}. \end{aligned} \quad (2.4)$$

Above test has Type I error rate α for any arm size. In order to determine the appropriate arm size per stage m , we need to specify the power requirement under the specified value of δ ,

$$1 - \beta = P_{\delta}\{|Z_1| \geq C_p(K, \alpha)\} + \dots + P_{\delta}\{|Z_1| < C_p(K, \alpha), \dots, |Z_K| \geq C_p(K, \alpha)\}. \quad (2.5)$$

To be conservative, we round m upward to obtain an integer number, so that the attained power is a little greater than the target value $1 - \beta$.

The maximum arm size of the group sequential test needed to achieve target power depends on K , α , and β , and is proportional to σ^2/δ^2 . As it is shown in equation (2.2), the arm size of a fixed sample test with the same Type I error rate and power is also proportional to σ^2/δ^2 . As such, it is sufficient to specify the ratio of the maximum arm size of the group sequential test to the arm size of fixed sample test. This ratio $R_p(K, \alpha, \beta)$ can be applied to all δ and σ^2 . Table 2.2 below sourced from Jennison and Turnbull (1999) Table 2.2 illustrates the ratio for Pocock tests. $K = 1$ corresponding to the fixed sample test, and we can see that when $K > 1$, the maximum arm size for the group sequential design is larger than the arm size for the fixed sample test.

Once the ratio is available, we can determine the maximum arm size of the group sequential test by multiplying the ratio to the arm size for the fixed sample test. The arm size at each stage can then be obtained by dividing the maximum arm size by K .

$R_p(K, \alpha, \beta)$						
K	$1 - \beta = 0.8$			$1 - \beta = 0.9$		
	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
1	1.000	1.000	1.000	1.000	1.000	1.000
2	1.092	1.110	1.121	1.084	1.100	1.110
3	1.137	1.166	1.184	1.125	1.151	1.166
4	1.166	1.202	1.224	1.152	1.183	1.202
5	1.187	1.229	1.254	1.170	1.207	1.228
6	1.203	1.249	1.277	1.185	1.225	1.249
7	1.216	1.265	1.296	1.197	1.239	1.266
8	1.226	1.279	1.311	1.206	1.252	1.280
9	1.236	1.291	1.325	1.215	1.262	1.292
10	1.243	1.301	1.337	1.222	1.271	1.302
11	1.250	1.310	1.348	1.228	1.279	1.312
12	1.257	1.318	1.357	1.234	1.287	1.320
15	1.272	1.338	1.381	1.248	1.305	1.341
20	1.291	1.363	1.411	1.264	1.327	1.367

Table 2.2: Pocock tests: ratio of the maximum arm size for the group sequential test to the arm size of the fixed sample test

2.3.2 O'Brien and Fleming

O'Brien and Fleming (1979) propose a sequential procedure that has boundary values (in absolute value) decrease over the stages on the standardized normal Z scale. Figure 2.3 shows the plot for two-sided O'Brien and Fleming boundaries for $K=5$. As the trial progresses, it is easier to reject the null hypothesis. The test is defined through sequence of test statistics, and for two-sided test with efficacy boundaries, it takes the form:

After stage $k = 1, \dots, K - 1$

If $|Z_k| > C_B(K, \alpha)\sqrt{K/k}$ stop, reject H_0

Otherwise continue to stage $k + 1$

After stage K

If $|Z_k| > C_B(K, \alpha)$ stop, reject H_0

Otherwise stop, accept H_0

Figure 2.3: Two-sided O'Brien and Fleming Test with Efficacy Boundaries for $K=5$

Value of $C_B(K, \alpha)$ is determined to ensure the overall Type I error α ,

$$\begin{aligned} \alpha = & P_{\delta=0}\{|Z_1| \geq C_B(K, \alpha)\sqrt{K}\} + \dots \\ & + P_{\delta=0}\{|Z_1| < C_B(K, \alpha)\sqrt{K} \dots, |Z_K| \geq C_B(K, \alpha)\sqrt{K/k}\}. \end{aligned} \quad (2.6)$$

Similar as Pocock test, above test has Type I error α for any arm size. In order to determine the appropriate arm size per stage m , we need to specify the power requirement under the specified δ ,

$$\begin{aligned}
1 - \beta &= P_{\delta}\{|Z_1| \geq C_B(K, \alpha)\sqrt{K}\} + \dots \\
&+ P_{\delta}\{|Z_1| < C_B(K, \alpha)\sqrt{K}, \dots, |Z_K| \geq C_B(K, \alpha)\}.
\end{aligned}
\tag{2.7}$$

Table 2.3 below sourced from Jennison and Turnbull (1999) Table 2.4 illustrates the ratio of the maximum arm size for the group sequential O'Brien and Fleming tests to the arm size of the fixed sample test.

<i>K</i>	<i>R_B(K, α, β)</i>					
	<i>1 - β = 0.8</i>			<i>1 - β = 0.9</i>		
	<i>α = 0.01</i>	<i>α = 0.05</i>	<i>α = 0.10</i>	<i>α = 0.01</i>	<i>α = 0.05</i>	<i>α = 0.10</i>
1	1.000	1.000	1.000	1.000	1.000	1.000
2	1.001	1.008	1.016	1.001	1.007	1.014
3	1.007	1.017	1.027	1.006	1.016	1.025
4	1.011	1.024	1.035	1.010	1.022	1.032
5	1.015	1.028	1.040	1.014	1.026	1.037
6	1.017	1.032	1.044	1.016	1.030	1.041
7	1.019	1.035	1.047	1.018	1.032	1.044
8	1.021	1.037	1.049	1.020	1.034	1.046
9	1.022	1.038	1.051	1.021	1.036	1.048
10	1.024	1.040	1.053	1.022	1.037	1.049
11	1.025	1.041	1.054	1.023	1.039	1.051
12	1.026	1.042	1.055	1.024	1.040	1.052
15	1.028	1.045	1.058	1.026	1.042	1.054
20	1.030	1.047	1.061	1.029	1.045	1.057

Table 2.3: O'Brien and Fleming tests: ratio of the maximum arm size for the group sequential test to the arm size of the fixed sample test

Pocock and O'Brien and Fleming's tests are proposed for two-sided problem with efficacy boundaries, DeMets and Ware (1980) extend them to one-sided test with efficacy and futility boundaries.

2.3.3 Wang and Tsiatis

Wang and Tsiatis (1987) introduce a family of two-sided test indexed by a parameter Δ .

The test takes the form:

After stage $k = 1, \dots, K - 1$

If $|Z_k| > C_{WT}(K, \alpha, \Delta)(k/K)^{\Delta-1/2}$ stop, reject H_0

Otherwise continue to stage $k + 1$

After stage K

If $|Z_K| > C_{WT}(K, \alpha, \Delta)$ stop, reject H_0

Otherwise stop, accept H_0

Taking $\Delta = 1/2$ and $\Delta = 0$, the test will reduce to Pocock and O'Brien & Fleming respectively. Therefore, Pocock test and O'Brien and Fleming test are special cases of Wang and Tsiatis test.

Table 2.4 below sourced from Jennison and Turnbull (1999) Table 2.10 illustrates ratio of the maximum arm size for the group sequential Wang and Tsiatis tests to the arm size of the fixed sample test.

$R_{WT}(K, \alpha, \beta, \Delta)$						
K	$\alpha = 0.05$					
	$1 - \beta = 0.8$			$1 - \beta = 0.9$		
	$\Delta = 0.1$	$\Delta = 0.25$	$\Delta = 0.4$	$\Delta = 0.1$	$\Delta = 0.25$	$\Delta = 0.4$
1	1.000	1.000	1.000	1.000	1.000	1.000
2	1.016	1.038	1.075	1.014	1.034	1.068
3	1.027	1.054	1.108	1.025	1.050	1.099
4	1.035	1.065	1.128	1.032	1.059	1.117
5	1.040	1.072	1.142	1.037	1.066	1.129
6	1.044	1.077	1.152	1.041	1.071	1.138
7	1.047	1.081	1.159	1.044	1.075	1.145
8	1.050	1.084	1.165	1.046	1.078	1.151
9	1.052	1.087	1.170	1.048	1.081	1.155
10	1.054	1.089	1.175	1.050	1.083	1.159
11	1.055	1.091	1.178	1.051	1.085	1.163
12	1.056	1.093	1.181	1.053	1.086	1.166
15	1.059	1.097	1.189	1.055	1.090	1.172
20	1.062	1.101	1.197	1.058	1.094	1.180

Table 2.4: Wang and Tsiatis tests: ratio of the maximum arm size for the group sequential test to the arm size of the fixed sample test

Wang and Tsiatis's test considers efficacy boundaries. Pampallona and Tsiatis (1994) extend it group sequential tests with early stopping to reject for efficacy or accept for futility boundaries. For the test with parameter Δ , at stage k , stopping boundaries are

$$\begin{aligned} r_k &= \widetilde{C}_1(K, \alpha, \beta, \Delta) \left(\frac{k}{K} \right)^{\Delta - \frac{1}{2}} \\ a_k &= \delta \sqrt{I_k} - \widetilde{C}_2(K, \alpha, \beta, \Delta) \left(\frac{k}{K} \right)^{\Delta - \frac{1}{2}}. \end{aligned} \quad (2.8)$$

In order that the final boundaries coincide, $a_K = r_K = u$, the final information level needs to be

$$I_K = \frac{\{\widetilde{C}_1(K, \alpha, \beta, \Delta) + \widetilde{C}_2(K, \alpha, \beta, \Delta)\}^2}{\delta^2}, \quad (2.9)$$

where the constants $\widetilde{C}_1(K, \alpha, \beta, \Delta)$ and $\widetilde{C}_2(K, \alpha, \beta, \Delta)$ are determined by size and power conditions.

2. 4 Group Sequential Tests Error Spending Methods

Error spending function approach is a more flexible way to dealing with unpredictable information sequences. This method does not require the maximum number of analyses to be fixed in advance, and this method does not require equal increments of sample size. To tackle a case of maximum number of analyses is not fixed and unequal increments of sample size, we can set up the design with a planned assuming K and equal increments of sample size assumption first. Then boundaries are adjusted to handle the unequal sample size increments. The trial continues until either the stopping decision is made or the maximum arm size is reached.

2. 4. 1 Efficacy Stopping Rules

Slud and Wei (1982) introduce the initial idea about the error spending method. However, this design is only for fixed maximum number of analyses. Lan and DeMets (1983) extend this idea. Specifically, Lan and DeMets' method does not require the maximum number of analyses to be fixed in advance and it allows the error spent at analyses varies in response to the observed information levels. The group sequential design proposed by Lan and

DeMets is for the case of early stopping to reject the null hypothesis for two-sided test. This design is constructed based on the Type I error spending function.

Error spending function is an equation that serves to guide us in how much error to spend at a given analysis. Type I spending function $f(t)$ is used to partition Type I error into probabilities $\pi_{1,1}, \dots, \pi_{1,K}$, which sum to α . It is a non-decreasing function, and satisfies the conditions $f(0) = 0$ and $f(t) = \alpha$ for $t \geq 1$. Partial alpha spent at a given step is given by

$$\begin{aligned}\pi_{1,1} &= f(t_1) \\ \pi_{1,k} &= f(t_k) - f(t_{k-1}) \text{ for } k = 2, \dots, K.\end{aligned}\tag{2.10}$$

For maximum information trials, t_k is the current number of completed patients over the number of patients need to achieve power $1 - \beta$. For equal increments of sample size, we have

$$t_k = \frac{mk}{mK} = \frac{k}{K} \text{ for } k = 1, 2, \dots, K.\tag{2.11}$$

Hwang, Shih, and De Cani (1990) introduce a family of error spending functions indexed by a parameter γ ,

$$f(t) = \begin{cases} \frac{\alpha(1 - e^{-\gamma t})}{1 - e^{-\gamma}} & \text{for } \gamma \neq 0 \\ \alpha t & \text{for } \gamma = 0 \end{cases}\tag{2.12}$$

Jennison and Turnbull (1999) suggest the simple functions, parameterized by the power $\rho > 0$, and $f(t) = \alpha t^\rho (0 < t \leq 1)$. Figure 2.4 illustrates the Type I and Type II error spending functions for $\rho \in \{1, 2, 3\}$.

Figure 2.4: Type I and Type II error spending functions with $\alpha = 0.05, \beta = 0.2$

The K stopping boundary values are determined by the K Type I error probability equations,

$$\pi_{1,1} = P_{\delta=0}\{|Z_1| \geq r_1\} \quad (2.13)$$

$$\pi_{1,k} = P_{\delta=0}\{|Z_1| < r_1, \dots, |Z_{k-1}| < r_{k-1}, |Z_k| > r_k\} \text{ for } k = 2, \dots, K.$$

In addition, we have power equation to determine arm size for the specified δ ,

$$1 - \beta = P_\delta\{|Z_1| \geq r_1\} + \cdots +$$

$$P_\delta\{|Z_1| < r_1, \dots, |Z_{k-1}| < r_{k-1}, |Z_K| > r_K\}$$

Similar as fixed boundary shape methods, for Lan and DeMets' maximum information test, the ratio of the maximum arm size for the group sequential test to the arm size of the fixed sample test could apply for all δ and σ^2 . Table 2.5 below sourced from Jennison and Turnbull (1999) Table 7.1 illustrates the ratio $R_{LD}(K, \alpha, \beta)$ for two-sided tests with Type I error spending functions $f(t) = \min\{\alpha t^\rho, \alpha\}$. Once the ratio is available, we can determine the maximum arm size of the group sequential test by multiplying the ratio to the arm size for fixed sample test.

$R_{LD}(K, \alpha, \beta, \rho)$						
K	$1 - \beta = 0.8$			$1 - \beta = 0.9$		
	$\rho = 1$	$\rho = 2$	$\rho = 3$	$\rho = 1$	$\rho = 2$	$\rho = 3$
1	1.000	1.000	1.000	1.000	1.000	1.000
2	1.082	1.028	1.010	1.075	1.025	1.009
3	1.117	1.045	1.020	1.107	1.041	1.018
4	1.137	1.056	1.027	1.124	1.051	1.025
5	1.150	1.063	1.032	1.136	1.058	1.030
6	1.159	1.069	1.036	1.144	1.063	1.033
7	1.165	1.073	1.039	1.150	1.067	1.036
8	1.170	1.076	1.041	1.155	1.070	1.039
9	1.174	1.079	1.043	1.159	1.073	1.040
10	1.178	1.081	1.045	1.162	1.075	1.042
11	1.180	1.083	1.046	1.164	1.077	1.043
12	1.183	1.085	1.048	1.166	1.078	1.044
15	1.188	1.088	1.050	1.171	1.082	1.047
20	1.193	1.092	1.054	1.176	1.085	1.050

Table 2.5: Two-sided group sequential tests using Type I error spending functions: ratio of the maximum arm size for the group sequential test to the arm size of the fixed sample test

2. 4. 2 Efficacy and Futility Stopping Rules

Chang, Hwang and Shih (1998) extend Lan and Demets' method to group sequential design for the case of early stopping to reject for efficacy or accept for futility. Construction of such design based on both Type I and Type II error spending functions.

Similar as the Type I error spending function, Type II error spending function $g(t)$ is non-decreasing and satisfy $g(0) = 0$ and $g(t) = \beta$ for $t \geq 1$. The efficacy and futility group sequential boundaries are defined through both Type I and Type II error probabilities. Partial beta spending for a specified δ at a given step is given by

$$\begin{aligned}\pi_{2,1} &= g(t_1) \\ \pi_{2,k} &= g(t_k) - g(t_{k-1}) \text{ for } k = 2, \dots, K.\end{aligned}\tag{2.15}$$

For the simple functions parameterized by the power ρ , it is $g(t) = \beta t^\rho$ ($0 < t \leq 1$).

For a one-sided test, stopping boundaries (a_k, r_k) , $k = 1, 2, \dots, K$, and arm size are solved using numerical methods by solving successively

$$\begin{aligned}P_{\delta=0}\{Z_1 > r_1\} &= \pi_{1,1} \\ P_{\delta}\{Z_1 < a_1\} &= \pi_{2,1} \\ P_{\delta=0}\{a_1 < Z_1 < r_1, \dots, a_{k-1} < Z_{k-1} < r_{k-1}, Z_k > r_k\} &= \pi_{1,k}, \quad k = 2, \dots, K \\ P_{\delta}\{a_1 < Z_1 < r_1, \dots, a_{k-1} < Z_{k-1} < r_{k-1}, Z_k < a_k\} &= \pi_{2,k}, \quad k = 2, \dots, K.\end{aligned}\tag{2.16}$$

Equations about $\pi_{2,k}$ are under the specified δ . At final stage K , we force final boundary points coincide, i.e. $a_K = r_K = u$, so that test terminates by analysis K with error rates exactly α and β .

Table 2.6 below sourced from Jennison and Turnbull (1999) Table 7.6 illustrates ratio of the maximum arm size for the one-sided group sequential test to the arm size of the fixed sample test. The group sequential test has Type I and Type II error spending functions $f(t) = \min\{\alpha t^\rho, \alpha\}$ and $g(t) = \min\{\beta t^\rho, \beta\}$.

$R_{OS}(K, \alpha, \beta, \rho)$						
K	$1 - \beta = 0.8$		$1 - \beta = 0.9$		$1 - \beta = 0.95$	
	$\rho = 2$	$\rho = 3$	$\rho = 2$	$\rho = 3$	$\rho = 2$	$\rho = 3$
1	1.000	1.000	1.000	1.000	1.000	1.000
2	1.043	1.014	1.044	1.015	1.045	1.016
3	1.070	1.028	1.072	1.030	1.073	1.031
4	1.087	1.038	1.089	1.040	1.090	1.042
5	1.098	1.045	1.100	1.048	1.101	1.050
6	1.106	1.050	1.108	1.053	1.109	1.055
7	1.111	1.054	1.114	1.058	1.115	1.060
8	1.116	1.058	1.119	1.061	1.120	1.063
9	1.120	1.060	1.123	1.064	1.124	1.066
10	1.123	1.062	1.126	1.066	1.127	1.069
11	1.125	1.064	1.128	1.068	1.130	1.071
12	1.127	1.066	1.131	1.070	1.132	1.072
15	1.132	1.069	1.135	1.073	1.137	1.076
20	1.137	1.073	1.140	1.077	1.142	1.080

Table 2.6: One-sided group sequential tests using Type I and Type II error spending functions: ratio of the maximum arm size for the group sequential test to the arm size of the fixed sample test

2.5 Estimation Following Group Sequential Tests

Once data have been collected, and the “accept” or “reject” decision has been made. In general, further measures of the treatment difference is needed. In this section, we review methods to estimate the treatment difference.

2.5.1 Point Estimation

The MLE for the parameter δ is the sample mean difference, which is given by

$$\hat{\delta} = \bar{X}_T - \bar{X}_C. \quad (2.17)$$

Due to the selective nature of the stopping rules, $\hat{\delta}$ is a biased estimator of δ .

For a two-stage group sequential test with continuation region $C_1 = (a, r)$ at stage 1, Emerson (1988) derives the analytic expression for this bias, which is given by,

$$b(\delta) = E_{\delta}(\hat{\delta}) - \delta = \frac{\Delta_2}{I_2\sqrt{I_1}} \{\phi(r - \delta\sqrt{I_1}) - \phi(a - \delta\sqrt{I_1})\}. \quad (2.18)$$

And this bias only equals to zero when δ is $(a + r)/(2\sqrt{I_1})$.

2.5.2 Bias Reduction Methods for Point Estimates

2.5.2.1 Whitehead Method

In order to reduce the bias for MLE, Whitehead (1986) suggests subtracting an estimate of the bias of the MLE. The expected value of $\hat{\delta}$ is given by $E_{\delta}(\hat{\delta}) = \delta + b(\delta)$, and the ideal

bias-corrected estimate of δ is $\hat{\delta} - b(\delta)$. However, in practice, δ is unknown. We can use $\tilde{\delta}$ such that the expected value $E_{\tilde{\delta}}(\hat{\delta})$ would take the value $\hat{\delta}$ if the true parameter value was $\tilde{\delta}$, i.e., $E_{\tilde{\delta}}(\hat{\delta}) = \tilde{\delta} + b(\tilde{\delta}) = \hat{\delta}$, which means we choose $\tilde{\delta}$ to satisfy equation

$$\hat{\delta} = \tilde{\delta} + b(\tilde{\delta}). \quad (2.19)$$

Therefore, the bias-corrected estimator of δ is given by

$$\hat{\delta}_{bc_Whitehead} = \hat{\delta} - b(\tilde{\delta}). \quad (2.20)$$

For a two-stage one-sided group sequential design, equation (2.18) shows the expression of bias. So plug the bias equation into (2.19), for a two-stage one-sided group sequential design, it is equivalent to solve $\tilde{\delta}$ numerically where

$$\hat{\delta} = \tilde{\delta} + \frac{\Delta_2}{I_2\sqrt{I_1}} \{ \phi(r - \tilde{\delta}\sqrt{I_1}) - \phi(a - \tilde{\delta}\sqrt{I_1}) \}. \quad (2.21)$$

Then take solution of $\tilde{\delta}$ into equation (2.20) to get the bias-corrected MLE.

2. 5. 2. 2 Bootstrap Bias Reduction Method

Another method is to use parametric bootstrap approach to find the bias-corrected estimator, as suggested by Wang and Leung (1997). Here we assume variance is unknown, and estimate σ^2 along with other two parameters μ_0 , and δ . The procedure is as follows,

- Calculate initial MLE for μ_0, σ^2, δ
- Generate B bootstrap group sequential sample paths via simulation using $\widehat{\mu}_0, \widehat{\sigma}^2$, and $\widehat{\delta}$. In the j -th bootstrap simulation, a group sequential test is carried out using the original arm size per stage and stopping rule. Post termination, it will yield a bootstrap estimate for $\widehat{\delta}$, which are denoted as $\widehat{\delta}^*, j = 1, \dots, B$;
- The bootstrap bias estimate of δ is

$$\frac{\sum_{j=1}^B \widehat{\delta}_j^*}{B} - \widehat{\delta} \quad (2.22)$$

- The bootstrap bias-corrected MLE estimate of δ is

$$\widehat{\delta}_{bc_resampling} = 2\widehat{\delta} - \frac{\sum_{j=1}^B \widehat{\delta}_j^*}{B}. \quad (2.23)$$

The bootstrap bias reduction method is easy to implement. Therefore, in later section, we will apply this method to the proposed group sequential design when the treatment group responses potentially follow a mixture of normal distributions.

2. 5. 2. 3 Numerical Results

For the following three examples, group sequential designs are constructed using error spending function methods. As it is illustrated in the following three examples, both bias reduction methods reviewed above can reduce the bias considerably, and these two method have comparable performance.

Example 1 Consider $K = 2, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 2, \mu_0 = 0$, and power is 80% for detecting $\delta = 0.3$. The bias is obtained via 10000 simulated sample paths, and the size of bootstrap samples is 1000. Using the R code in Appendix C, the efficacy and futility boundaries, and arm size per stage are as follows,

```
## $lower
## [1] 0.1551464 1.6804027
##
## $upper
## [1] 2.241403 1.680403
##
## $arm_size
## [1] 72
```

Figure 2.5 plots the analytical bias for MLE from equation (2.18), numerical bias for MLE from Monte Carlo simulation, bias of the Whitehead method bias-corrected MLE, and the bias of the bootstrap method bias-corrected MLE. We can see that both Whitehead method and bootstrap resampling method can reduce the bias significantly (more than 50%), and they have comparable performance.

Figure 2.5: Bias of MLE and bias-corrected MLE for Two-Stage One-Sided Group Sequential Design: Example 1

Example 2 Consider $K = 2, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 2, \mu_0 = 0$, and power is 80% for detecting $\delta = 0.5$. The bias is obtained via 10000 simulated sample paths, and the size of bootstrap samples is 1000. Using the R code in Appendix C, the efficacy and futility boundaries, and arm size per stage are as follows,

```
## $lower
## [1] 0.157922 1.679403
##
## $upper
## [1] 2.241403 1.679403
##
## $arm_size
## [1] 26
```

Figure 2.6 plots the analytical bias for MLE from equation (2.18), numerical bias for MLE from Monte Carlo simulation, bias of the Whitehead method bias-corrected MLE, and the bias of the bootstrap method bias-corrected MLE. We can see that both Whitehead method and bootstrap method can reduce the bias significantly (more than 50%), and they have comparable performance.

Figure 2.6: Bias of MLE and bias-corrected MLE for Two-Stage One-Sided Group Sequential Design: Example 2

Example 3 Consider $K = 2, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 2, \mu_0 = 0$, and power is 80% for detecting $\delta = 0.7$. The bias is obtained via 10000 simulated sample paths, and the size of bootstrap samples is 1000. Using the R code in Appendix C, the efficacy and futility boundaries, and arm size per stage are as follows,

```
## $lower
## [1] 0.2313127 1.6744027
##
## $upper
## [1] 2.241403 1.674403
##
## $arm_size
## [1] 11
```

Figure 2.7 plots the analytical bias for MLE from equation (2.18), numerical bias for MLE from Monte Carlo simulation, bias of the Whitehead method bias-corrected MLE, and the bias of the bootstrap method bias-corrected MLE. We can see that both Whitehead method and bootstrap method can reduce the bias significantly (more than 50%), and they have comparable performance.

Figure 2.7: Bias of MLE and bias-corrected MLE for Two-Stage One-Sided Group Sequential Design: Example 3

2. 5. 3 Comparison of MLE and Bias-Corrected MLE

Emerson and Fleming (1990) compare several estimators for four-stage Pocock and O'Brien and Fleming tests. In this section, we focus on the group sequential design constructed using error spending methods, and discuss MLE and bias-corrected MLE for δ . The bias, variance, and MSE are compared for these two estimators.

We consider $K = 3, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 2, \mu_0 = 0$, and power is 80% for detecting the specified group mean difference δ . Results are obtained via 1000 simulated sample paths, and the size of bootstrap samples is 500. Examples below are for various specified values of δ , where $\delta \in \{0.3, 0.5, 0.7, 0.9, 1.1\}$.

As we can see from below graphs, the bias-corrected MLE can reduce bias significantly for the range of δ close to the designed target. Variance of the adjusted MLE is smaller than variance of MLE at the designed target. There is cross over in the range close to the designed target. As a result, bias-corrected MLE has smaller bias and MSE than MLE for the range of δ close to the designed target.

Figure 2.8: Bias, variance and MSE of MLE and bias-corrected MLE of (θ, δ) with $\delta = 0.3$

Figure 2.9: Bias, variance and MSE of MLE and bias-corrected MLE of (θ, δ) with $\delta = 0.5$

Figure 2.10: Bias, variance and MSE of MLE and bias-corrected MLE of (θ, δ) with $\delta = 0.7$

Figure 2.11: Bias, variance and MSE of MLE and bias-corrected MLE of (θ, δ) with $\delta = 0.9$

Figure 2.12: Bias, variance and MSE of MLE and bias-corrected MLE of (θ, δ) with $\delta = 1.1$

Chapter 3

One-Sided Group Sequential Designs to Detect a Mixture Alternative

3.1 A Fixed Sample Test

Jeske and Yao (in review) discussed the sample size determination and treatment effect estimation if a sub-population has a higher likelihood of benefiting from a treatment. Their setting was a fixed sample size and a nonparametric test. This dissertation focuses on a group sequential context and normal theory inference. Let X_{Ci} and X_{Ti} denote the response of subjects from the control and treatment group respectively. Suppose X_{Ci} are

independently identically distributed as $F \sim N(\mu_0, \sigma^2)$, for $i = 1, \dots, m$, and X_{Ti} comes from a mixture of normal distributions $G \sim (1 - \theta)N(\mu_0, \sigma^2) + \theta N(\mu_0 + \delta, \sigma^2)$ with known σ^2 . The whole parameter space is $\theta \in [0, 1], \delta \geq 0, \mu_0 \in R$. We consider the null hypothesis of $H_0: F = G$, against the one-sided alternative that G is a mixture of F and a shift of F .

Define the standardized test statistic Z as

$$Z = \frac{1}{\sqrt{2m\sigma^2}} \left(\sum_{i=1}^m X_{Ti} - \sum_{i=1}^m X_{Ci} \right) = \sqrt{\frac{m}{2\sigma^2}} (\bar{X}_T - \bar{X}_C). \quad (3.1)$$

Under the null hypothesis, Z has standard normal distribution. However, the distribution of Z under the mixture alternative is not straightforward. When m is large enough, the distribution is approximately,

$$Z \sim N \left(\theta \delta \sqrt{\frac{m}{2\sigma^2}}, 1 + \frac{\theta(1 - \theta)\delta^2}{2\sigma^2} \right). \quad (3.2)$$

For a Type I error probability α , we reject H_0 if $Z > Z_\alpha$, where Z_α is the upper α quantile for standard normal distribution. In order to have power $1 - \beta$ at a specified values of (θ, δ) , the required arm size m is approximately,

$$m = \frac{(\sqrt{2}Z_\alpha + \sqrt{2 + \theta(1 - \theta)(\delta/\sigma)^2}Z_\beta)^2}{\theta^2(\delta/\sigma)^2}. \quad (3.3)$$

To be conservative, we round m upward to obtain an integer number.

Table 3.1 below shows the required arm size for some choices of θ and δ/σ based on (3.3). The numbers that are in parentheses are the near exact required arm sizes based on Monte Carlo simulation. Starting with the required arm size for $\theta = 1$ (pure shift setting), the power of the test was evaluated based on 100,000 simulated data sets for a sequence of increasing (by one) arm sizes until the power became greater than 0.8. The required arm sizes based on the approximate formula (3.3) and the Monte Carlo simulation are almost same, which demonstrates that the approximation in (3.2) works well in typical settings.

θ	Treatment Effect Shift Size, δ/σ			
	0.25	0.5	0.75	1
0.5	794 (792)	200 (200)	90 (91)	52 (52)
0.6	551 (551)	139 (139)	63 (63)	36 (36)
0.7	405 (406)	102 (102)	46 (46)	27 (27)
0.8	310 (310)	78 (79)	35 (35)	20 (20)
0.9	245 (244)	62 (62)	28 (28)	16 (16)
1	198 (199)	50 (50)	22 (22)	13 (13)

Table 3.1: Arm size for a 5% level fixed sample one-sided test based on Z that will achieve 80% power for the mixture alternative. Numbers in parentheses are arm sizes for a 5% test based Monte Carlo simulation (simulation size 100,000)

3.2 Group Sequential Tests

In a group sequential design with a maximum of K stages, the data is analyzed after every group of $2m$ patients have been accrued. For the analysis at stage k , the test statistic is computed as

$$Z_k = \frac{1}{\sqrt{2mk\sigma^2}} \left(\sum_{i=1}^{mk} X_{Ti} - \sum_{i=1}^{mk} X_{Ci} \right). \quad (3.4)$$

Under the null hypothesis, the sequence of test statistics has multivariate normal distribution. However, the distribution of the test statistics under the mixture alternative is not straightforward. When m is large, distribution for the sequence of test statistics $\{Z_1, \dots, Z_K\}$ can be approximated by a multivariate normal distribution,

$$\begin{pmatrix} Z_1 \\ Z_2 \\ \dots \\ Z_K \end{pmatrix} \sim MVN \left(\begin{pmatrix} \theta\delta\sqrt{m/(2\sigma^2)} \\ \theta\delta\sqrt{2m/(2\sigma^2)} \\ \dots \\ \theta\delta\sqrt{Km/(2\sigma^2)} \end{pmatrix}, \left(1 + \frac{\theta(1-\theta)\delta^2}{2\sigma^2} \right) \begin{pmatrix} 1 & \sqrt{1/2} & \dots & \sqrt{1/K} \\ \sqrt{1/2} & 1 & \dots & \sqrt{2/K} \\ \dots & \dots & \dots & \dots \\ \sqrt{1/K} & \sqrt{2/K} & \dots & 1 \end{pmatrix} \right). \quad (3.5)$$

A general one-sided group sequential test is defined by pairs of constants (a_k, r_k) with $a_k < r_k$ for $k = 1, \dots, K-1$ and $a_K = r_K = u$. Figure 3.1 shows a general picture of the stopping boundaries, which takes the form:

After stage $k = 1, \dots, K-1$

If $Z_k \geq r_k$ stop, reject H_0

If $Z_k \leq a_k$ stop, accept H_0

Otherwise continue to stage $k+1$

After stage K

If $Z_K \geq u$ stop, reject H_0

If $Z_K < u$ stop, accept H_0

Here, $a_K = r_K = u$ ensures that the test terminates at stage K .

Figure 3.1: General picture for one-sided group sequential design stopping boundaries with $K=5$

For the error spending function approach, the stopping boundaries are determined through error spending functions. Error spending functions $f(t)$ and $g(t)$ are defined for Type I and Type II errors respectively, which are non-decreasing and satisfy $f(0) = 0, g(0) = 0$ and $f(t) = \alpha, g(t) = \beta$ for $t \geq 1$. Chang, Hwang, and Shih (1990) use conservative error spending functions introduced by Hwang, Shih, and DeCani (1990). We consider simple power family functions suggested by Jennison and Turnbull (1999), and it takes the form

$$f(t) = \min\{\alpha, \alpha t^\rho\} \text{ and } g(t) = \min\{\beta, \beta t^\rho\}, t \in [0, 1]$$

for a chosen value of $\rho > 0$. The choice of ρ controls how much Type I error and Type II error probability we want to spend at each stage. For equal error spending at each stage, we take $\rho = 1$. A relatively large ρ value is recommended if practitioners want to be stingy about Type I error and Type II error at early stages. For the case of equal increments of sample size, t at stage k is $t_k = k/K$.

The stopping boundaries and arm size are determined using numerical methods by solving the $2K$ equations below successively. Starting with a value of m , the first $2K - 2$ equations are solved to obtain $(a_k, r_k), k = 1, 2, \dots, K - 1$. Then the last two equation are checked to see if they yield the same solution for u . If the solutions for u are not the same, we increase m and try again. The difference between the pure shift setting and mixture setting is where the power is designed. For the pure shift setting, the power is designed for a specified value of δ . For the proposed setting, the power is designed at the mixture alternative with specification of the mixing proportion θ and mean shift parameter δ . This difference will be reflected in Type II error probability equations. Equation below shows the Type I and Type II error probability equations,

$$\begin{aligned}
P_{F=G}\{Z_1 \geq r_1\} &= \pi_{1,1} \\
P_{\theta,\delta}\{Z_1 \leq a_1\} &= \pi_{2,1} \\
P_{F=G}\{a_1 < Z_1 < r_1, \dots, a_{k-1} < Z_{k-1} < r_{k-1}, Z_k \geq r_k\} &= \pi_{1,k}, k = 2, \dots, K - 1 \\
P_{\theta,\delta}\{a_1 < Z_1 < r_1, \dots, a_{k-1} < Z_{k-1} < r_{k-1}, Z_k \leq a_k\} &= \pi_{2,k}, k = 2, \dots, K - 1 \\
P_{F=G}\{a_1 < Z_1 < r_1, \dots, a_{K-1} < Z_{K-1} < r_{K-1}, Z_K \geq u\} &= \pi_{1,K} \\
P_{\theta,\delta}\{a_1 < Z_1 < r_1, \dots, a_{K-1} < Z_{K-1} < r_{K-1}, Z_K < u\} &= \pi_{2,K},
\end{aligned} \tag{3.6}$$

where $\pi_{1,k}$ and $\pi_{2,k}$ are Type I and Type II error respectively spent at stage k ,

$$\begin{aligned}
\pi_{1,1} &= f(t_1) \\
\pi_{1,k} &= f(t_k) - f(t_{k-1}) \text{ for } k = 2, 3, \dots, K \\
\pi_{2,1} &= g(t_1) \\
\pi_{2,k} &= g(t_k) - g(t_{k-1}) \text{ for } k = 2, 3, \dots, K.
\end{aligned} \tag{3.7}$$

For the pure shift setting, stopping boundaries and arm size can be determined using SAS proc seqdesign. For the proposed mixture alternative setting, an R program with user documentation is provided in Appendix C. Generalization of the sub-density computational approach to our mixture setting is described in Appendix A. Table 3.2 shows the runtime of the R function for $K \in \{3, 5, 10, 15\}$. When $K = 15$, the runtime was only 141.69 s, which is computationally efficient.

K	Runtime
3	25.24 s
5	59.96 s
10	141.69 s

Table 3.2: Runtime for $\alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \delta = 0.5, \theta = 0.7, \rho = 2$, and specified $(\theta, \delta) = (0.7, 0.5)$

3.3 Properties of the Proposed Group Sequential Designs

From equation (3.6), it follows that the maximum sample size per arm mK depends on the specified $\alpha, \beta, \theta, \rho, K$, and δ/σ . Table B.1-Table B.3 in Appendix B show the ratio of the maximum sample size per arm to the arm size for the fixed sample test for the case $\alpha =$

0.05, $\beta = 0.2$, and $\rho = 1, 2, 3$. We can see there is very little dependence on the specified θ or δ/σ , although dependence on K is evident. Similar results are observed for different choices of α, β and ρ . Therefore, similar to the pure shift setting, it is useful to tabulate the ratio of the maximum arm size for the group sequential test to the arm size of the fixed sample test for values of K, α, β , and ρ . Values of this ratio $R(K, \alpha, \beta, \rho)$ are listed in Table 3.3 for $K \in \{2, 3, 4, 5\}$, $\alpha \in \{0.01, 0.05, 0.1\}$, $1 - \beta \in \{0.8, 0.85, 0.9\}$, and $\rho \in \{1, 2, 3\}$.

α	K	$1 - \beta = 0.8$			$1 - \beta = 0.9$			$1 - \beta = 0.95$		
		$\rho = 1$	$\rho = 2$	$\rho = 3$	$\rho = 1$	$\rho = 2$	$\rho = 3$	$\rho = 1$	$\rho = 2$	$\rho = 3$
0.01	2	1.135	1.043	1.015	1.132	1.043	1.016	1.129	1.043	1.016
	3	1.191	1.070	1.030	1.188	1.071	1.032	1.183	1.072	1.032
	4	1.223	1.087	1.041	1.219	1.089	1.043	1.213	1.089	1.044
	5	1.244	1.098	1.049	1.238	1.100	1.051	1.232	1.100	1.052
0.05	2	1.145	1.043	1.014	1.143	1.045	1.015	1.139	1.045	1.016
	3	1.207	1.070	1.028	1.203	1.072	1.030	1.198	1.073	1.031
	4	1.242	1.088	1.038	1.237	1.089	1.040	1.230	1.090	1.042
	5	1.264	1.099	1.046	1.258	1.101	1.048	1.251	1.102	1.051
0.1	2	1.148	1.043	1.013	1.146	1.044	1.014	1.143	1.045	1.015
	3	1.212	1.069	1.026	1.209	1.072	1.029	1.203	1.072	1.030
	4	1.248	1.086	1.036	1.243	1.089	1.039	1.237	1.089	1.041
	5	1.271	1.097	1.043	1.266	1.100	1.046	1.258	1.101	1.049

Table 3.3: Proposed one-sided group sequential tests: ratio of the maximum arm size for the group sequential test to the arm size of the fixed sample test with Type I error rate 1%, 5%, and 10%

Table 3.3 implies that the mixture alternative setting and the pure shift setting have same ratio of maximum arm size for the group sequential test to the arm size for the fixed sample test. Since the fixed sample test arm size is bigger for the mixture setting than the pure shift setting. It follows that the arm size for group sequential test mixture alternative setting is also larger than the arm size in the pure shift setting.

Appendix B discusses the fact that when the specified values of θ and δ/σ are in practical ranges, they have a small impact on the stopping boundaries. Thus, the stopping boundaries for the pure shift setting could serve as good approximations for the stopping boundaries in the mixture setting.

3.4 Examples Illustration

Consider a group sequential design with a maximum number of analyses $K = 3$, Type I error rate 5%, Type II error rate 20%, variance $\sigma^2 = 1$, and $\rho = 2$ for both Type I and Type II spending functions. Assume we wish to detect the mixture that has $\theta = 0.7$ and $\delta = 0.5$.

Using the R code in Appendix C, it is determined that arm size per stage is $m = 37$, which is significantly larger than the arm size $m = 18$ that is needed to detect a pure shift ($\theta = 1$). The blue solid lines and numbers listed in Figure 3.2 illustrate the stopping boundaries. In addition, the stopping boundaries for the pure shift setting are also plotted in Figure 3.2. The insignificant difference between the stopping boundaries for the mixture setting and pure shift setting illustrate adequacy of the previously discussed (section 3.3) near-equivalence of the two sets of stopping boundaries.

Table 3.4 compares the target and achieved Type I and Type II error probabilities at each stage. The achieved Type I and Type II error probabilities are obtained through simulation. The sample path is simulated 100,000 times with stopping boundaries and m stated above, then $\pi_{1,k}$ and $\pi_{2,k}$ are calculated based on the simulation results. As it is

shown in the table, the target and achieved Type I and Type II error probabilities are very close.

Figure 3.2: Stopping boundaries for $K = 3, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \delta = 0.5, \theta = 0.7, \rho = 2$

	Type I error $\pi_{1,k}$		Type II error $\pi_{2,k}$	
	Target	Simulated	Target	Simulated
$k = 1$	0.0056	0.0053	0.0222	0.0227
$k = 2$	0.0167	0.0168	0.0667	0.0687
$k = 3$	0.0278	0.0282	0.1111	0.1053
Overall	0.0500	0.0503	0.2000	0.1968

Table 3.4: Target and achieved Type I and Type II error at each stage for $K = 3, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 2$, and specified $(\theta, \delta) = (0.7, 0.5)$

Chapter 4

Implementation Details and Practical Concerns

4.1 Unknown σ^2

In practice, σ^2 is unknown and an estimate needs to be used instead. Jennison and Turnbull (1999) section 4.4 discuss the group sequential t-tests for a pure shift setting. Following that, we introduce the group sequential one-sided t-tests for the mixture setting, which is formulated by replacing the unknown σ^2 in Z_k with an estimator s_k^2 to get,

$$T_k = \frac{1}{\sqrt{2mks_k^2}} \left(\sum_{i=1}^{mk} X_{Ti} - \sum_{i=1}^{mk} X_{Ci} \right). \quad (4.1)$$

Table 4.1 below compares the estimates of σ^2 using control group responses only, and estimator of σ^2 using the pooled variance estimates form control and treatment group responses.

design target		Type I error rate (s.e.)		Type II error rate (s.e.)	
theta	delta	Control	Control+ Treatment	Control	Control+ Treatment
1	1	0.0752 (0.0026)	0.1014 (0.0030)	0.2091 (0.0041)	0.1431 (0.0035)
0.75	1	0.0607 (0.0024)	0.0901 (0.0029)	0.1796 (0.0038)	0.1434(0.0035)
0.5	1	0.0567 (0.0023)	0.0829 (0.0027)	0.1907 (0.0039)	0.1795 (0.0038)
0.25	1	0.0522 (0.0022)	0.0759 (0.0026)	0.1996 (0.0040)	0.1849 (0.0039)
1	0.5	0.0553 (0.0023)	0.083 (0.0027)	0.1960 (0.0040)	0.1576 (0.0036)
0.75	0.5	0.0524 (0.0022)	0.0822 (0.0027)	0.2014 (0.0040)	0.1636 (0.0037)
0.5	0.5	0.0493 (0.0022)	0.0725 (0.0026)	0.2030 (0.0040)	0.1834 (0.0039)

Table 4.1: Type I and Type II error rate for two variance estimates. $K = 2, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 1, \mu_0 = 0$, simulation size is 10,000

We can see that the estimate using control group data can maintain Type I and Type II error rate more close to the target value, and this suggest to use,

$$s_k^2 = \frac{1}{mk - 1} \sum_{i=1}^{mk} (X_{Ci} - \bar{X}_C^k)^2. \quad (4.2)$$

Since the observed information depends on σ^2 , it will not be possible to guarantee both Type I and Type II error probabilities using pre-specified arm size. Following Jennison and Turnbull (1999), the priority is to maintain the Type I error rate close to its target value. The starting point is the stopping rule with critical values a_k and r_k . When σ^2 is known, the rejection region at stage k is $Z_k \geq r_k$, and the marginal rejection probability

is $1 - \Phi(r_k)$. In order to keep the same marginal rejection probability at stage k , the test static T_k should use a rejection boundary of $t_{mk-1,1-\Phi(r_k)}$. Making this change at each stage gives the stopping rule,

After stage $k = 1, \dots, K - 1$

If $T_k \geq t_{mk-1,1-\Phi(r_k)}$ stop, reject H_0

If $T_k \leq t_{mk-1,1-\Phi(a_k)}$ stop, accept H_0

Otherwise continue to stage $k + 1$

After stage K

If $T_K \geq t_{mk-1,1-\Phi(u)}$ stop, reject H_0

If $T_K \leq t_{mk-1,1-\Phi(u)}$ stop, accept H_0

where $t_{v,q}$ denotes the upper q tail-point of a t-distribution on v degrees of freedoms, i.e. $P(T > t_{v,q}) = q$ when $T \sim t_v$. When degrees of freedoms go to infinity, t-tests boundaries converge to z-test boundaries.

Table 4.2 below shows properties for group sequential t-tests for $K = 3, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 1$, and various specified values of θ and δ . The Type I error and Type II error are estimated by simulation, using 10000 replications. Similar to the pure shift setting, our empirical results show that group sequential t-tests can maintain the Type I error pretty accurately, with only a slight erosion in power.

θ	δ	Type I error rate (s.e.)	Type II error rate (s.e.)
1	1	0.0752 (0.0026)	0.2091 (0.0041)
0.75	1	0.0607 (0.0024)	0.1796 (0.0038)
0.5	1	0.0567 (0.0023)	0.1907 (0.0039)
0.25	1	0.0522 (0.0022)	0.1996 (0.0040)
1	0.5	0.0553 (0.0023)	0.1960 (0.0040)
0.75	0.5	0.0524 (0.0022)	0.2014 (0.0040)
0.5	0.5	0.0493 (0.0022)	0.2030 (0.0040)

Table 4.2: Properties of Group Sequential t-Tests: Type I and Type II error rate

4.2 Treatment Effect Estimation

In a mixture setting, the treatment effect is thought of as the pair (θ, δ) . We propose to use the maximum likelihood estimator (MLE) and method of moment estimator (MOM) for treatment effect (θ, δ) . Here we assume variance is unknown, and estimate σ^2 along with other three parameters μ_0 , θ , and δ .

4.2.1 MLE and Bias-Corrected MLE

In the pure shift setting, it is known that the post termination MLE of treatment effect is biased. Simulation results shown in Figure 4.1-Figure 4.6 indicate the MLE for (θ, δ) in the mixture setting are also biased. In order to reduce the bias of MLE, we considered using bootstrap bias correction (Wang & Leung, 1997). At the termination of the trial, we have maximum likelihood estimators $\widehat{\mu}_0, \widehat{\sigma}^2, \widehat{\theta}$, and $\widehat{\delta}$ obtained from the pooled control and treatment group responses. It should be noted that s_k^2 in equation (4.2)

is not used as an estimator for σ^2 here, and four parameters are optimized to maximize the log likelihood function. The bootstrap bias reduction algorithm is outlined below,

- 1) Generate B bootstrap group sequential sample paths via simulation using $\widehat{\mu}_0, \widehat{\sigma}^2, \widehat{\theta}$, and $\widehat{\delta}$. Each sample path will yield a post termination estimate of θ and δ , which are denoted as $\widehat{\theta}_j^*$ and $\widehat{\delta}_j^*, j = 1, \dots, B$;
- 2) The bootstrap bias estimates of θ and δ are,

$$\begin{aligned} \frac{\sum_{j=1}^B \widehat{\theta}_j^*}{B} - \widehat{\theta}, \\ \frac{\sum_{j=1}^B \widehat{\delta}_j^*}{B} - \widehat{\delta}; \end{aligned} \tag{4.3}$$

- 3) The bootstrap bias-corrected MLE estimates of θ and δ are

$$\begin{aligned} \widehat{\theta}_{bc} &= 2\widehat{\theta} - \frac{\sum_{j=1}^B \widehat{\theta}_j^*}{B}, \\ \widehat{\delta}_{bc} &= 2\widehat{\delta} - \frac{\sum_{j=1}^B \widehat{\delta}_j^*}{B}. \end{aligned} \tag{4.4}$$

For the following discussion, the group sequential tests are constructed using error spending function methods. We consider $K = 3, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 2, \mu_0 = 0$, and various specified values of (θ, δ) .

Following graphs compare the bias, variance, and mean squared error (MSE) of the MLE and the bias-corrected MLE of (θ, δ) . As it is shown, for both θ and δ , the

bootstrap bias-corrected MLE can reduce bias significantly if we compare it with MLE at the designed target values (except for case that designed theta equals 0.5, the bias of MLE for theta and adjusted MLE for theta are almost same). However, the increase in variance results in higher MSE. This is consistent with what people found in the non-parametric fixed sample test with mixture alternative setting, see (Jeske & Yao, in review). Therefore, we suggest using the MLE instead of the bias-corrected MLE.

Figure 4.1: Bias, variance and MSE of MLE and bootstrap bias-corrected MLE of (θ, δ) . $K = 3, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 2, \mu_0 = 0$, specified values of $(\theta, \delta) = (0.3, 0.8)$, $B = 1,000$, and 1,000 sample paths.

Figure 4.2: Bias, variance and MSE of MLE and bootstrap bias-corrected MLE of (θ, δ) . $K = 3, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 2, \mu_0 = 0$, specified values of $(\theta, \delta) = (0.5, 0.8)$, $B = 1,000$, and 1,000 sample paths.

Figure 4.3: Bias, variance and MSE of MLE and bootstrap bias-corrected MLE of (θ, δ) . $K = 3$, $\alpha = 0.05$, $\beta = 0.2$, $\sigma^2 = 1$, $\rho = 2$, $\mu_0 = 0$, specified values of $(\theta, \delta) = (0.8, 0.8)$, $B = 1,000$, and 1,000 sample paths.

Figure 4.4: Bias, variance and MSE of MLE and bootstrap bias-corrected MLE of (θ, δ) . $K = 3$, $\alpha = 0.05$, $\beta = 0.2$, $\sigma^2 = 1$, $\rho = 2$, $\mu_0 = 0$, specified values of $(\theta, \delta) = (0.3, 1)$, $B = 1,000$, and 1,000 sample paths.

Figure 4.5: Bias, variance and MSE of MLE and bootstrap bias-corrected MLE of (θ, δ) . $K = 3, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 2, \mu_0 = 0$, specified values of $(\theta, \delta) = (0.5, 1)$, $B = 1,000$, and 1,000 sample paths.

Figure 4.6: Bias, variance and MSE of MLE and bootstrap bias-corrected MLE of (θ, δ) . $K = 3, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 2, \mu_0 = 0$, specified values of $(\theta, \delta) = (0.8, 1)$, $B = 1,000$, and 1,000 sample paths.

4.2.2 MOM and Bias-Corrected MOM

Besides MLE, we also consider method of moment (MOM) estimator for treatment effect.

Based on the control and treatment group data, MOM estimator is given by,

$$\left\{ \begin{array}{l} \mu_F = \mu_0 \\ \sigma_F^2 = \sigma^2 \\ \mu_G = \mu_0 + \theta\delta \\ \sigma_F^2 = \sigma^2 + \theta(1 - \theta)\delta \end{array} \right. \Rightarrow \left\{ \begin{array}{l} \widehat{\mu}_0 = \overline{X}_C \\ \widehat{\sigma}^2 = s_C^2 \\ \hat{\theta} = \left\{ 1 + \frac{(s_T^2 - s_C^2)_+}{(\overline{X}_T - \overline{X}_C)_+^2 + \varepsilon} \right\}^{-1} \\ \hat{\delta} = (\overline{X}_T - \overline{X}_C)_+ \left\{ 1 + \frac{(s_T^2 - s_C^2)_+}{(\overline{X}_T - \overline{X}_C)_+^2 + \varepsilon} \right\} \end{array} \right. \quad (4.5)$$

where $t_+ = t$ if $t > 0$ and 0 otherwise. $(\overline{X}_C, s_C^2)$ are the mean and variance of the control group observations, $(\overline{X}_T, s_T^2)$ are the same and variance for the treatment group observations, and ε is a small positive number that bounds the denominators away from zero.

For the following examples, we consider $K = 3, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 2, \mu_0 = 0, \varepsilon = 0.001$, and various specified values of (θ, δ) . The bootstrap bias reduction method using $B = 1000$ replications, and the simulation size 1000. As it is shown in the graphs, for both θ and δ , the bootstrap bias-corrected MOM can reduce bias significantly if we compare it with MOM at the designed target values (except for case that designed θ equals 0.5 and δ equals 0.8, the bias of MOM for theta and adjusted MOM for δ are almost same). However, the variance for the bias-corrected MOM is larger than MOM. As a result, the MSE for the bias-corrected MOM is larger than MOM for the designed treatment effect. This is consistent with results we see in previous section about MLE. Bias-corrected estimator can reduce bias, however, MSE gets larger. As it is shown in Table 4.3, for most of the cases, MLE has better performance than MOM in terms of bias, and MSE. Therefore, across all four treatment effect estimators, we suggest to use MLE.

Figure 4.7: Bias, variance and MSE of MOM and bootstrap bias-corrected MOM of (θ, δ) . $K = 3, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 2, \mu_0 = 0$, specified values of $(\theta, \delta) = (0.3, 0.8)$, $B = 1,000$, and 1,000 sample paths.

Figure 4.8: Bias, variance and MSE of MOM and bootstrap bias-corrected MOM of (θ, δ) . $K = 3, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 2, \mu_0 = 0$, specified values of $(\theta, \delta) = (0.5, 0.8)$, $B = 1,000$, and 1,000 sample paths.

Figure 4.9: Bias, variance and MSE of MOM and bootstrap bias-corrected MOM of (θ, δ) . $K = 3, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 2, \mu_0 = 0$, specified values of $(\theta, \delta) = (0.8, 0.8)$, $B = 1,000$, and 1,000 sample paths.

Figure 4.10: Bias, variance and MSE of MOM and bootstrap bias-corrected MOM of (θ, δ) . $K = 3, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 2, \mu_0 = 0$, specified values of $(\theta, \delta) = (0.3, 1)$, $B = 1,000$, and 1,000 sample paths.

Figure 4.11: Bias, variance and MSE of MOM and bootstrap bias-corrected MOM of (θ, δ) . $K = 3, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 2, \mu_0 = 0$, specified values of $(\theta, \delta) = (0.5, 1)$, $B = 1,000$, and 1,000 sample paths.

Figure 4.12: Bias, variance and MSE of MOM and bootstrap bias-corrected MOM of (θ, δ) . $K = 3, \alpha = 0.05, \beta = 0.2, \sigma^2 = 1, \rho = 2, \mu_0 = 0$, specified values of $(\theta, \delta) = (0.8, 1)$, $B = 1,000$, and 1,000 sample paths.

design target		θ					
δ	θ	Bias		Variance		MSE	
		MLE	MLE_bc	MLE	MLE_bc	MLE	MLE_bc
1	0.5	0.0750	0.0647	0.0989	0.1743	0.1045	0.1785
1	0.8	-0.0283	0.0066	0.0754	0.1375	0.0762	0.1375
1	0.9	-0.1008	-0.0688	0.0729	0.1295	0.0831	0.1342
δ	θ	Bias		Variance		MSE	
		MOM	MOM_bc	MOM	MOM_bc	MOM	MOM_bc
1	0.5	-0.2265	-0.1513	0.0834	0.2201	0.1348	0.2429
1	0.8	-0.1793	-0.1073	0.0687	0.1843	0.1008	0.1958
1	0.9	-0.1602	-0.0885	0.0560	0.1449	0.0817	0.1528
design target		δ					
δ	θ	Bias		Variance		MSE	
		MLE	MLE_bc	MLE	MLE_bc	MLE	MLE_bc
1	0.5	0.1251	-0.0167	0.2946	0.4016	0.3103	0.4018
1	0.8	0.2564	0.0692	0.3149	0.4243	0.3807	0.4291
1	0.9	0.3165	0.1316	0.3434	0.4520	0.4436	0.4693
δ	θ	Bias		Variance		MSE	
		MOM	MOM_bc	MOM	MOM_bc	MOM	MOM_bc
1	0.5	0.3489	0.0942	0.3400	0.8017	0.4617	0.8106
1	0.8	0.4015	0.0927	0.6690	1.8214	0.8302	1.8300
1	0.9	0.3940	0.0626	0.4814	1.0860	0.6366	1.0899

Table 4.3: Bias, variance and MSE comparisons between MLE and MOM

Chapter 5

Conclusions

5.1 Conclusion

For traditional group sequential designs, we assume that both the control and treatment group responses come from normal distributions that potentially have different means. And we power the design to detect a pure mean shift. In this dissertation, we argue that this alternative is often unrealistic. Instead, we consider an alternative where only a percentage of treated patients have a higher likelihood of benefiting from a treatment, and propose the use of a mixture alternative as a way to capture this realism. The required arm size to detect a specified mean shift of δ for a percentage of patients is larger than that what is required to detect the mean shift of δ for all treated patients. The stopping boundaries depend on the

specified θ and δ/σ , the dependence is weak. However, the arm size depends significantly on both specified θ and δ/σ .

We also have discussed adjustment for the unknown σ^2 case, and introduced the group sequential t-tests. For treatment effect estimation, we have proposed MLE, MOM, bias-corrected MLE, and bias-corrected MOM. Bias-corrected MLE has the smallest bias, and MLE has the smallest MSE. Therefore, we suggest using the MLE instead of the bias-corrected MLE.

5.2 Future Work

For the post termination analysis, we only focused on the point estimation for treatment effect. Further, we could look at confidence interval for treatment effect and p-value for the test.

We assume the normal distribution for the control group observations, and a mixture of normal distributions for the treatment group observations. We can extend it to group sequential designs in non-parametric setting.

We can extend the group sequential designs to detect a mixture alternative for two-sided tests. For one-sided tests, we power the design for an alternative where the treatment group observations come from a mixture of normal distributions, and only one proportion parameter is used. For two-sided tests, we can consider the case that a subset of the

treatment group responses have higher probability to benefit from the treatment, and a subset of the treatment group patients have lower probability to benefit from the treatment than control. In other words, two-sided tests could be constructed using the mixture model with two proportion parameters.

Bibliography

- Anderson, K. M. (2009). *gsDesign: An R Package for Designing Group Sequential Clinical Trials Version 2.0 Manual*. Retrieved from https://r-forge.r-project.org/scm/viewvc.php/*checkout*/pkg/gsDesign/inst/doc/gsDesignManual.pdf?revision=183&root=gsdesign
- Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A*, 235-244.
- Chang, M. N., Hwang, I. K., & Shin, W. J. (1998). Group sequential designs using both type I and type II error probability spending functions. *Communications in Statistics - Theory and Methods*, 1323-1339.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale: Erlbaum Associates.
- Cytel Inc. (2010). *EAST Advanced Clinical Trial Design, Simulation and Monitoring*. Cambridge, MA: Cytel Inc.
- DeMets, D. L., & Ware, J. H. (1980). Group Sequential Methods for Clinical Trials with a One-Sided Hypothesis. *Biometrika*, 651-660.
- Emerson, S. S. (1988). *Parameter Estimation Following Group Sequential Hypothesis Testing*. Seattle: University of Washington.
- Emerson, S. S., & Fleming, T. R. (1990). Parameter Estimation Following Group Sequential Hypothesis Testing. *Biometrika*, 875-892.

- FDA. (2013). *Paving the Way for Personalized Medicine*. Retrieved from <https://www.fda.gov/downloads/scienceresearch/specialtopics/personalizedmedicine/ucm372421.pdf>.
- Haybittle, J. L. (1971). Repeated assessment of results in clinical trials of cancer treatment. *Br J Radiol*, 44, 793-797.
- Hwang, I. K., Shih, W. J., & De Cani, J. S. (1990). Group Sequential Designs Using a Family of Type I error Probability Spending Functions. *Statistics in Medicine*, 9, 1439-1445.
- Hwang, I. K., Shih, W. J., & De Cani, J. S. (1990). Group sequential design using a family of type I error probability spending functions. *Statistics in Medicine*, 1439-1445.
- Jennison, C., & Turnbull, B. W. (1999). *Group Sequential Methods with Applications to Clinical Trials*. London: CRC Press.
- Jeske, D. R., & Yao, W. (in review). Sample Size Calculations for Mixture Alternatives in a Control Group vs. Treatment Group Design. *Statistical Methods in Medical Research*.
- Lan, K. K., & DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, 70, 659-663.
- O'Brien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, 35, 549-556.
- Pampallona, S., & Tsiatis, A. A. (1994). Group Sequential Designs for One-Sided and Two-Sided Hypothesis Testing with Provision for Early Stopping in Favor of the Null Hypothesis. *Journal of Statistical Planning and Inference*, 19-35.
- Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S. V., . . . Smith, P. G. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and Design. *Brit. J. of Cancer*, 34, 585-612.

- Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64, 191-199.
- SAS Institute Inc. (2009). *SAS/STAT® 9.2 User's Guide*. Cary, NC: SAS Institute Inc.
- SenGupta, A. (2006). P3 approach to Intersection-Union testing of hypotheses. *S.N. Roy Birth Centenary volume, J. Statist. Plann. Inf.*
- Slud, E., & Wei, L. J. (1982). Two-Sample Repeated Significance Tests. *Journal of the American Statistical Association*, 862-868.
- Spear, B. B., Heath-Chiozzi, M., & Huff, J. (2001). Clinical application of pharmacogenetics. *Trends in Molecular Medicine*, 201-204.
- Spjøtvoll, E. (1968). Most powerful test for some non-exponential families. *Annals of Mathematical Statistics*, 39, 772-784.
- Wang, S. K., & Tsiatis, A. A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics*, 43, 193-200.
- Wang, S.-J., O'Neill, T. R., & Hung, H. J. (2007). Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharmaceutical Statistics*, 6, 227-244.
- Wang, Y., & Leung, D. (1997). Bias reduction via resampling for estimation following sequential tests. *Sequential Analysis*, 16, 249-267.
- Whitehead, J. (1986). On the Bias of Maximum Likelihood Estimation Following a Sequential Test. *Biometrika*, 573-581.
- Zhu, L., Ni, L., & Yao, B. (2011). Group Sequential Methods and Software Application. *The American Statistician*, 127-135.

Appendix

Appendix A Sub-density

The distribution of the sequence of test statics $\{Z_1, \dots, Z_K\}$ are approximately multivariate normal with mean and covariance,

$$\begin{aligned} E(Z_k) &= \theta\delta \sqrt{\frac{mk}{2\sigma^2}}, \quad k = 1, \dots, K \\ \text{cov}(Z_{k_1}, Z_{k_2}) &= \left(1 + \frac{\theta(1-\theta)\delta^2}{2\sigma^2}\right) \sqrt{k_1/k_2}, \quad 1 \leq k_1 \leq k_2 \leq K. \end{aligned} \tag{A.1}$$

This allows use of the recursive formula of Armitage, McPherson and Rowe (1969).

Specifically, for $k = 2, \dots, K$,

$$Z_k \sqrt{\frac{mk}{2\sigma^2}} - Z_{k-1} \sqrt{\frac{m(k-1)}{2\sigma^2}} \sim N\left(\theta\delta \frac{m}{2\sigma^2}, \left(1 + \frac{\theta(1-\theta)\delta^2}{2\sigma^2}\right) \frac{m}{2\sigma^2}\right), \tag{A.2}$$

and this distribution is independent of $Z_1, \dots, Z_{k-1}$. Defining

$$G_k(z; \theta, \delta) = P_{\theta, \delta} \{a_1 < Z_1 < r_1, \dots, a_{k-1} < Z_{k-1} < r_{k-1}, Z_k \geq z\}, \quad (\text{A.3})$$

and $g_k(z; \theta, \delta)$ to be the derivative of $G_k(z; \theta, \delta)$ with respect to z , this is called sub-density of Z_k at stage k . Let $\phi(z)$ denote the standard normal density, for $k = 2, \dots, K$.

Then g_k is given recursively by

$$\begin{aligned} g_1(z; \theta, \delta) &= \phi\left(z - \theta\delta\sqrt{\frac{m}{2\sigma^2}}\right), \\ g_k(z; \theta, \delta) &= \int_{a_{k-1}}^{r_{k-1}} g_{k-1}(u; \theta, \delta) \frac{\sqrt{\frac{mk}{2\sigma^2}}}{\sqrt{\left(1 + \frac{\theta(1-\theta)\delta^2}{2\sigma^2}\right) \frac{m}{2\sigma^2}}} \phi\left(\frac{z\sqrt{\frac{mk}{2\sigma^2}} - u\sqrt{\frac{m(k-1)}{2\sigma^2}} - \theta\delta\frac{m}{2\sigma^2}}{\sqrt{\left(1 + \frac{\theta(1-\theta)\delta^2}{2\sigma^2}\right) \frac{m}{2\sigma^2}}}\right) du \end{aligned} \quad (\text{A.4})$$

The stopping boundaries (a_k, r_k) , $k = 1, 2, \dots, K-1$, and u are calculated in sequence.

When (a_i, r_i) , $i = 1, 2, \dots, k-1$ are known, the above formula can be used to evaluate g_k numerically by a succession of $k-1$ univariate integrals, and hence for the specified values of (θ, δ) , one can solve

$$\begin{aligned} \pi_{1,k} &= \int_{r_k}^{\infty} g_k(z_k; \theta, \delta | \theta\delta = 0) dz_k \\ \pi_{2,k} &= \int_{-\infty}^{a_k} g_k(z_k; \theta, \delta) dz_k \end{aligned} \quad (\text{A.5})$$

to obtain r_k , and a_k for the group sequential design with mixture alternative.

Appendix B Ratio and Stopping Boundaries

The variance covariance matrix is,

$$\left(1 + \frac{\theta(1-\theta)\delta^2}{2\sigma^2}\right) \begin{pmatrix} 1 & \sqrt{1/2} & \dots & \sqrt{1/K} \\ \sqrt{1/2} & 1 & \dots & \sqrt{2/K} \\ \dots & \dots & \dots & \dots \\ \sqrt{1/K} & \sqrt{2/K} & \dots & 1 \end{pmatrix}. \quad (\text{B.1})$$

Cohen (1988) suggested that 0.2 be considered a 'small' effect size, 0.5 represents a 'medium' effect size and 0.8 a 'large' effect size. If we assume δ/σ is 1, then the multiplier in the variance covariance matrix satisfies,

$$1 \leq 1 + \frac{\theta(1-\theta)\delta^2}{2\sigma^2} \leq 1 + \frac{0.25}{2} = 1.125 \quad (\text{B.2})$$

As such, the choices of the specified θ and δ/σ do not have a big impact for the variance covariance matrix of the test statistics $\{Z_1, \dots, Z_K\}$, and

$$\begin{pmatrix} Z_1 \\ Z_2 \\ \dots \\ Z_K \end{pmatrix} \sim MVN \left(\begin{pmatrix} \theta\delta\sqrt{m/(2\sigma^2)} \\ \theta\delta\sqrt{2m/(2\sigma^2)} \\ \dots \\ \theta\delta\sqrt{Km/(2\sigma^2)} \end{pmatrix}, 1.125 \begin{pmatrix} 1 & \sqrt{1/2} & \dots & \sqrt{1/K} \\ \sqrt{1/2} & 1 & \dots & \sqrt{2/K} \\ \dots & \dots & \dots & \dots \\ \sqrt{1/K} & \sqrt{2/K} & \dots & 1 \end{pmatrix} \right). \quad (\text{B.3})$$

Under $H_0: F = G$,

$$\begin{pmatrix} Z_1 \\ Z_2 \\ \dots \\ Z_K \end{pmatrix} \sim MVN \left(\begin{pmatrix} 0 \\ 0 \\ \dots \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sqrt{1/2} & \dots & \sqrt{1/K} \\ \sqrt{1/2} & 1 & \dots & \sqrt{2/K} \\ \dots & \dots & \dots & \dots \\ \sqrt{1/K} & \sqrt{2/K} & \dots & 1 \end{pmatrix} \right), \quad (\text{B.4})$$

Under the specified value of θ and δ , Type II error probabilities become,

$$P_{\theta,\delta}\{a_1 < Z_1 < r_1, \dots, a_{k-1} < Z_{k-1} < r_{k-1}, Z_k \leq a_k\} \\ \approx P_{\theta,\delta} \left\{ \frac{a_1 - \theta\delta\sqrt{\frac{1}{K}}\sqrt{\frac{mK}{2\sigma^2}}}{\sqrt{1.125}} < Z_1^* < \frac{r_1 - \theta\delta\sqrt{\frac{1}{K}}\sqrt{\frac{mK}{2\sigma^2}}}{\sqrt{1.125}}, \dots, Z_k^* \leq \frac{a_k - \theta\delta\sqrt{\frac{k}{K}}\sqrt{\frac{mK}{2\sigma^2}}}{\sqrt{1.125}} \right\} \quad (\text{B.5})$$

where $Z_i^* = \frac{1}{\sqrt{1.125}} \left\{ Z_i - \theta\delta\sqrt{\frac{i}{K}}\sqrt{\frac{mK}{2\sigma^2}} \right\}$ for $i = 1, \dots, k$, and Z_i^* is standard normal at the

specified values of θ and δ . Solving for m is equivalent to solving first for $\lambda = \theta\delta\sqrt{\frac{mK}{2\sigma^2}}$

in (B.5) and then computing $m = 2\lambda^2\sigma^2/(\theta^2\delta^2K)$.

$$P_{H_0}\{Z_1 \geq r_1\} = \pi_{1,1} \\ P_{\theta,\delta} \left\{ Z_1^* \leq \frac{a_1 - \lambda\sqrt{1/K}}{\sqrt{1.125}} \right\} = \pi_{2,1} \\ P_{H_0}\{a_1 < Z_1 < r_1, \dots, Z_k \geq r_k\} = \pi_{1,k}, k = 2, \dots, K-1 \\ P_{\theta,\delta} \left\{ \frac{a_1 - \lambda\sqrt{1/K}}{\sqrt{1.125}} < Z_1^* < \frac{r_1 - \lambda\sqrt{1/K}}{\sqrt{1.125}}, \dots, Z_k^* \leq \frac{a_k - \lambda\sqrt{k/K}}{\sqrt{1.125}} \right\} = \pi_{2,k}, k = \\ 2, \dots, K-1 \\ P_{\theta=0}\{a_1 < Z_1 < r_1, \dots, Z_K \geq u\} = \pi_{1,K} \quad (\text{B.6})$$

$$P_{\theta,\delta} \left\{ \frac{a_1 - \lambda\sqrt{1/K}}{\sqrt{1.125}} < Z_1^* < \frac{r_1 - \lambda\sqrt{1/K}}{\sqrt{1.125}}, \dots, Z_K^* < \frac{u - \lambda}{\sqrt{1.125}} \right\} = \pi_{2,K}.$$

Since distribution of $\{Z_1, \dots, Z_K\}$ under the null hypothesis and $\{Z_1^*, \dots, Z_K^*\}$ under

the mixture alternative are both approximately $MVN \left(\begin{pmatrix} 0 \\ 0 \\ \dots \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sqrt{1/2} & \dots & \sqrt{1/K} \\ \sqrt{1/2} & 1 & \dots & \sqrt{2/K} \\ \dots & \dots & \dots & \dots \\ \sqrt{1/K} & \sqrt{2/K} & \dots & 1 \end{pmatrix} \right)$, the

solution to the stopping boundaries and λ only depend on K, α, β , and ρ . Once λ is solved for, we can determine the maximum sample size per arm, $mK = 2\lambda^2\sigma^2/(\theta^2\delta^2)$, which is proportional to $\sigma^2/(\theta^2\delta^2)$. For the fixed sample test in the mixture setting, the arm size approximation formula is shown (3.3). Again using (B.2), the arm size will be approximately proportional to $\sigma^2/(\theta^2\delta^2)$. It follows that the ratio of the maximum arm size for the group sequential test to the arm size for the fixed sample test primarily only depends on K, α, β , and ρ .

To illustrate the results, Table C.1-Table C.3 show the ratio of the maximum arm size for the group sequential test to the arm size for the fixed sample test for the case $\alpha = 0.05, \beta = 0.2$, and $\rho = 1, 2, 3$. What can be seen in below tables is that unless δ/σ is large, the values in the table only depend on K .

θ	K	Treatment Effect Shift Size, δ/σ						
		0.25	0.5	0.75	1	1.5	2	2.5
1	2	1.145	1.145	1.145	1.145	1.145	1.145	1.145
	3	1.207	1.207	1.207	1.207	1.207	1.207	1.207
	4	1.242	1.242	1.242	1.242	1.242	1.242	1.242
	5	1.264	1.264	1.264	1.264	1.264	1.264	1.264
0.9	2	1.145	1.145	1.145	1.145	1.144	1.144	1.142
	3	1.207	1.207	1.207	1.207	1.206	1.205	1.201
	4	1.241	1.241	1.241	1.241	1.240	1.238	1.234
	5	1.264	1.264	1.264	1.263	1.262	1.260	1.255
0.8	2	1.145	1.145	1.145	1.145	1.144	1.142	1.138
	3	1.207	1.207	1.207	1.206	1.205	1.200	1.193
	4	1.241	1.242	1.241	1.241	1.238	1.232	1.223
	5	1.264	1.264	1.263	1.263	1.260	1.253	1.241
0.7	2	1.145	1.145	1.145	1.145	1.143	1.140	1.134
	3	1.207	1.207	1.207	1.206	1.203	1.196	1.186
	4	1.242	1.241	1.241	1.240	1.236	1.227	1.213
	5	1.264	1.264	1.263	1.262	1.257	1.247	1.229
0.6	2	1.145	1.145	1.145	1.144	1.142	1.139	1.132
	3	1.207	1.207	1.206	1.206	1.202	1.194	1.182
	4	1.241	1.242	1.241	1.240	1.234	1.224	1.207
	5	1.264	1.264	1.263	1.262	1.255	1.242	1.221
0.5	2	1.145	1.145	1.145	1.145	1.142	1.138	1.131
	3	1.207	1.207	1.206	1.206	1.201	1.193	1.180
	4	1.241	1.241	1.241	1.240	1.234	1.223	1.205
	5	1.264	1.264	1.263	1.261	1.255	1.241	1.219

Table B.1: Ratio of the maximum arm size for the group sequential test to the arm size for the fixed sample test for 5% level and will achieve 80% power for the mixture alternative with $\rho = 1$

θ	K	Treatment Effect Shift Size, δ/σ						
		0.25	0.5	0.75	1	1.5	2	2.5
1	2	1.043	1.043	1.043	1.043	1.043	1.043	1.043
	3	1.070	1.070	1.070	1.070	1.070	1.070	1.070
	4	1.088	1.088	1.088	1.088	1.088	1.088	1.088
	5	1.099	1.099	1.099	1.099	1.099	1.099	1.099
0.9	2	1.043	1.043	1.043	1.043	1.043	1.042	1.040
	3	1.070	1.070	1.071	1.070	1.069	1.067	1.064
	4	1.088	1.087	1.087	1.087	1.086	1.084	1.079
	5	1.099	1.098	1.099	1.098	1.096	1.094	1.089
0.8	2	1.043	1.044	1.043	1.043	1.042	1.039	1.035
	3	1.070	1.070	1.070	1.069	1.067	1.063	1.056
	4	1.087	1.087	1.087	1.086	1.084	1.078	1.069
	5	1.099	1.099	1.098	1.097	1.094	1.087	1.077
0.7	2	1.043	1.043	1.043	1.043	1.040	1.036	1.030
	3	1.071	1.07	1.069	1.069	1.065	1.059	1.049
	4	1.087	1.087	1.087	1.085	1.081	1.073	1.060
	5	1.099	1.099	1.098	1.096	1.091	1.082	1.067
0.6	2	1.043	1.043	1.043	1.042	1.040	1.035	1.028
	3	1.071	1.070	1.070	1.068	1.065	1.057	1.045
	4	1.087	1.087	1.086	1.085	1.080	1.070	1.055
	5	1.099	1.098	1.097	1.096	1.090	1.078	1.060
0.5	2	1.043	1.043	1.043	1.043	1.040	1.035	1.027
	3	1.071	1.070	1.070	1.069	1.064	1.056	1.043
	4	1.087	1.087	1.086	1.085	1.079	1.069	1.053
	5	1.098	1.099	1.098	1.096	1.089	1.077	1.058

Table B.2: Ratio of the maximum arm size for the group sequential test to the arm size for the fixed sample test for 5% level and will achieve 80% power for the mixture alternative with $\rho = 2$

θ	K	Treatment Effect Shift Size, δ/σ						
		0.25	0.5	0.75	1	1.5	2	2.5
1	2	1.014	1.014	1.014	1.014	1.014	1.014	1.014
	3	1.028	1.028	1.028	1.028	1.028	1.028	1.028
	4	1.038	1.038	1.038	1.038	1.038	1.038	1.038
	5	1.046	1.046	1.046	1.046	1.046	1.046	1.046
0.9	2	1.014	1.014	1.014	1.014	1.013	1.012	1.011
	3	1.028	1.028	1.028	1.028	1.027	1.025	1.023
	4	1.038	1.038	1.038	1.038	1.037	1.035	1.031
	5	1.046	1.046	1.045	1.045	1.044	1.041	1.038
0.8	2	1.014	1.014	1.014	1.014	1.012	1.010	1.007
	3	1.028	1.028	1.028	1.027	1.025	1.022	1.016
	4	1.038	1.038	1.038	1.037	1.035	1.031	1.023
	5	1.046	1.046	1.045	1.044	1.041	1.036	1.028
0.7	2	1.014	1.014	1.014	1.013	1.012	1.009	1.004
	3	1.028	1.028	1.028	1.027	1.024	1.019	1.011
	4	1.038	1.038	1.037	1.036	1.033	1.027	1.016
	5	1.045	1.045	1.045	1.044	1.039	1.032	1.020
0.6	2	1.014	1.014	1.014	1.014	1.011	1.007	1.001
	3	1.028	1.028	1.028	1.027	1.023	1.017	1.007
	4	1.038	1.038	1.037	1.036	1.032	1.024	1.012
	5	1.045	1.046	1.044	1.043	1.038	1.029	1.014
0.5	2	1.014	1.014	1.014	1.014	1.011	1.007	1.001
	3	1.028	1.028	1.028	1.026	1.023	1.016	1.006
	4	1.038	1.038	1.037	1.036	1.031	1.023	1.010
	5	1.046	1.046	1.045	1.043	1.038	1.028	1.013

Table B.3: Ratio of the maximum arm size for the group sequential test to the arm size for the fixed sample test for 5% level and will achieve 80% power for the mixture alternative with $\rho = 3$

Appendix C The R Code for the Stopping Boundaries

C.1. Manual

This is the guide to function **GSDMix()** written in R to implement the Group Sequential Design to detect a Mixture alternative.

GSDMix *Determining stopping boundaries and arm size per stage for Group Sequential Design to detect a mixture alternative*

Description

GSDMix() is used to determine stopping boundaries and trial size required for a group sequential design in mixture setting.

Usage

GSDMix(K,alpha,beta,rho,theta,effectsize)

Arguments

K	maximum number of analyses planned, including interim and final.
alpha	Type I error rate for the design
beta	Type II error rate for the design
rho	exponent in error spending functions

theta	the specified mixing proportion parameter in the mixture distribution
effectsize	the specified standardized mean shift parameter, which is δ/σ

Details

GSDMix() is used for one-sided group sequential design in mixture setting. The stopping boundaries are early stopping to reject for efficacy or early stopping to accept for futility. The input *rho* is the power exponent for Type I and Type II error spending functions. The group sequential design in mixture setting will be reduced to group sequential design in pure shift setting if choosing *theta*=1.

Value

lower	boundaries of acceptance for the group sequential design with mixture alternative
upper	boundaries of rejection for the group sequential design with mixture alternative
arm_size	sample size per stage per arm

Examples

```
GSDMix(3,0.05,0.2,2,0.8,0.5)
## $lower
## [1] -0.5332111  0.7047889  1.7031848
##
## $upper
## [1] 2.539185 2.068185 1.703185
##
```

```
## $arm_size
## [1] 28
```

C.2. R code

```
GSDMix=function(K,alpha,beta,rho,theta,effectsize){

  powerdiff=function(m){
    pi1=vector("numeric",K)
    pi2=vector("numeric",K)
    b=vector("numeric",2*K+1)
    a=vector("numeric",K)
    r=vector("numeric",K)
    t=vector("numeric",K)
    In=vector("numeric",K)

    for (i in 1:K){
      t[i]=i/K
    }

    pi1[1]=alpha*(t[1])^rho
    pi2[1]=beta*(t[1])^rho

    for (i in 2:K){
      pi1[i]=alpha*(t[i])^rho-alpha*(t[i-1])^rho
      pi2[i]=beta*(t[i])^rho-beta*(t[i-1])^rho
    }

    #####
    ratiomean=theta
    ratiovariance=1+theta*(1-theta)*effectsize^2/2

    for (i in 1:K){
      In[i]=m*i/2
    }

    r[1]=qnorm(1-pi1[1])
    a[1]=ratiomean*effectsize*sqrt(In[1])+sqrt(ratiovariance)*qnorm(pi2
[1])

    n=320
    h1=matrix(numeric(0),K,n+1)
    h2=matrix(numeric(0),K,n+1)
    z=matrix(numeric(0),K,n+1)
    w=vector("numeric",n+1)
```

```

f=vector("numeric",K)
g=vector("numeric",K)

g[1]=pi2[1]

for( i in 2:K){
  ### weight for Trapezoidal rule
  w[1]=(r[i-1]-a[i-1])/(2*n)
  w[n+1]=(r[i-1]-a[i-1])/(2*n)
  for (j in 2:n) {
    w[j]=(r[i-1]-a[i-1])/n
  }
  h=(r[i-1]-a[i-1])/n

  for (j in 1:(n+1)){
    z[i,j]=a[i-1]+(j-1)*h
  }
  if(i==2){
    for (j in 1:(n+1)) {
      h1[i,j]=w[j]*dnorm(z[i,j])
      h2[i,j]=w[j]*dnorm((z[i,j]-ratiomean*effectsize*sqrt(In[i-1]))/sqrt(ratiovariance))/sqrt(ratiovariance)
    }
  }else{
    for (j in 1:(n+1)) {
      h1[i,j]=0
      h2[i,j]=0
      for(l in 1:(n+1)){
        h1[i,j]=h1[i,j]+h1[i-1,l]*w[j]*sqrt(In[i-1]/(In[i-1]-In[i-2]))*dnorm((z[i,j]*sqrt(In[i-1])-z[i-1,l]*sqrt(In[i-2]))/sqrt(In[i-1]-In[i-2]))
        h2[i,j]=h2[i,j]+h2[i-1,l]*w[j]*sqrt(In[i-1]/((In[i-1]-In[i-2])*ratiovariance))*dnorm((z[i,j]*sqrt(In[i-1])-z[i-1,l]*sqrt(In[i-2]))/sqrt((In[i-1]-In[i-2])*ratiovariance)-ratiomean*effectsize*sqrt((In[i-1]-In[i-2])/ratiovariance))
      }
    }
  }
  d=-0.001
  r[i]=r[i-1]
  f[i]=0
  for (j in 1:(n+1)) {
    f[i]=f[i]+h1[i,j]*pnorm((z[i,j]*sqrt(In[i-1])-r[i]*sqrt(In[i]))/sqrt(In[i]-In[i-1]))
  }
}

```

```

e1=(f[i]-pi1[i])^2

r[i]=r[i]+d
f[i]=0
for (j in 1:(n+1)) {
  f[i]=f[i]+h1[i,j]*pnorm((z[i,j]*sqrt(In[i-1])-r[i]*sqrt(In[i]))
/sqrt(In[i]-In[i-1]))
}
e2=(f[i]-pi1[i])^2

if (e1<e2) {
  d=-d
}

e1=0.9
e2=1

while (e1<e2){
  r[i]=r[i]+d
  e2=e1
  f[i]=0
  for (j in 1:(n+1)) {
    f[i]=f[i]+h1[i,j]*pnorm((z[i,j]*sqrt(In[i-1])-r[i]*sqrt(In[i]
))/sqrt(In[i]-In[i-1]))
  }
  e1=(f[i]-pi1[i])^2
}
r[i]=r[i]-d

if(i==K){
  a[i]=r[i]
  g[i]=0
  for (j in 1:(n+1)) {
    g[i]=g[i]+h2[i,j]*pnorm((-z[i,j]*sqrt(In[i-1])+a[i]*sqrt(In[i]
))-ratiomean*effectsize*(In[i]-In[i-1]))/sqrt((In[i]-In[i-1])*ratiovari
ance))
  }
  pK=g[i]
}else{
  d=0.001
  a[i]=a[i-1]
  g[i]=0
  for (j in 1:(n+1)) {
    g[i]=g[i]+h2[i,j]*pnorm((-z[i,j]*sqrt(In[i-1])+a[i]*sqrt(In[i]
))-ratiomean*effectsize*(In[i]-In[i-1]))/sqrt(In[i]-In[i-1]))
  }
}

```

```

    e1=(g[i]-pi2[i])^2

    a[i]=a[i]+d
    g[i]=0
    for (j in 1:(n+1)) {
        g[i]=g[i]+h2[i,j]*pnorm((-z[i,j]*sqrt(In[i-1])+a[i]*sqrt(In[i]
))-ratiomean*effectsize*(In[i]-In[i-1]))/sqrt((In[i]-In[i-1])*ratiovari
ance))
    }
    e2=(g[i]-pi2[i])^2

    if (e1<e2) {
        d=-d
    }

    e1=0.9
    e2=1

    while (e1<e2){
        a[i]=a[i]+d
        e2=e1
        g[i]=0
        for (j in 1:(n+1)) {
            g[i]=g[i]+h2[i,j]*pnorm((-z[i,j]*sqrt(In[i-1])+a[i]*sqrt(In
[i]))-ratiomean*effectsize*(In[i]-In[i-1]))/sqrt((In[i]-In[i-1])*ratioa
riance))
        }
        e1=(g[i]-pi2[i])^2
    }
    a[i]=a[i]-d
}
}
1-sum(g)-(1-beta)
}
n=2*(qnorm(1-alpha)+qnorm(1-beta)*sqrt(1+theta*(1-theta)*effectsize^2
/2))^2/(theta*effectsize)^2
l=max(n/K,2)
u=n/K*2
m=uniroot(powerdiff,lower=l,upper=u)$root
m=ceiling(m)

ratiomean=theta
ratiovariance=1+theta*(1-theta)*effectsize^2/2

pi1=vector("numeric",K)
pi2=vector("numeric",K)

```

```

b=vector("numeric",2*K+1)
a=vector("numeric",K)
r=vector("numeric",K)
t=vector("numeric",K)
In=vector("numeric",K)

for (i in 1:K){
  t[i]=i/K
}

pi1[1]=alpha*(t[1])^rho
pi2[1]=beta*(t[1])^rho

for (i in 2:K){
  pi1[i]=alpha*(t[i])^rho-alpha*(t[i-1])^rho
  pi2[i]=beta*(t[i])^rho-beta*(t[i-1])^rho
}
for (i in 1:K){
  In[i]=m*i/2
}

r[1]=qnorm(1-pi1[1])
a[1]=ratiomean*effectsize*sqrt(In[1])+sqrt(ratiovariance)*qnorm(pi2[1
])

n=320
h1=matrix(numeric(0),K,n+1)
h2=matrix(numeric(0),K,n+1)
z=matrix(numeric(0),K,n+1)
w=vector("numeric",n+1)
f=vector("numeric",K)
g=vector("numeric",K)

for( i in 2:K){
  ### weight for Trapezoidal rule
  w[1]=(r[i-1]-a[i-1])/(2*n)
  w[n+1]=(r[i-1]-a[i-1])/(2*n)
  for (j in 2:n) {
    w[j]=(r[i-1]-a[i-1])/n
  }
  h=(r[i-1]-a[i-1])/n

  for (j in 1:(n+1)){
    z[i,j]=a[i-1]+(j-1)*h
  }
  if(i==2){
    for (j in 1:(n+1)) {

```

```

        h1[i,j]=w[j]*dnorm(z[i,j])
        h2[i,j]=w[j]*dnorm((z[i,j]-ratiomean*effectsize*sqrt(In[i-1]))/
sqrt(ratiovariance))/sqrt(ratiovariance)
    }
    }else{
        for (j in 1:(n+1)) {
            h1[i,j]=0
            h2[i,j]=0
            for(l in 1:(n+1)){
                h1[i,j]=h1[i,j]+h1[i-1,l]*w[j]*sqrt(In[i-1]/(In[i-1]-In[i-2]))
)*dnorm((z[i,j]*sqrt(In[i-1])-z[i-1,l]*sqrt(In[i-2]))/sqrt(In[i-1]-In[i-2]))
-2]))
                h2[i,j]=h2[i,j]+h2[i-1,l]*w[j]*sqrt(In[i-1]/((In[i-1]-In[i-2])
)*ratiovariance))*
+dnorm((z[i,j]*sqrt(In[i-1])-z[i-1,l]*sqrt(In[i-2]))/sqrt((
In[i-1]-In[i-2])*ratiovariance)-ratiomean*effectsize*sqrt((In[i-1]-In[i-2])
)/ratiovariance))
            }
        }
    }
    d=-0.001
    r[i]=r[i-1]
    f[i]=0
    for (j in 1:(n+1)) {
        f[i]=f[i]+h1[i,j]*pnorm((z[i,j]*sqrt(In[i-1])-r[i]*sqrt(In[i]))/s
qrt(In[i]-In[i-1]))
    }

    e1=(f[i]-pi1[i])^2

    r[i]=r[i]+d
    f[i]=0
    for (j in 1:(n+1)) {
        f[i]=f[i]+h1[i,j]*pnorm((z[i,j]*sqrt(In[i-1])-r[i]*sqrt(In[i]))/s
qrt(In[i]-In[i-1]))
    }
    e2=(f[i]-pi1[i])^2

    if (e1<e2) {
        d=-d
    }

    e1=0.9
    e2=1

    while (e1<e2){
        r[i]=r[i]+d

```



```

    e2=e1
    f[i]=0
    for (j in 1:(n+1)) {
        f[i]=f[i]+h1[i,j]*pnorm((z[i,j]*sqrt(In[i-1])-r[i]*sqrt(In[i]))
/sqrt(In[i]-In[i-1]))
    }
    e1=(f[i]-pi1[i])^2
}
r[i]=r[i]-d

if(i==K){
    a[i]=r[i]
    g[i]=0
    for (j in 1:(n+1)) {
        g[i]=g[i]+h2[i,j]*pnorm((-z[i,j]*sqrt(In[i-1])+a[i]*sqrt(In[i])
-ratiomean*effectsize*(In[i]-In[i-1]))/sqrt((In[i]-In[i-1])*ratiovarian
ce))
    }
    pK=g[i]
}else{
    d=0.001
    a[i]=a[i-1]
    g[i]=0
    for (j in 1:(n+1)) {
        g[i]=g[i]+h2[i,j]*pnorm((-z[i,j]*sqrt(In[i-1])+a[i]*sqrt(In[i])
-ratiomean*effectsize*(In[i]-In[i-1]))/sqrt(In[i]-In[i-1]))
    }

    e1=(g[i]-pi2[i])^2

    a[i]=a[i]+d
    g[i]=0
    for (j in 1:(n+1)) {
        g[i]=g[i]+h2[i,j]*pnorm((-z[i,j]*sqrt(In[i-1])+a[i]*sqrt(In[i])
-ratiomean*effectsize*(In[i]-In[i-1]))/sqrt((In[i]-In[i-1])*ratiovarian
ce))
    }
    e2=(g[i]-pi2[i])^2

    if (e1<e2) {
        d=-d
    }

    e1=0.9
    e2=1

    while (e1<e2){

```

```

        a[i]=a[i]+d
        e2=e1
        g[i]=0
        for (j in 1:(n+1)) {
            g[i]=g[i]+h2[i,j]*pnorm((-z[i,j]*sqrt(In[i-1])+a[i]*sqrt(In[i
]))-ratiomean*effectsize*(In[i]-In[i-1]))/sqrt((In[i]-In[i-1])*ratiovari
ance))
        }
        e1=(g[i]-pi2[i])^2
    }
    a[i]=a[i]-d
}

for (i in 1:K){
    b[2*i-1]=a[i]
    b[2*i]=r[i]
    b[2*K+1]=m
}
return(list(lower=a,upper=r,arm_size=m))
}

```