Unconditional Exact Tests for Binomial Proportions in the Group Sequential Setting

Navneet R. Hakhu

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2014

Reading Committee:

Scott S. Emerson, Chair

Marco Carone

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Unconditional Exact Tests for Binomial Proportions
in the Group Sequential Setting

Navneet R. Hakhu

Chair of the Supervisory Committee:
Professor Scott S. Emerson
Department of Biostatistics

Exact inference for independent binomial outcomes in small samples is complicated by the presence of a mean-variance relationship that depends on nuisance parameters, discreteness of the outcome space, and departures from normality. Although large sample theory based on Wald, score, and likelihood ratio (LR) tests are well developed, suitable small sample methods are necessary when "large" samples are not feasible. Fisher's exact test, which conditions on an ancillary statistic to eliminate nuisance parameters, however its inference based on the hypergeometric distribution is "exact" only when a user is willing to base decisions on flipping a biased coin for some outcomes. Thus, in practice, Fisher's exact test tends to be conservative due to the discreteness of the outcome space.

To address the various issues that arise with the asymptotic and/or small sample tests, Barnard (1945, 1947) introduced the concept of unconditional exact tests that use exact distributions of a test statistic evaluated over all possible values of the nuisance parameter. For test statistics derived based on asymptotic approximations, these "unconditional exact tests" ensure that the realized type 1 error is less than or equal to the nominal level. On the other hand, an unconditional test based on the conservative Fisher's exact test statistic can better achieve the nominal type 1 error. In fixed sample settings, it has been found that unconditional exact tests are preferred to conditional exact tests (Mehta and Senchaudhuri, 2003). In this thesis we first illustrate the behavior of candidate unconditional

exact tests in the fixed sample setting, and then extend the comparisons to the group sequential setting. Adjustment of the fixed sample tests is defined as choosing the critical value that minimizes the conservativeness of the actual type 1 error without exceeding the nominal level of significance. We suggest three methods of using critical values derived from adjusted fixed sample tests to determine the rejection region of the outcome space when testing binomial proportions in a group sequential setting: (1) at the final analysis time only; (2) at analysis times after accrual of more than 50% of the maximal sample size; and (3) at every analysis time. We consider (frequentist) operating characteristics (Emerson, Kittelson, and Gillen, 2007) when evaluating group sequential designs: overall type 1 error; overall statistical power; average sample number (ASN); stopping probabilities; and error spending function. We find that using the fixed sample critical values is adequate provided they are used at each of the interim analyses. We find relative behavior of Wald, chi square, LR, and Fisher's exact test statistics all depend on the sample size and randomization ratio, as well as the boundary shape function used in the group sequential stopping rule. Owing to its tendency to behave well across a wide variety of settings, we recommend implementation of the unconditional exact test using the adjusted Fisher's exact test statistic at every analysis. Because the absolute behavior of that test varies according to the desired type 1 error, the randomization ratio, and the sample size, we recommend that the operating characteristics for each candidate stopping rule be evaluated explicitly for the chosen unconditional exact test.

TABLE OF CONTENTS

Page

List of Figures iii

List of Tables v

Abbreviations and Notation vi

Chapter 1: Introduction and Overview 1
1.1 AML trial 2

Chapter 2: Background: Tests of binomial proportions in small samples 4
2.1 Binomial proportions 4
2.2 Target of inference, $\theta = p_1 - p_0$ 5
2.3 Exact distribution of $\hat{\theta} = \hat{p}_1 - \hat{p}_0$ 6
2.4 Asymptotic results: Z tests for means (proportions) 6
2.5 Fisher's (conditional) exact test 10
2.6 Randomized (versus nonrandomized) tests 11
2.7 Evaluation of tests in fixed sample setting 13
2.8 Barnard's unconditional exact test 15
2.9 Rejection region 16

Chapter 3: Background: Group sequential tests using asymptotic theory 23
3.1 One sample normal setting 23
3.2 Naïve approach: Use asymptotic results with constant mean-variance relationship 27

Chapter 4: Tests of binomial proportions in sequential testing 37
4.1 Alternative approaches: Three methods of using critical values derived from adjusted fixed sample unconditional exact tests 37
4.2 Evaluation of alternative approaches: Identifying a recommended approach to use in clinical practice 38

Chapter 5: Discussion 80

Bibliography 82

Appendix A: Derivation of the fixed sample distribution of $\hat{\theta}$ via transformation (Jacobian) method 84

Appendix B: Comparison of variances according to parametrization of nuisance parameter 86

LIST OF FIGURES

Figure Number Page

2-1 FD (Unadj) . . . 19
2-2 FD (Unadj w/ Adj) . . . 22

3-1 Valid Unadj Chi Square: n_1=n_0=60 . . . 29
3-2 GSD (Poc): analyses=4, ratio=1, N=240 (Unadj w/ Adj) . . . 31
3-3 AML Trial; GSD (OBF): analyses=4, ratio=1, N=180 (Unadj) . . . 33
3-4 AML Trial; GSD (OBF): analyses=4, ratio=1, N=180 (Unadj) . . . 34
3-5 AML Trial; GSD (OBF): analyses=4, ratio=1, N=180 (Unadj) . . . 35
3-6 AML Trial; GSD (OBF): analyses=4, ratio=1, N=180 (Unadj) . . . 36

4-1 GSD (OBF): analyses=4 (Unadj w/ 3 different Adj) . . . 42
4-2 GSD (Poc): analyses=4 (Unadj w/ 3 different Adj) . . . 43
4-3 GSD (OBF): analyses=4, N=96 (Adj) . . . 45
4-4 GSD (OBF): analyses=4, N=96 (Adj) . . . 46
4-5 GSD (OBF): analyses=4, N=192 (Adj) . . . 47
4-6 GSD (OBF): analyses=4, N=192 (Adj) . . . 48
4-7 GSD (Poc): analyses=4, N=96 (Adj) . . . 49
4-8 GSD (Poc): analyses=4, N=96 (Adj) . . . 50
4-9 GSD (Poc): analyses=4, N=192 (Adj) . . . 51
4-10 GSD (Poc): analyses=4, N=192 (Adj) . . . 52
4-11 Power GSD (OBF) . . . 54
4-12 Power GSD (Poc) . . . 55
4-13 Stop Probs GSD (OBF) . . . 57
4-14 Stop Probs GSD (Poc) . . . 58
4-15 GSD (OBF): analyses=4, ratio=1, N=96, θ=0.3, ω=0.17 . . . 59
4-16 GSD (OBF): analyses=4, ratio=2, N=96, θ=0.3, ω=0.17 . . . 60
4-17 GSD (OBF): analyses=4, ratio=3, N=96, θ=0.3, ω=0.17 . . . 61
4-18 GSD (Poc): analyses=4, ratio=1, N=96, θ=0.3, ω=0.17 . . . 62
4-19 GSD (Poc): analyses=4, ratio=2, N=96, θ=0.3, ω=0.17 . . . 63
4-20 GSD (Poc): analyses=4, ratio=3, N=96, θ=0.3, ω=0.17 . . . 64

4-21 ASN GSD (OBF) 66
4-22 ASN GSD (Poc) 67
4-23 ASN vs. symm P (unified fam boundary); GSD: ratio=1, Pwr=97.5%, θ=0.3, ω=0.33 70
4-24 ASN vs. symm P (unified fam boundary); GSD: ratio=1, Pwr=97.5%, θ=0.3, ω=0.50 71
4-25 ASN vs. symm P (unified fam boundary); GSD: ratio=1, Pwr=97.5%, θ=0.5, ω=0.33 72
4-26 ASN vs. symm P (unified fam boundary); GSD: ratio=1, Pwr=97.5%, θ=0.5, ω=0.50 73
4-27 ASN vs. symm P (unified fam boundary); GSD: ratio=1, Pwr=80%, θ=0.3, ω=0.33 74
4-28 ASN vs. symm P (unified fam boundary); GSD: ratio=1, Pwr=80%, θ=0.3, ω=0.50 75
4-29 ASN vs. symm P (unified fam boundary); GSD: ratio=1, Pwr=80%, θ=0.5, ω=0.33 76
4-30 ASN vs. symm P (unified fam boundary); GSD: ratio=1, Pwr=80%, θ=0.5, ω=0.50 77

LIST OF TABLES

Table Number Page

2-1 FD: N=30 (Unadj) 14
2-2 Rejection region: $n_1 = 5$, $n_0 = 3$ 20
2-3 FD: $n_0 = 3$, $n_1 = 5$ (Unadj w/ Adj) 21
2-4 FD: N=30 (Unadj w/ Adj) 21
3-1 Symmetric Pocock GSD: N=240, ratio=1, $\theta_{Dsn} = 1$ 29
4-1 ASN vs. symm P (unified fam boundary); GSD: ratio=1, ω=0.50 78
4-2 ASN vs. symm P (unified fam boundary); GSD: ratio=2, ω=0.50 79

ABBREVIATIONS AND NOTATION

ADJ: adjusted

ALT: alternative

$\mathcal{B}$: Binomial distribution with n Bernoulli trials and probability of "success" (response/outcome) p.

CHISQ: chi square

DIFF: difference

DSMB: Data safety and monitoring board

DSN: design

FAM: family

FD: fixed (RCT) design (i.e., only one analysis time)

FISH: Fisher's exact (test)

GSD: group sequential (RCT) design

LR: likelihood ratio

$\mathcal{N}$: Normal (Gaussian) distribution with mean μ and variance σ^2.

N: maximal/total sample size

NBR: maximal number of analysis times

OBF: O'Brien-Fleming (monitoring guideline / stopping rule/boundary)

POC: Pocock (monitoring guideline / stopping rule/boundary)

PWR: (statistical) power

RATIO: randomization ratio, $n_1 : n_0$

RCT: randomized clinical trial

SEQOC: sequential operating characteristics

SIMS: simulations

SYMM: symmetric

UNADJ: unadjusted

ACKNOWLEDGMENTS

There are many people to thank for their impact on my life, and in particular for the completion of this thesis. It is an impossible task to write down every single individual or group who has impacted and/or influenced me in some capacity. As a result, I mention here those who have had the greatest and most direct impact on the completion of this thesis. First, I thank my advisor and chair of my reading committee, teacher, and friend Scott Emerson for his continued support, encouragement, and guidance over the years. Second, I thank my other reading committee member Marco Carone for his review and comments that were invaluable. Third, I thank my Research Assistant supervisor Jim Hughes, the Graduate Program Advisor Gitana Garofalo, and my friends and fellow classmates (Elisa Sheng, Jason Liang, Clara Domínguez-Islas, William Koh, Bob Salim, and Chloe Krakauer) who supported me throughout this process. Lastly, I thank my sister and brother-in-law (Nisha and Patrick Doherty) and Mom and Dad (Nalini and Jai Hakhu), whose words of encouragement and love have supported me over the course of my life. To professors, classmates, friends, family members, those who are living, those who have departed, those who I have known briefly, and those who I have known my entire life, I convey my sincerest appreciation and gratitude: Thank you, each and every one of you.

DEDICATION

to Mom and Dad

Chapter 1

INTRODUCTION AND OVERVIEW

Randomized clinical trials (RCTs) represent an important tool in the clinical research directed toward improving public health. Well designed and conducted RCTs, when ethical and feasible, are considered the gold standard for assessing cause and effect. Randomization is an important aspect of the design and implementation of RCTs to be able, on average, to have comparable treatment arms. In many such RCTs, the primary measure of treatment outcome is the difference in event probabilities between two treatment arms. In this thesis, we focus on the methods used to analyze the results of such RCTs when conducted with small to moderate sample sizes in the phase 2 (screening) or phase 3 (confirmatory) settings.

In Chapter 1 we motivate the two sample binomial proportions setting with an example from an acute myelogenous leukemia (AML) RCT. Using this example we discuss the application of our research extending established methods in the fixed sample setting to the group sequential setting. Chapter 2 focuses on tests for binomial proportions in small samples, illustrating the inherent problems with the typical approaches used for analysis. We discuss four common methods, namely the Wald, score (chi square), likelihood ratio (LR), and (conditional) Fisher's exact tests. We thus motivate and introduce Barnard's (1945, 1947) concept of unconditional exact tests and the role it has had compared to conditional exact tests in fixed sample settings (Mehta and Senchaudhuri, 2003). Chapter 3 introduces group sequential tests using asymptotic theory. We start with the one sample normal setting to illustrate the additional considerations of the sequential nature of the design as compared to the fixed design. We first focus on what was done in the AML trial: a naïve approach that used asymptotic results for the chi square test, assuming a constant mean-variance relationship when deriving power curves (and confidence intervals). We illustrate problems with such an approach and use the AML trial as motivation for using the unconditional exact test, where adjusting the critical value, instead of using the normal approximation critical

value, ensures that the overall type 1 error of an RCT does not exceed the pre-specified design nominal level. In Chapter 4 we suggest three alternative forms of adjustment when using the unconditional exact test in sequential testing, namely adjusting only at the final analysis time, at each analysis that occurs after accruing more than 50% of the maximal sample size (statistical information) of the study, and at every analysis. We evaluate the three forms of adjustment for a variety of group sequential designs according to different features and operating characteristics. Chapter 5 concludes this thesis with a discussion of our findings, including limitations, next steps, and recommendations we make for designing and analyzing RCTs with binary outcomes and small to moderate sample sizes in clinical practice.

1.1 AML trial

As a motivating example, we consider a phase 3 RCT from 1984–1989 which compared two anthracycline chemotherapy agents for the treatment of acute myelogenous leukemia (AML) among adult patients from Memorial Sloan Kettering Cancer Center. Patients were randomized in 1:1 fashion to receive either the new treatment (idarubicin) or the then current standard (daunorubicin). The primary outcome was whether each patient achieved complete remission (a binary outcome). The target of inference was the difference between patients randomized to receive idarubicin or daunorubicin with respect to the proportions achieving a complete remission.

There was an adaptive component of the trial: a group sequential stopping rule was introduced after the start of the study. Emerson and Banks (1994) discuss the amendment of the original study protocol (fixed design with only one analysis at the end of the study) to a group sequential design with O'Brien-Fleming monitoring guidelines (1979), having a maximum of 4 analysis times, and 80% statistical power to detect difference in proportions of 0.20. Under that design, a maximal sample size of 180 patients (90 in each arm) would be accrued to the study. The four analysis times corresponded to total sample sizes of 25, 45, 65, and 90 per arm, respectively.

Statistical inference consisted of sequential analyses using the Z test for proportions based on asymptotic results. Such a test is equivalent to the chi square test. The first formal

interim analysis occurred after accrual of 45 patients to each arm at which time 35 patients (78%) on the idarubicin arm had complete remission compared to 25 patients (56%) on the daunorubicin arm. The observed numbers (proportions) of patients with complete remission correspond to a chi square test Z statistic of 2.236, which was smaller than the group sequential efficacy critical value (upper monitoring guideline) of 2.863. Based on the group sequential O'Brien-Fleming monitoring guideline, this was insufficient evidence of efficacy (or futility) to warrant termination of the study. The next interim analysis occurred after 65 patients had been accrued to each arm. At that time, 51 patients (78%) on the idarubicin arm had a complete remission compared to 38 patients (58%) on the daunorubicin arm. The observed numbers (proportions) of patients with complete remission correspond to a chi square test Z statistic of 2.454 which exceeded the group sequential efficacy critical value of 2.382. Based on the results of the primary efficacy analysis as well as secondary analyses (number of complete remissions after one course of treatment and number of resistant disease after two courses of chemotherapy), the Data Safety and Monitoring Board (DSMB) recommended termination of the trial in favor of adopting idarubicin over daunorubicin as chemotherapy treatment for AML patients.

In this thesis, we plan to focus on the validity of using asymptotic results for hypothesis testing in the binomial proportions setting with small to moderate sample sizes; full estimation, including confidence intervals, is not addressed in this thesis. The question we ask: Are asymptotic results appropriate to use in group sequential settings with a binary outcome and small to moderate sample sizes, and if they are not, what testing procedure should be used instead?

Chapter 2

BACKGROUND: TESTS OF BINOMIAL PROPORTIONS IN SMALL SAMPLES

In this chapter, we describe historical approaches of hypothesis testing when the difference of independent proportions is the target of inference. We introduce the Z test for means (proportions) followed by its variants based on the asymptotically equivalent Wald, score (chi square), and likelihood ratio (LR) tests. Also considered is the conditional Fisher's exact test, which conditions on an ancillary statistic. As we discuss these four tests, we discuss the problems that are inherent to binary outcomes, especially with small to moderate sample sizes: nuisance parameters; mean-variance relationship; discreteness of outcome space; and departures from normality. We evaluate the four tests and compare them to asymptotic results for a variety of fixed sample RCT designs to illustrate the need to use an alternative approach to handle the problems discussed. Concerns about the validity of asymptotic results in small to moderate samples lead to discussion of Barnard's concept of unconditional exact tests (1945, 1947). Considerations in the fixed sample setting are a prelude to Chapter 3 where we introduce and discuss the group sequential setting.

2.1 Binomial proportions

Consider two treatment arms $i = 0, 1$ consisting of n_i individuals per arm. As in the AML trial, suppose the binary (Bernoulli) outcome indicates achievement of complete remission for individual j in treatment arm i, denoted by

$$Y_{ij} \overset{indep}{\sim} \mathcal{B}(1, p_i)$$

where p_i is the probability of achieving complete remission, and assumed to be the same for all individuals within each respective treatment arm. The number of individuals who achieved complete remission in group i is the sum of the n_i independent Bernoulli random

variables denoted by

$$A_i \equiv \sum_{j=1}^{n_i} Y_{ij} \sim \mathcal{B}(n_i, p_i)$$

with mean $n_i p_i$ and variance $n_i p_i(1-p_i)$ for group $i = 0, 1$. In this thesis, we assume A_1 and A_0 are independent binomial random variables. Furthermore, we can summarize the outcomes in each treatment arm i by reporting an estimator of the sample mean (i.e., proportion) of individuals who achieved complete remission induction denoted by

$$\hat{p}_i \equiv \frac{A_i}{n_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \sim \left(p_i, \frac{p_i(1-p_i)}{n_i} \right)$$

where the specified mean and variance (i.e., the first two moments) of the distribution of $\hat{p}_i$ is derived from the properties of expectation and variance of independent random variables.

2.2 Target of inference, $\theta = p_1 - p_0$

With a binary outcome, several possible targets of inference include the risk difference or difference in proportions $(p_1 - p_0)$, the risk ratio or relative risk (p_1/p_0), and the odds ratio (ratio of odds of outcome in group 1, $p_1/(1-p_1)$, and odds of outcome in group 0, $p_0/(1-p_0)$). In this thesis, we are going to focus on questions of interest where the scientifically and clinically useful summary measure for quantification of a treatment effect is the difference in the proportion with outcome across two groups. Specifically we focus on the difference in proportions $\theta = p_1 - p_0$ as the target of inference, where p_1 is the unknown probability of response on the experimental treatment arm and p_0 is the unknown probability of response on the control arm. Additionally, we assume that we are not asking about effect modification, that randomization precludes confounding, and that there is no desire to adjust for precision variables.

We want to perform a hypothesis test in which we test the null hypothesis (H_0) against the design alternative hypothesis (H_A)

$$H_0 : \theta \leq \theta_0 \quad vs. \quad H_A : \theta \geq \theta_A.$$

In this thesis, we focus on testing where $\theta_0 = 0$. Let T denote a test statistic that is a function of a_0 and a_1. Of interest in this hypothesis test is to have for some critical value

c_{θ_0} a low false positive error rate

$$\alpha = \Pr(T \geq c_{\theta_0} \mid \theta = \theta_0)$$

(i.e., low type 1 error, the probability of rejecting the null hypothesis when the null hypothesis is true) and low false negative error rate (i.e., type 2 error) that we parametrize as the statistical power

$$\beta = \Pr(T \geq c_{\theta_0} \mid \theta = \theta_A)$$

(i.e., probability of rejecting the null hypothesis when the alternative hypothesis is true). More generally, the power function is denoted by $Pwr(\theta) \equiv \beta(\theta)$ for arbitrary values of θ.

2.3 Exact distribution of $\hat{\theta} = \hat{p}_1 - \hat{p}_0$

A natural place to start is with the exact distribution of $\hat{\theta}$. In Appendix A, we show the derivation of the fixed sample distribution of $\hat{\theta}$ via transformation (Jacobian) method. Our purpose in illustrating some of the details of the derivation is to show that deriving the exact distribution function of $\hat{\theta}$ is difficult for two reasons: (1) integrating out terms within gamma functions (i.e., originally as part of the binomial coefficients); and (2) even if we could manage to obtain a closed form expression for the distribution function of $\hat{\theta}$, we would still need to know the true values of p_1 and p_0. Of course, for the latter case, we could plug in estimates of p_1 and p_0, namely the maximum likelihood estimates (MLEs) $\hat{p}_1 = a_1/n_1$ and $\hat{p}_0 = a_0/n_0$, respectively. However, with the difficulty of obtaining a closed form expression for the distribution function of $\hat{\theta}$, plugging in MLEs for p_1 and p_0 does not help.

2.4 Asymptotic results: Z tests for means (proportions)

Alternatively, we can consider the asymptotic distribution of $\hat{\theta}$. By the Levy Central Limit Theorem (CLT),

$$\frac{\hat{p}_i - p_i}{\sqrt{\frac{p_i(1-p_i)}{n_i}}} \longrightarrow_d \mathcal{N}(0,1),$$

alternatively written as

$$\hat{p}_i \dot{\sim} \mathcal{N}\left(p_i, \frac{p_i(1-p_i)}{n_i}\right).$$

The maximum likelihood estimator (MLE) of p_i is $\hat{p}_i = A_i/n_i$ for group $i = 0, 1$. Using properties of independent (approximately) normal random variables, $\hat{\theta} = \hat{p}_1 - \hat{p}_0$ has an approximate normal distribution

$$\hat{\theta} \dot{\sim} \mathcal{N}\left(\theta = p_1 - p_0,\ \frac{p_0(1-p_0)}{n_0} + \frac{p_1(1-p_1)}{n_1}\right).$$

A consequence of the above result is the Z test for proportions based on the following normalized quantity which has a standard normal distribution

$$Z_{\mathcal{A}} = \frac{\hat{\theta} - \theta_0}{\sqrt{\frac{p_0(1-p_0)}{n_0} + \frac{p_1(1-p_1)}{n_1}}} \dot{\sim} \mathcal{N}(0,1)$$

where $\hat{\theta} = A_1/n_1 - A_0/n_0$. The problem with the asymptotic distributions of $\hat{\theta}$ and $Z_{\mathcal{A}}$ is the dependence on unknown parameters p_0 and p_1. One approach to account for the unknown true values of p_0 and p_1 is to plug in estimates of the truth yielding an estimator for the normalized quantity.

2.4.1 Wald

We first consider using the MLEs $\hat{p}_0 = A_0/n_0$ and $\hat{p}_1 = A_1/n_1$ as plug-in estimators for p_0 and p_1, respectively. For an observed pair of values $(A_0, A_1) = (a_0, a_1)$, the Wald Z statistic is defined as

$$Z_{\mathcal{W}} = \frac{\hat{\theta} - \theta_0}{\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_0} + \frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}}$$

$$\overset{(\theta_0=0)}{=} \frac{\frac{a_1}{n_1} - \frac{a_0}{n_0}}{\sqrt{\frac{a_0(n_0-a_0)}{(n_0)^3} + \frac{a_1(n_1-a_1)}{(n_1)^3}}}.$$

2.4.2 Alternative parameterization

The Wald test uses unrestricted MLEs $\hat{p}_1$ and $\hat{p}_0$ as plug-in estimators which do not make use of the relationship between p_1 and p_0 under the null hypothesis. Similar to how we define the population parameter $\theta = p_1 - p_0$, we define the following relationship between the MLE under the null hypothesis $\hat{p}_1^{(0)}$ and $\hat{p}_0^{(0)}$ as

$$\theta_0 = \hat{p}_1^{(0)} - \hat{p}_0^{(0)} \quad \overset{(\theta_0=0)}{\Longrightarrow} \quad \hat{p}_1^{(0)} = \hat{p}_0^{(0)}$$

where we estimate $\hat{p}_0^{(0)}$ using MLE restricted to the null hypothesis. Inherent to our initial parameterization of the target of inference are ***nuisance parameters*** p_1 and p_0. We can, however, focus on p_0 as the nuisance parameter since p_1 is determined once θ and p_0 are known. We note that θ does not discriminate between two studies with an estimated difference in proportions of 0.3, where in the first study $p_1 = 0.4$ and $p_0 = 0.1$, and where in the second study $p_1 = 0.8$ and $p_0 = 0.5$. However, according to the asymptotic results provided earlier, the exact values of p_0 and p_1 play an important role in the variance of the test statistic, and thus we must account for the nuisance parameter.

An alternative parameterization of the nuisance parameter especially useful for $\theta_0 = 0$ (as is presumed in this thesis) is a weighted average of the two proportions

$$\omega = \frac{n_0 p_0 + n_1 p_1}{n_0 + n_1},$$

noting that $\hat{\omega}$ corresponds to the restricted MLE as is used in the common Z test of binomial proportions. In this case we find it useful to consider the hypothesis test of interest

$$H_0 : p_1 = p_0 = \omega \quad \text{vs.} \quad H_A : \begin{Bmatrix} p_1 = \omega + \theta_A/2 \\ p_0 = \omega - \theta_A/2 \end{Bmatrix}$$

since under the null hypothesis, $(n_0\omega + n_1\omega)/(n_0 + n_1) = \omega = p_0 = p_1$, and for any design alternative, the difference in proportions $p_1 - p_0 = (\omega + \theta_A/2) - (\omega - \theta_A/2) = \theta$. See Appendix B for variances of the estimator $\hat{\theta}$ according to nuisance parameter parametrization (p_0 vs. ω). The variance of $\hat{\theta}$ under the ω parameterization does not have a linear term in θ as compared to the variance of $\hat{\theta}$ under the p_0 parameterization.

The above approach is primarily of interest to us as we design a study under the alternative hypothesis in such a way that corresponds closely to the restricted MLE under the null hypothesis. We want to consider a sequence of alternatives that move outward from some central null hypothesis, rather than fixing, say, p_0 and considering p_1 deviating from that. While there are multiple ways this can be done, we suspect parameterizing according to ω may better handle the ***mean-variance relationship***, inherent to binary outcomes since the variance is a function of the mean.

2.4.3 Score (chi square)

Now under the null hypothesis, the restricted MLE for p_0 is

$$\hat{p}_0^{(0)} = \hat{p}_1^{(0)} = \hat{\omega} \equiv \frac{n_0\hat{p}_0 + n_1\hat{p}_1}{n_0 + n_1} = \frac{A_0 + A_1}{n_0 + n_1}.$$

For an observed pair of values $(A_0, A_1) = (a_0, a_1)$, the score (chi square) Z statistic is defined as

$$Z_{\mathcal{S}} = \frac{\hat{\theta} - \theta_0}{\sqrt{\frac{\hat{p}_0^{(0)}(1-\hat{p}_0^{(0)})}{n_0} + \frac{\hat{p}_1^{(0)}(1-\hat{p}_1^{(0)})}{n_1}}}$$

$$\overset{(\theta_0=0)}{=} \frac{\frac{a_1}{n_1} - \frac{a_0}{n_0}}{\sqrt{\frac{\hat{\omega}(1-\hat{\omega})}{n_0} + \frac{\hat{\omega}(1-\hat{\omega})}{n_1}}} = \frac{\frac{a_1}{n_1} - \frac{a_0}{n_0}}{\sqrt{\frac{(a_0+a_1)(n_0+n_1-a_0-a_1)}{n_0 n_1 (n_0+n_1)}}}.$$

2.4.4 Likelihood ratio

A third Z statistic is based on the LR motivated by the Neyman-Pearson lemma. For an observed pair of values $(A_0, A_1) = (a_0, a_1)$ and samples sizes n_0 and n_1, the likelihood function is

$$L(\theta, p_0) = \binom{n_0}{a_0} p_0^{a_0}(1-p_0)^{n_0-a_0} \cdot \binom{n_1}{a_1}(p_0+\theta_0)^{a_1}(1-p_0-\theta_0)^{n_1-a_1}.$$

The LR is defined as

$$\hat{\lambda} = \frac{L\left(\hat{\theta} = \frac{a_1}{n_1} - \frac{a_0}{n_0},\ \hat{p}_0 = \frac{a_0}{n_0}\right)}{L\left(\theta_0 = 0,\ \hat{p}_0 = \hat{p}_0^{(0)}\right)} = \frac{\left[\frac{(a_0)^{a_0}(n_0-a_0)^{n_0-a_0}}{(n_0)^{n_0}}\right] \cdot \left[\frac{(a_1)^{a_1}(n_1-a_1)^{n_1-a_1}}{(n_1)^{n_1}}\right]}{\left[\frac{(a_0+a_1)^{a_0+a_1}(n_0+n_1-a_0-a_1)^{n_0+n_1-a_0-a_1}}{(n_0+n_1)^{n_0+n_1}}\right]}$$

and the LR signed Z statistic is defined as

$$Z_{\mathcal{L}} = sgn(\hat{\theta})\sqrt{2\log\hat{\lambda}}$$

where

$$sgn(\hat{\theta}) = \begin{cases} +1, & \hat{\theta} \geq 0 \\ -1, & \hat{\theta} < 0\,. \end{cases}$$

2.4.5 Properties common to asymptotic tests

In small to moderate samples we have to pay attention to issues which arise from having a mean-variance relationship and discreteness of the outcome space, whereas asymptotic tests indicate these issues are not a problem with large (enough) samples. From asymptotic theory, we know the Wald, chi square, and LR tests are all asymptotically equivalent: each test's Z statistic has an approximately standard normal distribution under the null hypothesis of $\theta_0 = 0$. Therefore, the one-sided P value corresponding to each test is defined as

$$\begin{aligned} P_{\mathcal{W}} &= 1 - \Phi(z_{\mathcal{W}}), \\ P_{\mathcal{S}} &= 1 - \Phi(z_{\mathcal{S}}), \text{ and} \\ P_{\mathcal{L}} &= 1 - \Phi(z_{\mathcal{L}}) \end{aligned}$$

where $z_{\mathcal{W}}$, $z_{\mathcal{S}}$, and $z_{\mathcal{L}}$ are the observed values of the respective Z statistics and Φ denotes the standard normal cumulative distribution function. However, in small to moderate samples, the sampling distribution of these statistics will not be exactly normal and will still depend upon the nuisance parameters. We further note (and as we will later demonstrate) in small to moderate samples the hypothesis tests based on these statistics can be either conservative or anti-conservative depending on the value of our nuisance parameter ω.

2.5 Fisher's (conditional) exact test

Instead of relying on asymptotic results, an alternative is to use Fisher's exact test which conditions on ancillary statistic $A_0 + A_1$, eliminating the nuisance parameter ω. Fisher's exact test is the two-sided uniformly most power unbiased randomized test. For an observed pair of values $(A_0, A_1) = (a_0, a_1)$ and sample sizes n_0 and n_1, the (one-sided) Fisher's exact test P value statistic can be defined as

$$P_{\mathcal{F}} = \sum_{i=0}^{n_0} \sum_{j=0}^{n_1} \frac{\binom{n_0}{i}\binom{n_1}{j}}{\binom{n_0+n_1}{i+j}} \cdot \mathbf{1}_{[i \leq a_0]} \cdot \mathbf{1}_{[i+j=a_0+a_1]},$$

where $A_0|(A_0 + A_1)$ has a hypergeometric distribution under the null hypothesis.

2.6 Randomized (versus nonrandomized) tests

In the previous sections, we described derivations of the Wald, chi square, LR, and Fisher's exact test statistics. We have yet to discuss how to achieve the nominal type 1 error exactly. We first suppose we knew the value of nuisance parameter ω. To illustrate how to achieve the nominal type 1 error exactly recall from testing theory the Neyman-Pearson lemma which says that for simple null H_0 versus simple alternative H_A hypotheses, ordering the outcome space according to the LR test statistic λ yields the most powerful level α test with test function

$$\phi = \begin{cases} 1 & \text{if } \lambda > c \\ \gamma & \text{if } \lambda = c \\ 0 & \text{if } \lambda < c \end{cases}$$

where $\gamma \in (0, 1)$ and nominal type 1 error

$$\alpha = E[\phi \mid H_0] = \Pr(\lambda > c \mid H_0) + \gamma \cdot \Pr(\lambda = c \mid H_0).$$

For both LR and Fisher's exact tests, achievement of the nominal type 1 error exactly requires randomizing the decision to reject the observed test statistic (outcome) value by the proportion γ (i.e., "flip a biased coin") when there is nonzero probability that the observed test statistic is equal to critical value c. When we implement a randomized test, we typically generate a uniform random variable $U \sim \mathcal{U}(0, 1)$ that is independent of A_0 and A_1. We reject the null hypothesis when $U < \gamma$. Extending results from the Neyman-Pearson lemma for one-sided hypothesis tests of simple hypotheses to composite hypotheses is done by additionally showing monotone likelihood ratio and invoking the Karlin-Rubin theorem. Such a procedure shows that the LR test is the uniformly most powerful one-sided level α test when randomized. A similar randomized test procedure could be applied to the Wald and chi square tests to deal with the discreteness of the distribution. However, implied in the above procedure is that we know the exact distribution of the statistics and can find critical value c exactly. This is not the case due to our nuisance parameter.

Fisher's exact test is similarly affected by discreteness of the outcome space, but because

its distribution does not depend on the nuisance parameter we can achieve the exact type 1 error through the use of a randomized test.

We thus have two choices for critical values:

(1) use randomized tests to exactly achieve the nominal level; or

(2) use nonrandomized tests and not achieve the nominal level.

Which choice we implement depends in part on whether we know the value of the nuisance parameter; of course, in practice, we typically do not know. When we do not know the value of nuisance parameters, we obtain a type 1 error exactly equal to the nominal level only with randomized tests. An issue people have with randomized tests is that they allow for the possibility of different conclusions of analyses despite the observed outcome being the same. Therefore, despite theory indicating randomized tests as optimal, nonrandomized tests are preferred over randomized tests in clinical practice. If we knew the value of the nuisance parameter, performing a nonrandomized test would yield a conservative test. On the other hand, by asymptotic theory using the normal approximation as shown earlier, when we do not know the value of the nuisance parameter we can end up with either anti-conservative or conservative tests. For Fisher's exact test, however, we condition on an ancillary statistic obtaining an exact randomized test, and thus, always yielding a conservative nonrandomized test.

We note that we have not addressed Yates' continuity correction (1934) to the chi square test, which is another approach sometimes advocated to deal with the discreteness of the outcome space. The use of Yates' continuity correction typically results in a test that behaves very similar to Fisher's exact test, and thus we will not explore that further.

In the following section and remainder of this thesis, we focus on evaluating nonrandomized tests noting that 1) Fisher's exact test will tend to be conservative, and 2) with small to moderate sample sizes the Wald, chi square, and LR tests can behave poorly (i.e., have different operating characteristics such as type 1 error) due to how each test handles the presence of a mean-variance relationship that depends on nuisance parameters, discreteness

of the outcome space, and departures from normality (i.e., standard normal distribution poorly approximates the true distribution of the statistics).

2.7 Evaluation of tests in fixed sample setting

2.7.1 Simulation study: setup and results

We plan to compare the type 1 error obtained from the Wald, chi square, LR, and Fisher's exact tests (termed 'unadjusted') to the asymptotic result (i.e., nominal level in fixed sample setting). Since Fisher's exact test computes a P value as its test statistic, for comparison of the four tests, we will ultimately convert the Wald, chi square, and LR Z statistics to their respective corresponding fixed sample P values based on the normal approximation as given in previous equations. We evaluate the type 1 error for fixed sample RCT designs across a range of possible values of nuisance parameter ω, based on 10^5 simulations. With this number of simulations, the central 95% of predicted P values for the two considered one-sided nominal levels of 0.025 and 0.005 are under the null hypothesis (0.02403, 0.02597) and (0.00456, 0.00544), respectively. We consider total sample sizes of 30, 90, and 180 individuals under 1:1 and 2:1 randomization assignments in Figure 2-1.

In Figure 2-1 we consider the fixed RCT with sample size of 30 and $\omega = 0.25$. Table 2-1 shows the simulated type 1 error for the Wald, chi square, LR, and Fisher's exact tests against two nominal levels with two randomization ratios. We find that for this particular value of the nuisance parameter:

- Wald test is anti-conservative in every scenario;
- chi square test is anti-conservative for the 1:1 randomization and 0.025 nominal level scenario, but conservative in the other scenarios;
- LR test is anti-conservative in every scenario; and
- Fisher's exact test is markedly conservative in every scenario.

We note that the pattern of conservative versus anti-conservative tests varies with the value of ω.

In Figure 2-1, we see similar tendencies of each test across the range of values of the nuisance parameter, including markedly higher type 1 error for the Wald, chi square, and LR tests with 2:1 randomization compared to 1:1 randomization. The observed asymmetry in the curves with unequal sample sizes might have been expected in those one-sided tests owing to the greater discreteness (and hence higher probability of observing $\hat{p}_0 = 0$ or 1) for the group with the smaller sample size, relative to the lesser relative chance of observing $\hat{p}_1 = 0$ or 1. The type 1 error computed from Fisher's exact test does not appear as affected by unequal sample sizes in the way that the other tests are affected. However, there is often extreme conservatism due to the discreteness.

Table 2-1: Comparison of type 1 error computed for the Wald, chi square, LR, and Fisher's exact (nonrandomized, unadjusted) tests for nuisance parameter $\omega = p_0 = p_1 = 0.25$ using asymptotic results based on 10^5 simulations to nominal level, according to randomization ratio for sample size of 30.

	sample size = 30			
	1:1 randomization		2:1 randomization	
nominal level	0.025	0.005	0.025	0.005
type 1 error	*Unadj*	*Unadj*	*Unadj*	*Unadj*
Wald	0.030	0.009	0.063	0.035
Chi square	0.027	0.004	0.015	0.001
LR	0.029	0.009	0.040	0.012
Fisher's exact	0.009	0.001	0.006	0.0002

2.7.2 Overall comments

We found that unadjusted tests, either based on asymptotic results or based on conditioning on an ancillary statistic, do not have the desired type 1 error. For any fixed value of nuisance parameter ω as the sample size increases excursions from the nominal level (above the solid gold reference line is anti-conservative and below is conservative) diminish. Wald, chi square, and LR tests vary between being conservative and anti-conservative across the range of possible values of ω. Fisher's exact test is quite conservative as a result of performing

the nonrandomized version of the test in the presence of discreteness of the outcome space. Results for different sample sizes, nominal levels, and randomization ratios were similar in that neither asymptotic approximations nor conditioning on ancillary statistic appear to be suitable approaches in the binomial proportions setting with small to moderate sample sizes. Departures from normality is a problem when asymptotic tests' statistics do not have an approximately normal distribution. Hence, using a critical value based on the normal approximation may yield less than ideal inference (i.e., markedly conservative or anti-conservative results).

2.8 Barnard's unconditional exact test

Barnard (1945, 1947) introduced the concept of unconditional exact tests as an alternative to the conditional Fisher's exact test. The unconditional exact test can use any test statistic. Instead of conditioning on an ancillary statistic to eliminate the nuisance parameter, the unconditional exact test handles the nuisance parameter by choosing the worst case over all possible values of the nuisance parameter for a given test statistic. For null hypothesis $\theta = \theta_0$, the P value for observing outcome $(A_0, A_1) = (a_0, a_1)$ for a given value of the nuisance parameter ω is

$$P_\omega = \sum_{i=0}^{n_0} \sum_{j=0}^{n_1} \Pr\left[\, (A_0, A_1) = (i, j) \mid \theta_0, \omega \,\right] \cdot \mathbf{1}_{[t(i,j) \geq t(a_0,a_1)]}$$

where $t(a_0, a_1)$ is the observed value of a test statistic T. Barnard's unconditional exact P value is the supremum (i.e., worst case) of P_ω values

$$P_B = \sup_\omega \{P_\omega\}$$

over the range of $\omega \in [0, 1]$. In the following section we highlight the differences between unadjusted tests based on asymptotic results, conditional Fisher's exact test, and unconditional exact tests using a toy example to demonstrate how to compute type 1 error and P values with the different approaches.

2.9 Rejection region

With careful study, Table 2-2 illustrates Barnard's procedure. We depict the rejection region of outcomes for treatment arms with extremely small sample sizes of $n_1 = 5$ and $n_0 = 3$. Each (a_0, a_1) cell contains the joint binomial cell probability under two specially chosen values of ω as described later. Also presented are the Wald Z, chi square Z, LR Z, and Fisher's exact test P value statistics. (Here we present test statistics as they would typically be computed, though we note that we could also have converted them to their "nominal" unadjusted P values.)

The conditional Fisher's exact P value is based on the hypergeometric distribution. When using unadjusted Z tests, the type 1 error is computed assuming the Z statistic has an approximately standard normal distribution. On the other hand, Barnard's approach computes the type 1 error by summing up the binomial cell probabilities of the cells in the outcome space associated with having test statistic larger than its corresponding critical value (i.e., cell probabilities for rejected cells).

The Z critical value is 1.96 when using unadjusted Wald, chi square, and LR tests based on asymptotic results and the P critical value is the nominal level 0.025 when using Fisher's exact test. Shaded cells (green=Wald, red=chi square, grey=LR, and blue=Fisher's exact) correspond to test statistics which are more extreme than the unadjusted critical value, either $Z > 1.96$ or $P < 0.025$. The unadjusted type 1 error is the sum of the rejected (joint) binomial cell probabilities. We see from Table 2-3 for both $\omega_{\mathcal{W}} = 0.44$ and $\omega_{\mathcal{S}} = 0.67$ the Wald, chi square, and LR tests all have anti-conservative unadjusted type 1 error whereas Fisher's exact test has markedly conservative unadjusted type 1 error.

Instead of using the unadjusted critical value, suppose we use an adjusted critical value found by ordering the Z test statistics from largest to smallest and P test statistics from smallest to largest, and for each test keep rejecting cells until obtaining the maximum cumulative cell probability without exceeding the nominal level of 0.025. In this setting, it turns out that for all four tests, each tests' respective adjusted critical values is based on rejecting outcome pairs (0,4) and (0,5). Since we have extremely small sample sizes, it is not surprising that the adjusted type 1 error for each ω is the same among the four tests, 0.021

and 0.017, respectively. From the following table we find that the Wald, chi square, and LR tests had anti-conservative unadjusted type 1 error, and after using each tests' respective adjusted critical value, these tests all had conservative type 1 error. On the other hand, using an adjusted critical value for Fisher's exact test lessened the conservativeness of the type 1 error, without exceeding the nominal level.

Now that we have described how we obtain an adjusted critical value, we want to compare computation of P values based on asymptotic results, conditional Fisher's exact test, and Barnard's concept of unconditional exact tests. Suppose the observed outcome is $(a_0, a_1) = (1, 5)$. As discussed in previous sections, we can compute the unadjusted P values for the Wald, chi square, and LR tests assuming each test statistic has an approximately standard normal distribution under the null hypothesis of no difference in proportions. The unadjusted Fisher's exact P value is computed assuming the the conditional distribution has a hypergeometric distribution under the null hypothesis of no difference in proportions.

As an alternative to computing unadjusted P values, we will compute the adjusted P values using Barnard's unconditional exact approach. If we consider how we ordered the Z and P statistics earlier, we sum all the binomial cell probabilities corresponding to all test statistics at least as extreme as that observed for $(a_0, a_1) = (1, 5)$. Such an approach yields an adjusted P value based on a given test statistic for a single nuisance parameter value, denoted by P_ω. To obtain Barnard's unconditional exact P value, we need to compute the set $\{P_\omega\}$ for all possible values of ω and take the worst case (i.e., supremum of the set) to obtain P_B. Table 2-3 includes both unadjusted and adjusted P values. However, since we only considered two particular values of ω, the respective P_ω values do not necessarily correspond to Barnard's unconditional exact P_B. Recall that for illustrative purposes, we chose $\omega_{\mathcal{W}} = 0.44$ because this value of the nuisance parameter corresponds to Barnard's unconditional exact P_B when using the Wald test statistic. Similarly, $\omega_{\mathcal{S}} = 0.67$ is the value of the nuisance parameter that corresponds to Barnard's unconditional exact P_B when using the chi square test statistic. In fact, for $n_0 = 3$ and $n_1 = 5$, P_B for chi square, LR, and Fisher's exact test statistics are equivalent.

Overall, using Barnard's concept of unconditional exact tests is an approach that 1) incorporates using the exact binomial distribution instead of relying on asymptotic results especially with small to moderate samples, 2) ensures that the nominal type 1 error is not exceeded because it maximizes across all nuisance parameter values, and 3) has the flexibility of using any test statistic where some may better handle the mean-variance relationship while others may better handle the discreteness of the outcome space. In Figure 2-2 we evaluate the type 1 error for unadjusted and adjusted tests in the fixed sample setting as a precursor to examining unconditional exact tests in the group sequential setting.

In Figure 2-2 we consider again the fixed RCT with sample size of 30 and $\omega = 0.25$. Table 2-4 shows the simulated type 1 error for the Wald, chi square, LR, and Fisher's exact tests (both unadjusted and adjusted) against two nominal levels with two randomization ratios. Note that the adjusted critical values are computed using the exact binomial distribution over a range of values for ω, and thus the true inference for the adjusted tests are thus known to be conservative. However, in Figure 2-2, the results shown are subject to simulation error.

We find that for this particular value of the nuisance parameter, the adjusted Wald and adjusted LR are least conservative among the four adjusted tests. In Figure 2-2 we find that the adjusted versions of Wald, chi square, and Fisher's exact tend to have similar type 1 error across nuisance parameter values under 1:1 randomization. However, adjusted versions of Wald, chi square, and LR are too conservative under 2:1 randomization, whereas adjusted Fisher's exact appears to handle the discreteness of the outcome space from unequal sample sizes better, with type 1 error closer to the nominal level.

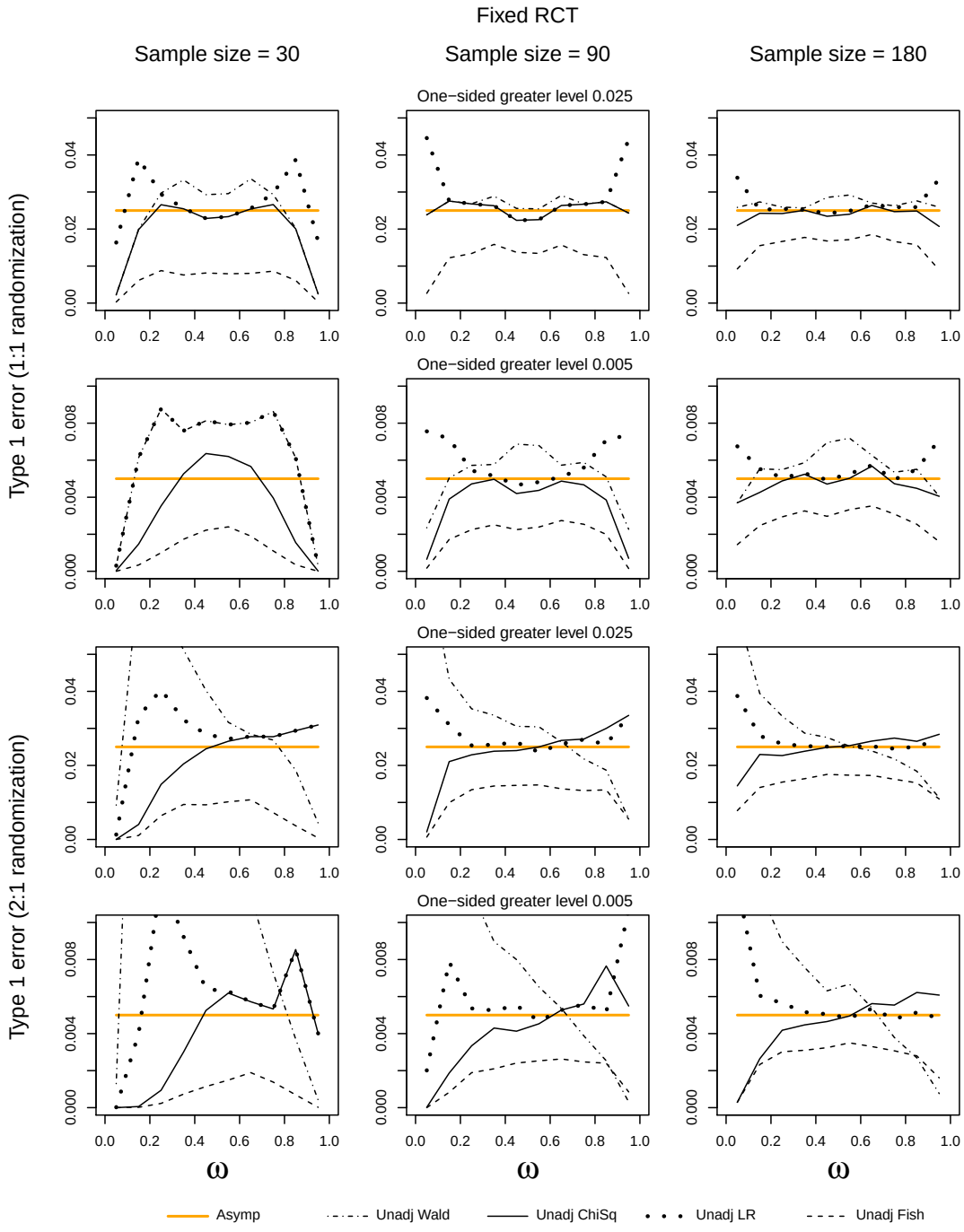


Figure 2-1: Consider FD with sample sizes of 30, 90, and 180, 1:1 and 2:1 randomization, and nominal levels of 0.025 and 0.005. Type 1 error is computed based on asymptotics (gold) and unadjusted (black) tests across a range of nuisance parameter ω from 10^5 sims.

Table 2-2: The rejection region for $n_0 = 3$ and $n_1 = 5$ includes the cell probability for each specified value of nuisance parameter $\omega = p_0 = p_1$. Shaded cells correspond to test statistics which are more extreme than the unadjusted critical value (either $Z > 1.96$ or $P < 0.025$). Values superscripted by $(\mathcal{W}, \mathcal{S})$ indicate rejected cells when using an adjusted critical value for the corresponding ω.

a_0 \ a_1	0	1	2	3	4	5	
0	0.010	0.038	0.060	0.047	0.018	0.003	$p(a_0, a_1 \mid \omega_{\mathcal{W}}{=}0.44)$
	0.0001	0.001	0.006	0.012	0.012	0.005	$p(a_0, a_1 \mid \omega_{\mathcal{S}}{=}0.67)$
	0.000	1.118	1.826	2.739	4.472 $^{(\mathcal{W},\mathcal{S})}$	Inf $^{(\mathcal{W},\mathcal{S})}$	$Z_{\mathcal{W}}$
	0.000	0.828	1.265	1.697	2.191 $^{(\mathcal{W},\mathcal{S})}$	2.828 $^{(\mathcal{W},\mathcal{S})}$	$Z_{\mathcal{S}}$
	0.000	1.012	1.506	1.963	2.467 $^{(\mathcal{W},\mathcal{S})}$	3.253 $^{(\mathcal{W},\mathcal{S})}$	$Z_{\mathcal{L}}$
	1.000	0.625	0.357	0.179	0.071 $^{(\mathcal{W},\mathcal{S})}$	0.018 $^{(\mathcal{W},\mathcal{S})}$	$P_{\mathcal{F}}$
1	0.023	0.090	0.141	0.111	0.043	0.007	$p(a_0, a_1 \mid \omega_{\mathcal{W}}{=}0.44)$
	0.001	0.009	0.035	0.072	0.073	0.030	$p(a_0, a_1 \mid \omega_{\mathcal{S}}{=}0.67)$
	-1.225	-0.409	0.191	0.763	1.433	2.449	$Z_{\mathcal{W}}$
	-1.380	-0.422	0.189	0.730	1.320	2.108	$Z_{\mathcal{S}}$
	-1.486	-0.417	0.189	0.736	1.327	2.276	$Z_{\mathcal{L}}$
	1.000	0.893	0.714	0.500	0.286	0.107	$P_{\mathcal{F}}$
2	0.018	0.070	0.111	0.087	0.034	0.005	$p(a_0, a_1 \mid \omega_{\mathcal{W}}{=}0.44)$
	0.002	0.018	0.072	0.146	0.148	0.060	$p(a_0, a_1 \mid \omega_{\mathcal{S}}{=}0.67)$
	-2.449	-1.433	-0.763	-0.191	0.409	1.225	$Z_{\mathcal{W}}$
	-2.108	-1.320	-0.730	-0.189	0.422	1.380	$Z_{\mathcal{S}}$
	-2.276	-1.327	-0.736	-0.189	0.417	1.486	$Z_{\mathcal{L}}$
	1.000	0.982	0.929	0.821	0.643	0.375	$P_{\mathcal{F}}$
3	0.005	0.018	0.029	0.023	0.009	0.001	$p(a_0, a_1 \mid \omega_{\mathcal{W}}{=}0.44)$
	0.001	0.012	0.049	0.099	0.100	0.041	$p(a_0, a_1 \mid \omega_{\mathcal{S}}{=}0.67)$
	-Inf	-4.472	-2.739	-1.826	-1.118	0.000	$Z_{\mathcal{W}}$
	-2.828	-2.191	-1.697	-1.265	-0.828	0.000	$Z_{\mathcal{S}}$
	-3.253	-2.467	-1.963	-1.506	-1.012	0.000	$Z_{\mathcal{L}}$
	1.000	1.000	1.000	1.000	1.000	1.000	$P_{\mathcal{F}}$

Table 2-3: Comparison of critical values, type 1 error, and P values computed for the Wald, chi square, LR, and Fisher's exact (nonrandomized, both unadjusted and adjusted) tests for sample sizes of $n_0 = 3$ and $n_1 = 5$, and observed outcome $(a_0, a_1) = (1, 5)$. The unadjusted Z critical value based on the normal approximation is 1.96 and the unadjusted P critical value based on the nominal level is 0.025.

		type 1 error				P value			
	critical value[1]	$\omega_{\mathcal{W}} = 0.44$		$\omega_{\mathcal{S}} = 0.67$			$\omega_{\mathcal{W}}$	$\omega_{\mathcal{S}}$	
Statistic	*Adj*	*Unadj*	*Adj*	*Unadj*	*Adj*	*Unadj*	*Adj*	*Adj*	P_B
Wald	(2.449, 4.472]	0.075	0.021	0.059	0.017	0.007	0.075	0.058	0.075
ChiSq	(2.108, 2.191]	0.028	0.021	0.047	0.017	0.018	0.028	0.046	0.046
LR	(2.276, 2.467]	0.075	0.021	0.059	0.017	0.011	0.028	0.046	0.046
Fish	[0.071, 0.107)	0.003	0.021	0.005	0.017	0.107	0.028	0.046	0.046

[1] Owing to the discreteness, any critical value within the interval gives identical results.

Table 2-4: Comparison of type 1 error computed for the Wald, chi square, LR, and Fisher's exact (nonrandomized, both unadjusted and adjusted) tests for nuisance parameter $\omega = p_0 = p_1 = 0.25$ based on 10^5 simulations to nominal level, according to randomization ratio for sample size of 30.

	sample size = 30							
	1:1 randomization				2:1 randomization			
nominal level	0.025		0.005		0.025		0.005	
type 1 error	*Unadj*	*Adj*	*Unadj*	*Adj*	*Unadj*	*Adj*	*Unadj*	*Adj*
Wald	0.030	0.021	0.009	0.003	0.063	0.024	0.035	0.002
Chi square	0.027	0.021	0.004	0.003	0.015	0.015	0.001	0.001
LR	0.029	0.021	0.009	0.003	0.040	0.024	0.012	0.002
Fisher's exact	0.009	0.021	0.001	0.003	0.006	0.015	0.0002	0.001

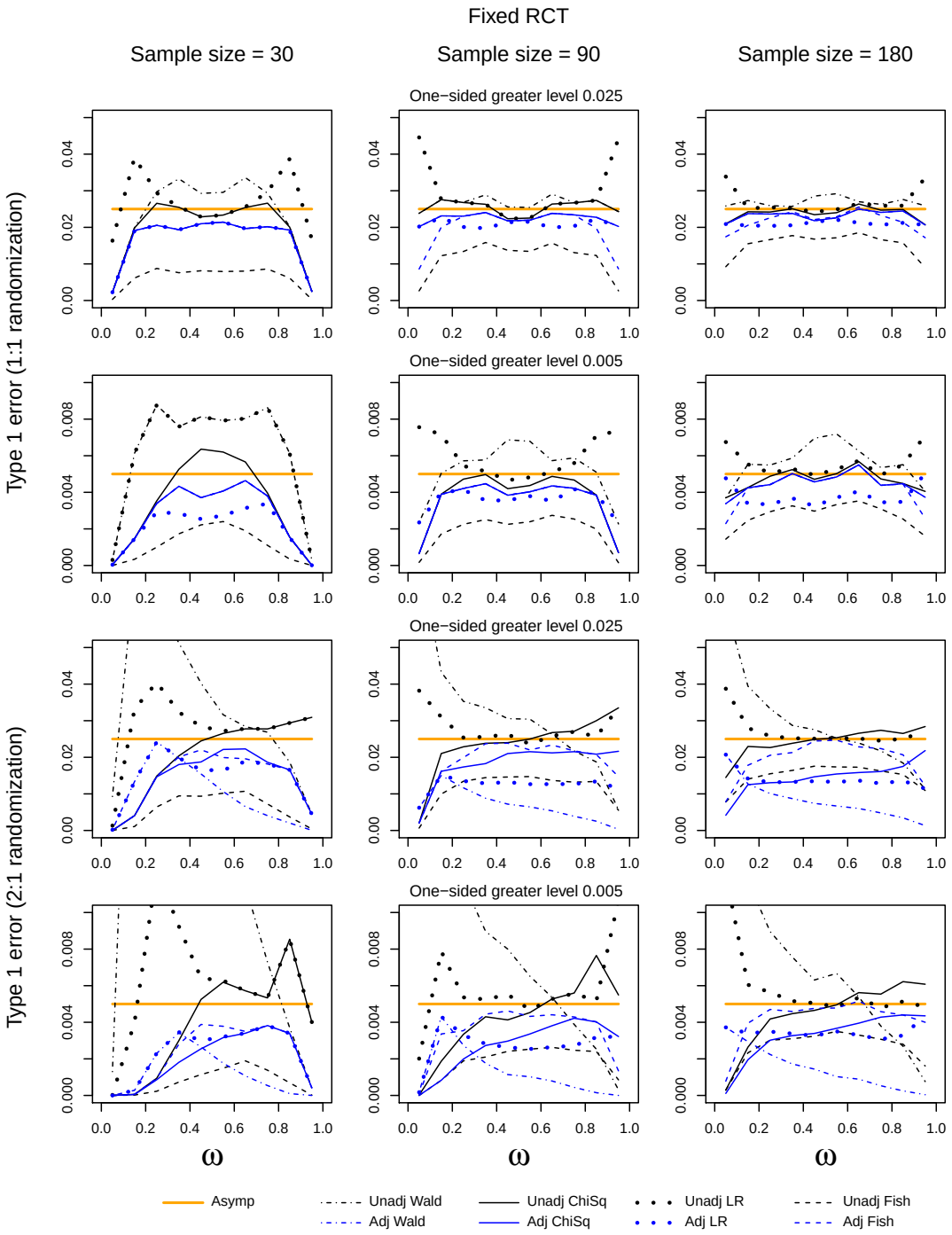


Figure 2-2: Consider FD with sample sizes of 30, 90, and 180, 1:1 and 2:1 randomization, and nominal levels of 0.025 and 0.005. Type 1 error is computed based on asymptotics (gold), unadjusted (black), and adjusted (blue) tests across a range of nuisance parameter ω from 10^5 sims. Note that the adjusted chi square and adjusted Fisher's sometimes overlap.

Chapter 3

BACKGROUND: GROUP SEQUENTIAL TESTS USING ASYMPTOTIC THEORY

In this chapter, we introduce group sequential RCTs by considering the one sample normal setting. Additionally, we explain what is "typically" done in sequential testing (use of asymptotics) and why we need a new approach/different option, namely the use of unconditional exact tests to (try to) handle the issues discussed in Chapter 2.

3.1 One sample normal setting

To motivate the group sequential setting, we consider the one sample setting (Emerson and Fleming, 1990; Kittelson and Emerson, 1999; Jennison and Turnbull, 2000; Emerson, Kittelson, and Gillen, 2007). Suppose the outcome of interest is independent among observations and comes from a normal distribution with mean μ and known variance σ^2:

$$Y_i \overset{indep}{\sim} \mathcal{N}(\mu, \sigma^2)$$

for observations $i = 1, 2, \ldots, N_J$, where N_J is the total number of observations if the RCT continues until the final analysis time. Without loss of generality, testing the one-sided hypothesis test (for known σ^2)

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_A : \mu > \mu_0$$

corresponds to "superiority" (efficacy) if $\mu > \mu_0$, "inferiority" (futility) if $\mu < \mu_0$, and "equivalence" (no difference) if $\mu = \mu_0$. Analysis times are indexed by $j = 1, 2, \ldots, J$. Sample sizes at each analysis time are cumulative

$$N_1 < N_2 < \cdots < N_J$$

for $j = 1, 2, \ldots, J$.

Now suppose you are interested in estimating the treatment effect at each analysis time. Since each subsequent analysis uses all prior and new observations, multiple comparisons is a problem. Four test statistics used to make inference about a treatment effect at the jth analysis are

- Partial sum statistic: $T_j \equiv \sum_{i=1}^{N_j} Y_i$;
- Sample mean statistic: $\overline{Y}_j \equiv T_j/N_j$;
- (Normalized) Z statistic: $Z_j \equiv \sqrt{N_j}(\overline{Y}_j - \mu_0)/\sigma$; and
- Fixed sample P value statistic: $P_j \equiv 1-\Phi(Z_j)$, where Φ is the cumulative distribution function of the standard normal distribution.

Using a standardized scale for treatment effect is common for group sequential design software. Define

$$X_i \equiv \frac{Y_i - \mu_0}{\sigma\sqrt{N_j}} \sim \mathcal{N}\left(\frac{\delta}{N_J}, \frac{1}{N_J}\right)$$

where $\delta \equiv \sqrt{N_J}(\mu - \mu_0)/\sigma$ denotes the standardized version of μ. The four standardized test statistics are

- Sample mean statistic: $\overline{X}_j \equiv \sqrt{N_J}\left(\frac{\overline{Y}_j - \mu_0}{\sigma}\right) \sim \mathcal{N}\left(\delta, \frac{1}{\Pi_j}\right)$, where $\Pi_j \equiv N_j/N_J$ is the proportion of statistical information (sample size) accrued by analysis time j;
- (Normalized) Z statistic: $Z_j \equiv \overline{X}_j\sqrt{\Pi_j} \sim \mathcal{N}(\delta\sqrt{\Pi_j}, 1)$;
- Fixed sample P value statistic: $P_j \equiv 1-\Phi(Z_j)$, where Φ is the cumulative distribution function of the standard normal distribution; and
- (Standardized) Partial sum statistic: $T_j \equiv \overline{X}_j\Pi_j \sim \mathcal{N}(\delta\Pi_j, \Pi_j)$.

After computing the test statistic for a given analysis time, we need to decide whether the RCT should continue or stop. Following Kittelson and Emerson (1999), in general, a continuation region $\mathcal{C}_j$ is defined as

$$\mathcal{C}_j = (a_j, b_j] \cup [c_j, d_j)$$

where critical values satisfy $-\infty \leq a_j \leq b_j \leq c_j \leq d_j \leq \infty$. In this thesis we focus on one-sided group sequential designs, so choose $b_j = c_j$, corresponding to continuation sets of the form $\mathcal{C}_j = (a_j, d_j)$. Additionally, we focus on symmetric group sequential designs that have equal type 1 and type 2 errors (Emerson and Fleming, 1989). Based on the definition of the continuation region appropriate for a specific test statistic $\{\overline{X}_j, Z_j, P_j, S_j\}$, we would stop an RCT at analysis time j if the value of the chosen test statistic is not in the continuation region appropriate for the chosen test statistic. The first such analysis time where the test statistic is outside of the continuation region, denoted by M. The sets $\{a_j\}_{j=1}^{J}$ and $\{d_j\}_{j=1}^{J}$ are termed the futility and efficacy boundaries, respectively, for analysis times indexed by $j = 1, 2, \ldots, J$. Although there are many boundary shape functions, we focus on the symmetric designs (Emerson and Fleming, 1989) using O'Brien-Fleming (1979) and Pocock (1977) boundary functions that span a spectrum of conservatism at the earliest analyses (Emerson and Fleming, 1989). The O'Brien-Fleming design represents the most commonly used efficacy boundary in RCTs, and the Pocock design is approximately optimal with respect to average sample size.

Pocock (1977) explored whether using boundaries derived in a one sample normal setting was valid on fixed sample P value (statistic) scale and found it reasonable. Mapping from this one sample normal setting to the two sample setting for binomial proportions is discussed in Whitehead (1997). In particular, in the two sample binomial setting with r:1 randomization, we have $\sigma^2 = \frac{p_1(1-p_1)}{r} - p_0(1-p_0)$. We note that it is not uncommon for software to calculate the asymptotic results such as power, based on "true" values of p_0 and p_1 when computing the variance: whether values of p_0 and p_1 are based on the null, alternative, or intermediate values, that "design" variance is used for the entire power curve regardless of values of θ. In this thesis, we make comparisons between Wald, chi square, LR, and Fisher's exact test using the fixed sample P value statistic for each. Therefore, we

define M as

$$M \equiv \min_{1 \leq j \leq J} \{j \ : \ P_j \notin \mathcal{C}_j\}.$$

Recall δ denotes the standardized version of μ (parameter of interest). The statistic (M, P_M) is minimal sufficient for δ, with the sampling distribution computed numerically (Armitage, McPherson, and Rowe, 1969).

We consider frequentist operating characteristics for evaluation of group sequential RCTs (Emerson, Kittelson, and Gillen, 2007), including type 1 error, power, error spending function, and average sample number (ASN). We vary the boundary shape, randomization ratio, and sample sizes.

Group sequential designs allow for interim analyses for efficacy and futility. In fixed RCT designs, the power function is a function of the parameter of interest, say θ. Two operating characteristics of fixed RCTs include the power function evaluated under the null hypothesis of θ_0 (type 1 error) and under any arbitrary value (statistical power). Analogously in group sequential designs, the stopping probability is the probability under θ, be it the null value θ_0 or any arbitrary value, that the chosen test statistic is at least as extreme as the critical value. As previously discussed, the critical values for one-sided group sequential designs are the boundaries of the continuation region $\mathcal{C}_j = (a_j, d_j)$. For one-sided greater hypothesis tests, a_j is the futility (i.e., lower) boundary and d_j is the efficacy (i.e., upper) boundary at analysis time j. For example, if we consider the (normalized) Z statistic, the upper stopping probability under arbitrary θ at analysis time j is $\Pr(Z_j > d_j \,|\, \theta)$. The type 1 error spending function is defined as the cumulative stopping probability at analysis time j

$$\sum_{k=1}^{j} \Pr(Z_k > d_k \,|\, \theta_0).$$

Note that under θ_0 the error spending function is the type 1 error "spent" to perform the j-th interim analysis (i.e., to have a look or another look at the interim data during the course of the study). The cumulative upper stopping probability under θ_0 up to and including the final planned analysis represents the overall type 1 error. The statistical power to detect the design alternative corresponds to the cumulative upper stopping probability under design alternative θ_A.

3.2 Naïve approach: Use asymptotic results with constant mean-variance relationship

In the AML trial, asymptotic results based on the chi square test were used with confidence intervals computed ignoring the impact the mean-variance relationship should have had as different values of θ were considered. Although the total sample size was "reasonably" large in the study, a potential problem in all group sequential RCTs is that of having smaller samples at earlier analyses. That is, even if the total sample size at the final analysis is large enough for asymptotic results to be a good approximation, sample sizes at interim analyses may not be large enough, potentially leading to inflation of the type 1 error. We mentioned using a constant mean-variance relationship with the asymptotic result for fixed samples in Chapter 2. When computing asymptotic results in this thesis, we use the intermediate values of p_0 and p_1 between the null and alternative values as the truth.

In the remainder of this section, we illustrate how using asymptotic results does not appear appropriate in sequential testing with a binary outcome, thus motivating the need for an alternative approach, namely unconditional exact tests. We will then compare the unconditional exact tests to asymptotic results to investigate any important differences according to the operating characteristics discussed in Chapter 2.

3.2.1 Fixed sample versus group sequential setting: Critical values and P values

Adjusting the critical value in fixed sample tests changes the realized type 1 error.

- In the case of Wald, chi square, and LR tests which can be anti-conservative, the realized type 1 error after adjustment is conservative because the adjustment is guaranteeing the type 1 error does not exceed the nominal level. From this process, the P value associated from the unadjusted test changes when using the adjusted test: since the unconditional exact test maximizes across the entire range of nuisance parameter values while ensuring that the worst case type 1 error does not exceed the nominal level, the adjusted test's P value is larger than the corresponding unadjusted test's P value.

- On the other hand, Fisher's exact test is always conservative when nonrandomized as discussed in Chapter 2. The realized type 1 error is still conservative after adjustment, however, the degree of conservatism is lower. Using the unconditional exact test with Fisher's exact test statistic the associated P value is smaller than the unadjusted test's P value.

How these tests behave after adjustment in the fixed sample setting affects inference in the group sequential setting with respect to: (1) stopping a RCT for efficacy; (2) stopping a RCT for futility; or (3) continuing a RCT to the next analysis. In the following example RCT, we want to illustrate how the decision based on an interim analysis may change according to the type of test and adjustment used. Since the Wald, chi square, and LR tests are asymptotically equivalent and behave in a similar manner as described above, we will restrict attention in this section to only the unadjusted and adjusted chi square tests out of those three. Since the impact of adjustment on Fisher's exact test behaves in the opposite way to the other three tests in the fixed sample setting, we will also compare unadjusted and adjusted Fisher's exact tests.

In the fixed sample setting, each analysis consists of one critical value and a corresponding P value. In the group sequential setting, this thesis focuses on designs with continuation regions where a_j is the futility critical value, d_j is the efficacy critical value, and $b_j = c_j$ indexed by $j = 1, ..., J$. So for each analysis prior to the Jth, instead of one critical value as in the fixed sample setting, there are two critical values a_j and d_j.

We consider an example RCT design to illustrate a potential problem that arises in group sequential designs which is dependent on which test is used. Suppose we design a group sequential RCT with maximal sample size of 240 individuals, a maximum of 4 analyses, and a symmetric Pocock boundary for efficacy and futility to test the null hypothesis that $p_1 = p_0 = 0.3$ against the alternative hypothesis that $p_1 = 0.6$ and $p_0 = 0.3$, a difference in proportions $\theta_A = 0.3$. Table 3-1 lists the efficacy and futility boundaries for each interim analysis according to the group sequential design we specified above. Boundary scales included in the table are the sample mean, partial sum, and fixed P value scales.

Table 3-1: Group sequential RCT with symmetric Pocock efficacy and futility boundaries for maximum of 4 analyses, maximal sample size of 240 with 1:1 randomization assignment. Boundary scales include sample mean, partial sum, and fixed P value.

	Symmetric Pocock boundary (P=0.5)					
	Futility			Efficacy		
N_j	Sample mean	Partial sum	Fixed P	Sample mean	Partial sum	Fixed P
60	0.000	0.00	0.5000	0.287	8.61	0.0101
120	0.084	5.04	0.1680	0.203	12.17	0.0101
180	0.121	10.91	0.0445	0.166	14.90	0.0101
240	0.143	17.21	0.0101	0.143	17.21	0.0101

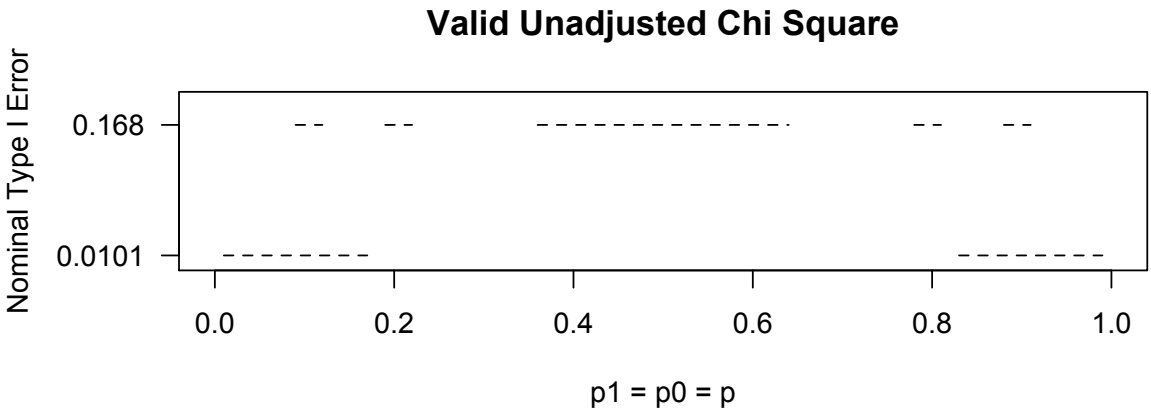


Figure 3-1: For sample sizes of $n_1 = n_0 = 60$ and equal proportions, the dashed lines correspond to valid tests, defined as those for which the respective realized type 1 error does not exceed the specified nominal level (either 0.0101 or 0.1680). Note that discreteness of the outcome space can lead to very different patterns as level of significance varies.

Figure 3-1 shows the proportions $p_1 = p_0$ for which the unadjusted chi square test produces a valid type 1 error: The realized type 1 error does not exceed the specified nominal level. The nominal levels of 0.0101 and 0.1680 were chosen as they correspond to the second interim analysis efficacy and futility fixed P value boundaries for the symmetric Pocock design in Table 3-1. From Figure 3-1 we find that for nominal level of 0.0101 corresponding

to the second analysis Pocock efficacy boundary, proportions $p_1 = p_0$ between 0.17 and 0.83 yield invalid type 1 error for the unadjusted chi square test. The range of valid proportions differs for the nominal level of 0.1680 corresponding to the second analysis Pocock futility boundary. These findings suggest the need for adjusted tests.

We want to compare the unadjusted and adjusted tests for both the chi square and Fisher's exact test. The asymptotic tests (Wald, chi square, and LR) can be either conservative and anti-conservative. Since unconditional exact tests have to consider the worst case realized type 1 error across nuisance parameters, the adjusted version of any of the asymptotic tests will yield conservative inference corresponding to higher P value. On the other hand, Fisher's exact test is conservative when implemented as a nonrandomized test. So the unconditional exact test using the adjusted Fisher's exact test statistic will yield less conservative inference corresponding to lower P value. The differences in what happens when we use adjusted versions of tests play an important role in sequential testing: decisions to continue studies, stop for efficacy, or stop for futility are based on the two critical values as well as the type of adjustment we make. We will illustrate how decisions may be altered according to the test including adjustment used.

We plan to show how adjustment using the unconditional exact test cannot only give different decisions compared to the corresponding unadjusted test, but also adjustments to different tests such as chi square and Fisher's exact test can yield different decisions. We shall show such an instance with the group sequential design specified above. We consider a Pocock design because its boundaries correspond to less extreme P values early on when the sample sizes are small than do the boundaries for the O'Brien-Fleming designs, and we thus anticipate worse behavior for a Pocock design.

Figure 3-2 plots the group sequential design we specified above with symmetric Pocock efficacy (lower on fixed P value scale) and futility (upper on fixed P value scale) boundaries which were also listed in Table 3-1. For illustration, we used the specified group sequential design and corresponding boundaries to identify which observed outcomes (a_0, a_1) cells help illustrate issues of concern. We will consider the second analysis time to conduct formal analysis. Based on the chosen statistics, we (or the DSMB) will have to decide whether to continue the RCT, or either stop the trial for an efficacious treatment benefit or for futility.

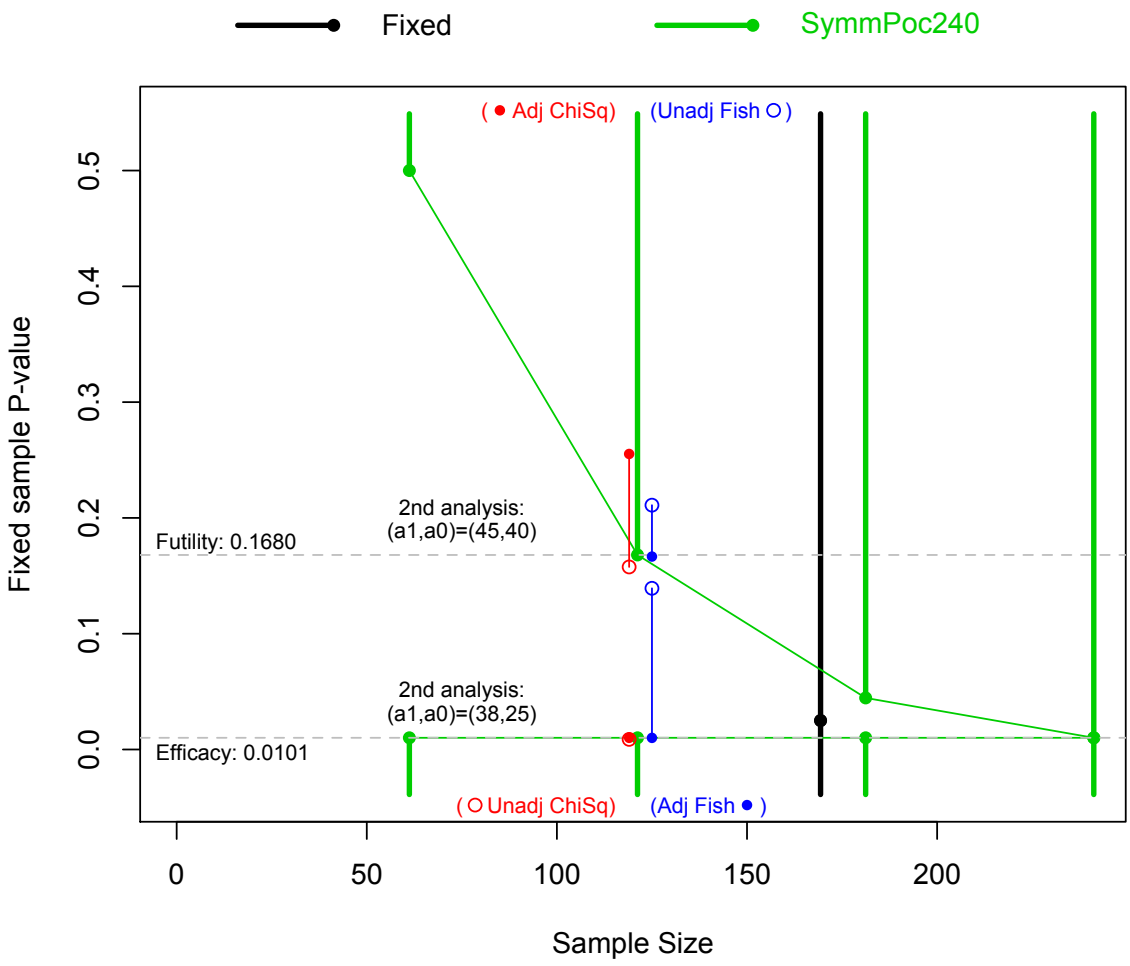


Figure 3-2: The effect of adjustment on the P values corresponding to two different outcomes: Decisions about stopping or continuing can differ according to the statistic used as the basis of an unconditional exact test. (See text)

For the second interim analysis where $n_1 = n_0 = 60$ the unadjusted and adjusted P values based on either chi square or Fisher's exact test are shown for two possibilities for observed outcomes (a_0, a_1): $(45, 40)$ and $(38, 25)$. If $(45, 40)$ is observed, the adjusted chi square P value would alter the decision made as compared to using the unadjusted chi square P value: Stop RCT for futility instead of continuing. Alternatively for the same observed

outcome, the adjusted Fisher's exact P value alter the decision made without adjustment in the opposite way of adjusting the chi square test: Continue RCT instead of stopping for futility. If $(38, 25)$ is observed, using either the unadjusted or adjusted chi square P values yield the same decision: Stop RCT for efficacy. However, using the adjusted Fisher's exact P value results in a different decision as compared to using the unadjusted Fisher's exact P value: Stop RCT for efficacy instead of continuing. The purpose of this example was to show how decisions at interim analyses can change depending on both the test and whether adjustment was used.

3.2.2 Evaluating group sequential designs

We consider a group sequential RCT with maximal number of analyses $m = 4$, 1:1 randomization ratio, power approximately 80% to discriminate between a null hypothesis of no difference in proportions and alternative hypothesis of $\theta_A = \theta_{Dsn} = 0.2$, yielding a total sample of 180 individuals (90 individuals per arm). The following Figure 3-3 shows the group sequential design with symmetric O'Brien-Fleming monitoring guidelines as well as the corresponding fixed sample design for illustrative purposes.

We focus on evaluating RCTs according to frequentist operating characteristics. First, we consider the statistical power as a function of the true difference in proportions. Figure 3-4 contains a plot of the power curves according to the group sequential designed specified earlier: (1) power computed based on asymptotics in red; (2) power computed based on 100,000 simulations in green; and (3) power computed based on fixed design in black. We notice a couple things:

- the fixed RCT has slightly higher power overall; and
- despite the power based on simulating RCTs and using the chi square test (similar to the AML trial) being the lowest among the three, these differences may not be material.

Another way to consider the potential "benefit" of using a group sequential design is to examine the expected sample size, or average sample number (ASN)–one of the group

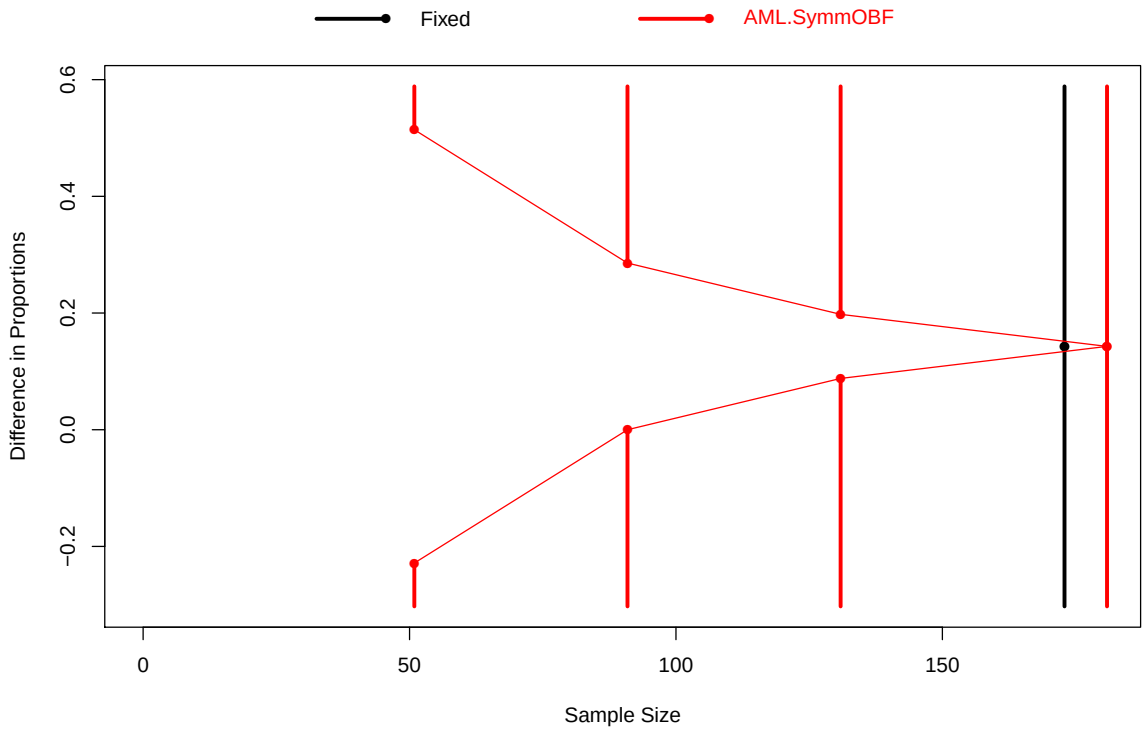


Figure 3-3: Comparison of FD (black) and the AML trial GSD with symmetric OBF (red) with maximum of 4 analyses and sample size of 180 with 1:1 randomization, 80% power to detect difference in proportions of 0.2, and variances computed based on nuisance parameter $\omega = 0.5$ for one-sided greater level 0.025 hypothesis test.

sequential operating characteristics. Figure 3-5 contains plots of ASN and 75th percentile of the sample size distribution as a function of the true difference in proportions. Red curves are based on using asymptotic results, green curves are based on 100,000 simulations and using the chi square test, and black curves are based on the fixed sample design. We notice that the ASN ranges from about 100 to 150 (individuals) for the group sequential design (both asymptotically and from simulations). However, we do find from simulations that the chi square test estimates ASN of approximately 140 (70 per arm). Although not markedly different (asymptotic versus simulated), agreement is not exact.

Another group sequential operating characteristic to consider is the stopping probability at each analysis for a range of differences in proportions. Figure 3-6 contains plots of

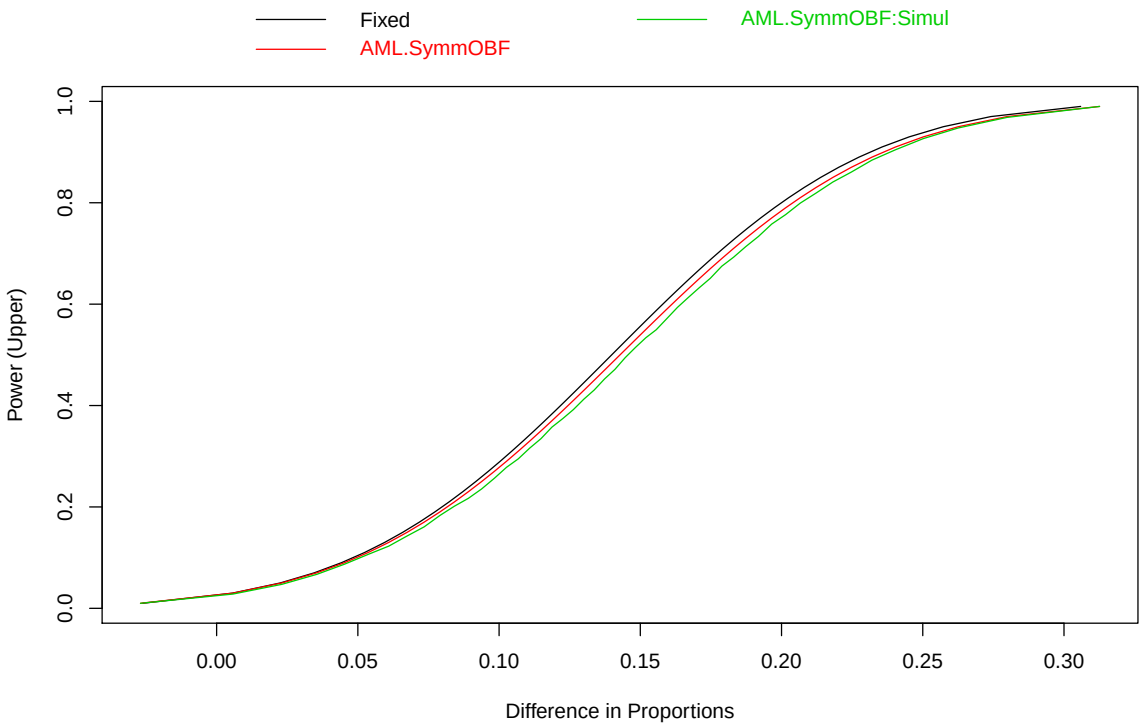


Figure 3-4: Consider FD and the AML trial GSD with symmetric OBF with maximum of 4 analyses and sample size of 180 with 1:1 randomization, 80% power to detect difference in proportions of 0.2, and nuisance parameter $\omega = 0.5$ for one-sided greater level 0.025 hypothesis test. Power (upper stopping probability) for range of differences in proportions is computed based on FD (black) and on GSD (asymptotic results in red and unconditional exact test using unadjusted chi square test statistic in green from 10^5 simulations). Comparisons between three designs are based on vertical separation.

the simulated (based on 100,000 simulations) stopping probabilities according to whether stopping for futility (lower on sample mean scale) or efficacy (upper on sample mean scale) and the stopping probability contribution at each analysis time–at the first analysis, at the second analysis given the trial continued after the first analysis, and so on until the final analysis, where the trial stops regardless of the value of the estimated treatment effect. Additionally, for comparison, the blue dotted/dashed and solid lines correspond to the asymptotic stopping probabilities for futility (lower) and efficacy (upper), respectively, at each analysis. From the following figure, we find that simulated stopping probabilities do

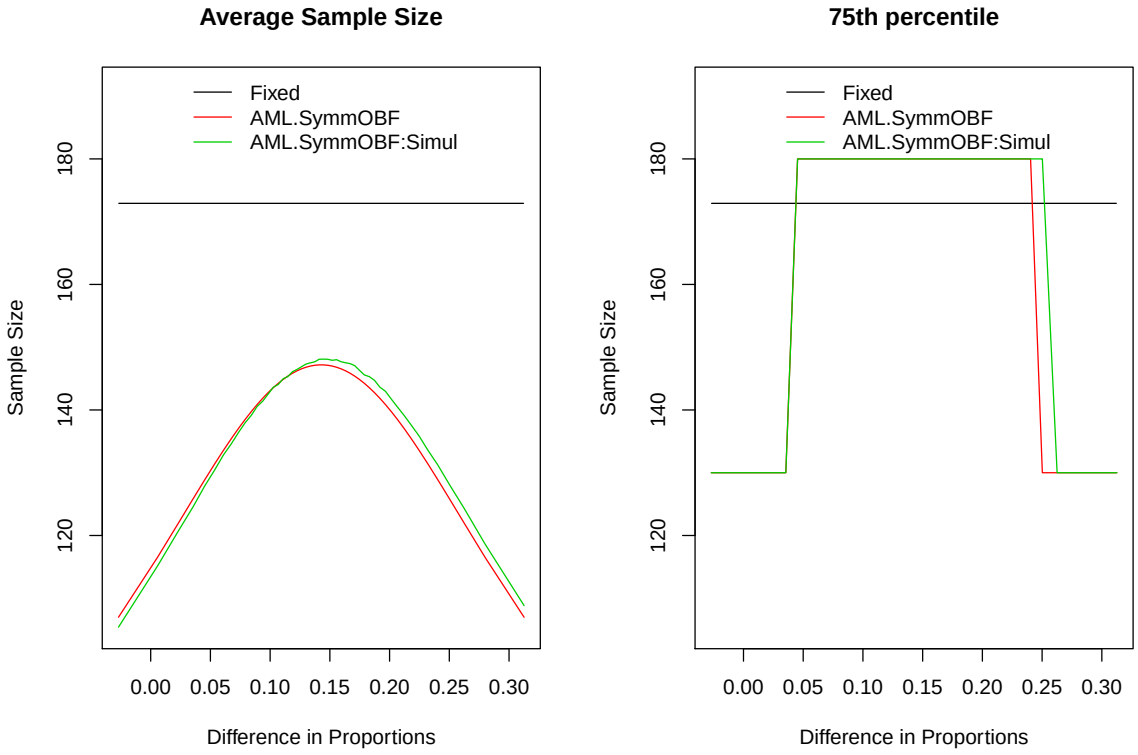


Figure 3-5: Consider FD and the AML trial GSD with symmetric OBF with maximum of 4 analyses and sample size of 180 with 1:1 randomization, 80% power to detect difference in proportions of 0.2, and nuisance parameter $\omega = 0.5$ for one-sided greater level 0.025 hypothesis test. Expected sample size (or average sample number, ASN) and 75th percentile of sample size distribution for range of differences in proportions are computed based on FD (black) and on GSD (asymptotic results in red and unadjusted chi square test in green based on 10^5 simulations).

not agree with asymptotic stopping probabilities for every difference in proportion.

Note that results for the group sequential design, mirroring the design of the AML trial, are based on moderate sample sizes at earlier interim analyses: the first formal interim analysis had 45 individuals per arm. Although we did not find markedly different results for our most recent example, based on the issues discussed in the previous section where the choice of test and whether adjustment is performed could lead to different decisions, consideration of using unconditional exact tests is warranted. In particular, our use of the conservative O'Brien-Fleming boundaries may protect against marked differences. In

Chapter 4 we discuss three alternative approaches of adjustment when implementing an unconditional exact test in small to moderate samples, and these are examined in a broader set of scenarios.

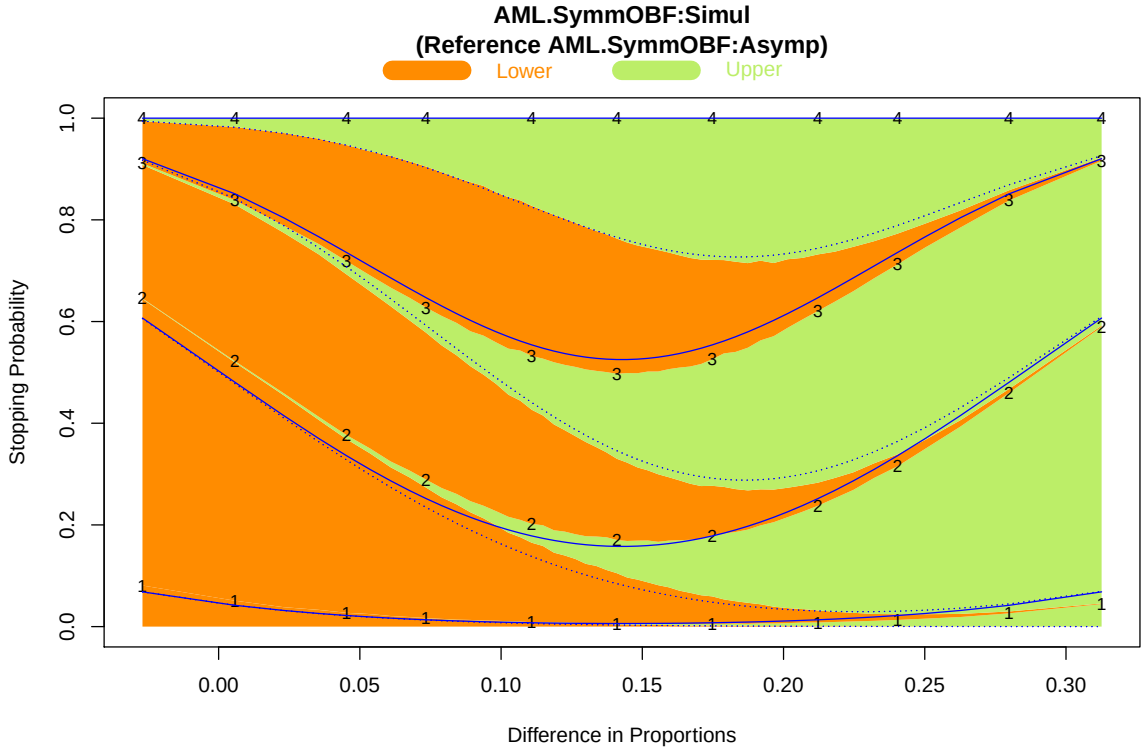


Figure 3-6: Consider the AML trial GSD with symmetric OBF with maximum of 4 analyses and sample size of 180 with 1:1 randomization, 80% power to detect difference in proportions of 0.2, and nuisance parameter $\omega = 0.5$ for one-sided greater level 0.025 hypothesis test. Stopping probability (lower/futility in orange and upper/efficacy in green for unconditional exact test using unadjusted chi square test statistic) at each analysis time for range of differences in proportions are computed based on 10^5 simulations). Asymptotic results (blue solid and dashed lines) serve as a reference. The discrepancies between the blue lines and the shaded regions are evidence that asymptotic results are not accurate.

Chapter 4

TESTS OF BINOMIAL PROPORTIONS IN SEQUENTIAL TESTING

In this chapter, we describe three methods of using critical values derived from adjusted fixed sample unconditional exact tests in group sequential designs having small to moderate sample sizes. We then evaluate the three alternative approaches using the Wald, chi square, LR, and Fisher's exact test statistics discussed in Chapter 2. After evaluating frequentist operating characteristics associated with the use of adjusted critical values in group sequential designs we make a recommendation for the approach to use in practice.

4.1 Alternative approaches: Three methods of using critical values derived from adjusted fixed sample unconditional exact tests

Recall, in Chapter 2 we defined adjustment of fixed sample tests as choosing the critical value that minimizes the conservativeness of the actual type 1 error without exceeding the nominal level of significance. We suggest three methods of using critical values derived from adjusted fixed sample tests to determine the rejection region of the outcome space when testing binomial proportions in a group sequential setting:

1. at the final analysis time only;

2. at analysis times after accrual of $> 50\%$ of the maximal sample size ($\Pi_j > 0.5$); and

3. at every analysis time.

Alternative approach 1 is considered in the hopes that any departures from the asymptotic error spending function at interim analyses can be corrected by adjustment of the critical value at the final analysis to achieve the correct overall type 1 error. Alternative approach 2 is a slightly more complex adjustment. The intuition behind choosing to use adjusted critical values when the proportion of maximal sample size $\Pi_j > 0.5$ is that at those later

times we attempt to correct for earlier departures better since we have larger sample sizes. Last, we suggest alternative approach 3 if the other two approaches do not guarantee overall type 1 error. This third approach uses adjustment at every analysis should prevent anti-conservative overall type 1 error. We do note, however, that this may not be equivalent to an approach that finds an adjusted group sequential critical value, which would be a more cumbersome implementation.

4.2 Evaluation of alternative approaches: Identifying a recommended approach to use in clinical practice

4.2.1 Outline for identifying a recommendation

We plan to evaluate the adequacy of using adjusted critical values 1) at the final analysis time, 2) at analysis times after accrual of more than 50% of the maximal sample size, and 3) at every analysis, derived from fixed sample unconditional exact tests based on the Wald, chi square, LR, and Fisher's exact tests' statistics in sequential analyses. In order to provide a recommendation based on our evaluation, we will compare the three alternative approaches to the four unadjusted tests as well as asymptotic results. Our goal is to demonstrate which approaches are suitable to use in the group sequential setting with small to moderate samples. The following list consists of the steps we plan to take in order to arrive at a recommended test for use in clinical practice. Note that all evaluations are based on one-sided greater tests, and we consider maximal samples sizes of 96 and 192 with up to a maximum of 4 analyses with symmetric group sequential boundaries.

1. We start with group sequential designs using 1) O'Brien-Fleming (early conservatism) and 2) Pocock (approximately efficient) boundaries, with 1:1, 2:1, and 3:1 randomization assignments of treatment and control. We consider all three adjustment approaches for each of the four test statistics as well as the four unadjusted tests. We will use asymptotic results as a global reference. We want to see if we can remove any of the candidate adjustment approaches from consideration among those, if any, that yield anti-conservative tests according to the overall type 1 error.

2. After ruling out any of the candidate adjustment approaches based on examining

the overall type 1 error, among those approaches that remain we plan to examine the overall power, including difference in power (simulated – asymptotic), for each test and remaining adjustment approach, and compare to the asymptotic results. Although we want to use tests that achieve the nominal level with the least amount of conservatism, we additionally want to use tests which have the most power.

3. At this point, we may have a reasonable idea for which of the candidate approaches we may want to consider using in practice if we were interested in the fixed sample setting. However, further examination is warranted of the additional operating characteristics inherent to the group sequential setting: stopping probabilities and error spending function; and sample size distribution: average sample number (ASN) and 75th percentile. We also want to assess whether there is agreement between the adjusted tests and asymptotic results.

4. Final consideration is understanding the efficiency of the group sequential design based on the chosen adjusted test to be used for analysis. We are interested in learning how the use of adjusted P values in the group sequential setting might alter the guidance provided by asymptotic results. In particular, we are interested in whether the results previously reported for the relative efficiency of particular boundary shapes (Emerson and Fleming, 1989) are different with adjusted inference.

4.2.2 Evaluation: overall type 1 error

Figures 4-1 and 4-2 contain plots of type 1 error similar to plots found in Chapter 2: Figure 4-1 focuses on group sequential designs with symmetric O'Brien-Fleming boundaries and Figure 4-2 focuses on group sequential designs with symmetric Pocock boundaries. In all scenarios, there are a maximum of 4 analyses, randomization assignments include 1:1, 2:1, and 3:1, and maximal sample sizes of 96 and 192. From these figures we find the following:

- As expected, anti-conservative and conservative type 1 error tended to lessen respectively toward the nominal level as maximal sample size increased from 96 to 192 regardless of group sequential design.

- Among O'Brien-Fleming designs, the unconditional exact tests with adjustment 1) at analyses after accrual of 50% of the maximal sample size and 2) at every analysis prevented anti-conservative type 1 error irrespective of the test statistic used.

- Among Pocock designs, adjustment at analyses after accrual of 50% of the maximal sample size did not prevent anti-conservative type 1 error as such an adjustment approach did with O'Brien-Fleming designs. Only adjustment at every analysis guaranteed prevention of anti-conservative type 1 error for Pocock designs.

We note that for these group sequential designs with 1:1 randomization, the number of individuals per arm are (12, 24, 36, 48) for maximal sample size of 96 and (24, 48, 72, 96) for maximal sample size of 192. When we refer to small to moderate samples in the group sequential setting, we consider the number of observations per arm at each analysis in addition to the maximal sample size, since methods of analysis used at earlier formal interim analyses may not be large enough for asymptotic results to be valid. The smallest samples sizes per arm and analysis occur with 3:1 randomization where the number of individuals in the control are (6, 8, 12) for maximal sample size of 96 and (12, 16, 24) for maximal sample size of 192.

Since our first objective is to eliminate anti-conservative inference, we remove the first two adjustment approaches 1) at final analysis only and 2) at analyses when $\Pi_j > 0.5$ from consideration as an improvement to the naïve approach using asymptotics since there were instances where the overall type 1 error exceeded the nominal level of significance when considering both O'Brien-Fleming and Pocock designs.

Before proceeding to the next section, we provide some comments regarding the tests adjusted at every analysis:

- For group sequential designs following 1:1 randomization, unconditional exact tests using adjusted Wald and chi square test statistics are closest to the nominal level. Fisher's exact test tends to be most conservative in O'Brien-Fleming designs whereas LR tends to be most conservative among adjusted tests in Pocock designs.

- For group sequential designs following either 2:1 or 3:1 randomization, the type 1 error of unconditional exact test using adjusted Fisher's exact test statistic is markedly closer to the nominal level than that for the adjusted Wald, chi square, and LR.

When randomization ratio is not 1:1, we see separation in overall type 1 error performance among the four test statistics. The adjusted Fisher's exact test statistic appears to be handling the discreteness from unequal sample sizes better, as its overall type 1 error behaves quite well across the range of values of nuisance parameter ω. Adjusted LR did not look promising even under 1:1 randomization, and based on the performances of adjusted Wald and adjusted chi square, these three test statistics do not appear to be viable options across a variety of group sequential designs. Under different randomization ratios, these findings are consistent whether using O'Brien-Fleming's early conservatism or Pocock's nearly efficient boundaries.

In the following section we will examine the power of the adjusted tests using the Wald, chi square, LR, and Fisher's exact test statistics, and try to identify which adjusted tests have the most power, and thus which adjusted tests we may want to recommend for analysis.

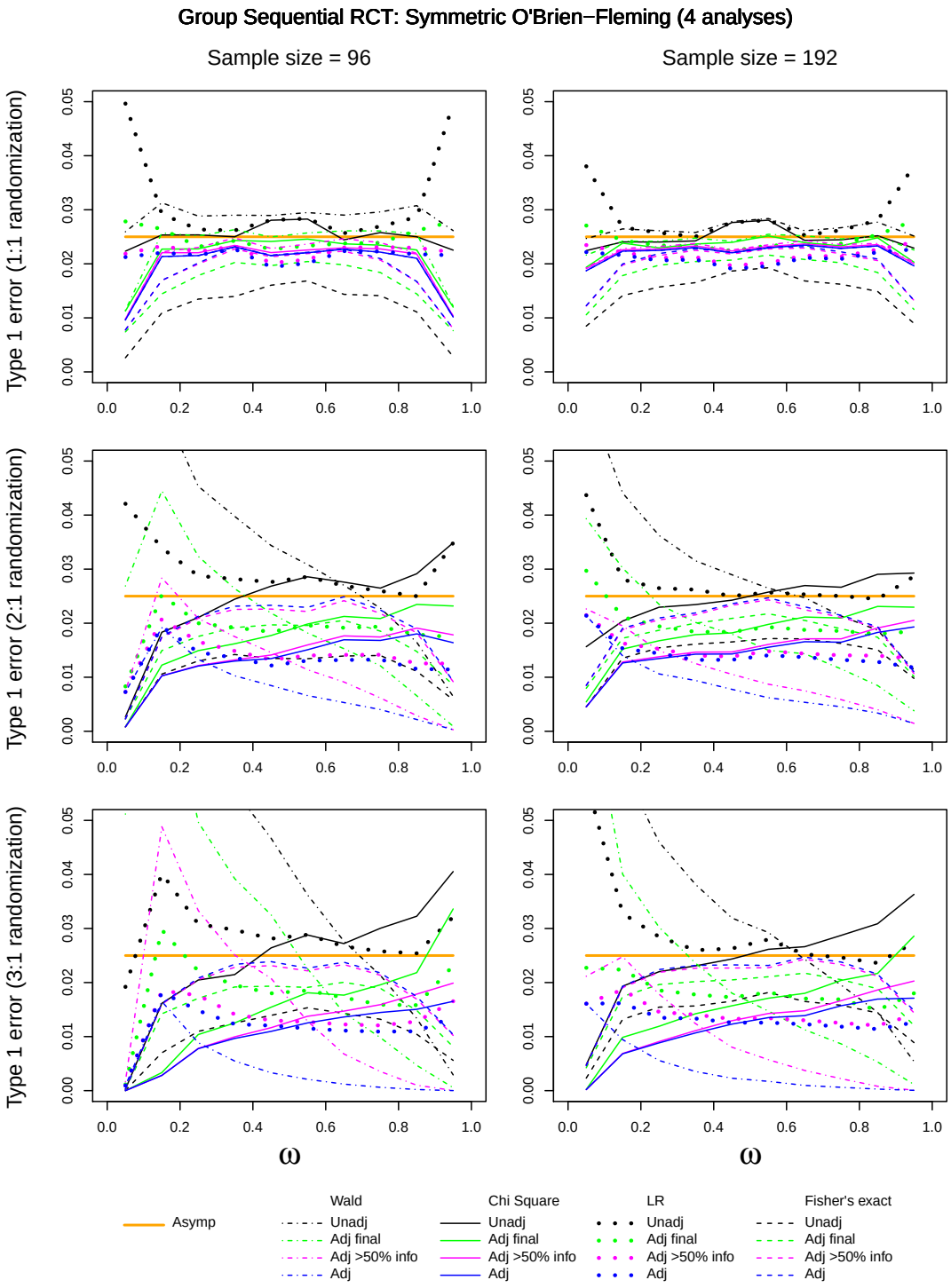


Figure 4-1: Symm OBF GSDs (up to 4 analyses) for one-sided greater 0.025 nominal level for specified sample size and randomization ratio. Type 1 error is computed based on asymptotics, unadj, and adj tests across a range of nuisance parameter ω from 10^5 sims.

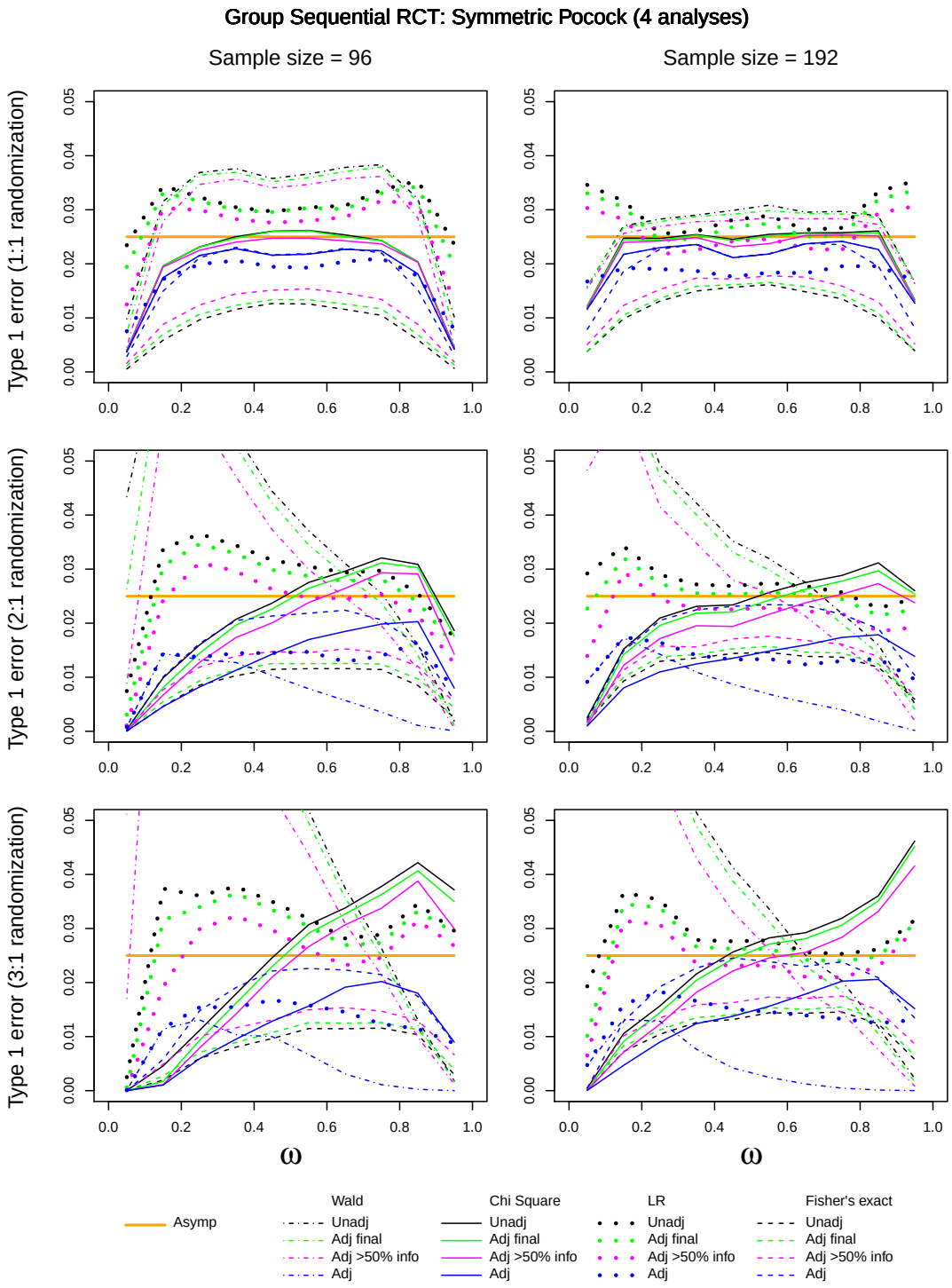


Figure 4-2: Symm Pocock GSDs (up to 4 analyses) for one-sided greater 0.025 nominal level for specified sample size and randomization ratio. Type 1 error is computed based on asymptotics, unadj, and adj tests across a range of nuisance parameter ω from 10^5 sims.

4.2.3 Evaluation: overall power

We consider the next important operating characteristic: power. As mentioned in hypothesis testing theory, a way to determine "optimal" tests is to choose the test with most power among those tests which achieve a particular nominal level. We want to identify which of the adjusted tests, based on Wald, chi square, LR, and Fisher's exact test statistics, tend to have the most power in ideally all, but if not then, in most settings. We will also consider several different design alternatives. We note that since we are focused on small to moderate sample sizes, in these settings trying to detect design alternatives which are quite small will correspond to a test having low power: we see this even when the design alternative is a 10% difference in the proportions with outcome between two arms.

We examine power and difference in power (simulated – asymptotic) for the designs considered in Figures 4-1 and 4-2 with Figures 4-3 to 4-10, where figures alternate between power and difference in power. From these figures we find the following:

- We find that in each of the group sequential designs considered with 1:1 randomization, the unconditional exact tests using the adjusted critical value at every analysis for Wald, chi square, LR, and Fisher's exact test statistics, respectively, all have approximately the same power for design alternative 0.3 and 0.5. For these design alternatives, differences relative to asymptotic results under 1:1 randomization indicate that the adjusted Fisher's exact test statistic yields the most power for lower values of ω and the chi square adjusted test has highest power for the higher values of ω.

- For group sequential designs with 2:1 and 3:1 randomizations, the adjusted Fisher's exact statistic yields markedly higher power than the adjusted chi square, LR, and Wald, respectively in terms of least to most.

From these findings, we will focus on comparing adjusted chi square and Fisher's exact tests.

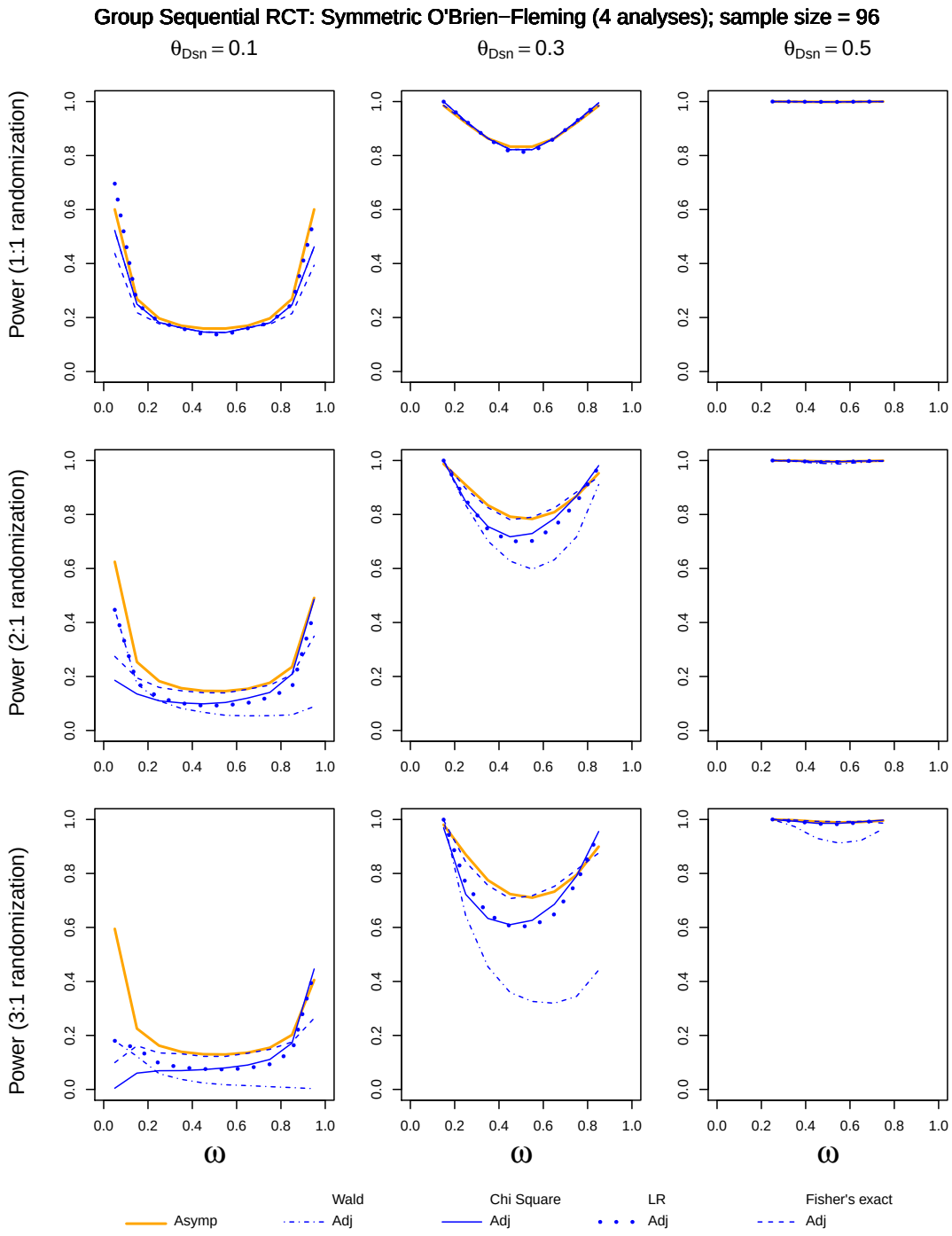


Figure 4-3: Symm OBF GSDs (up to 4 analyses, sample size of 96) for one-sided greater 0.025 nominal level for specified randomization ratio. Power is computed based on asymptotics and adjusted tests across a range of nuisance parameter ω from 10^5 sims.

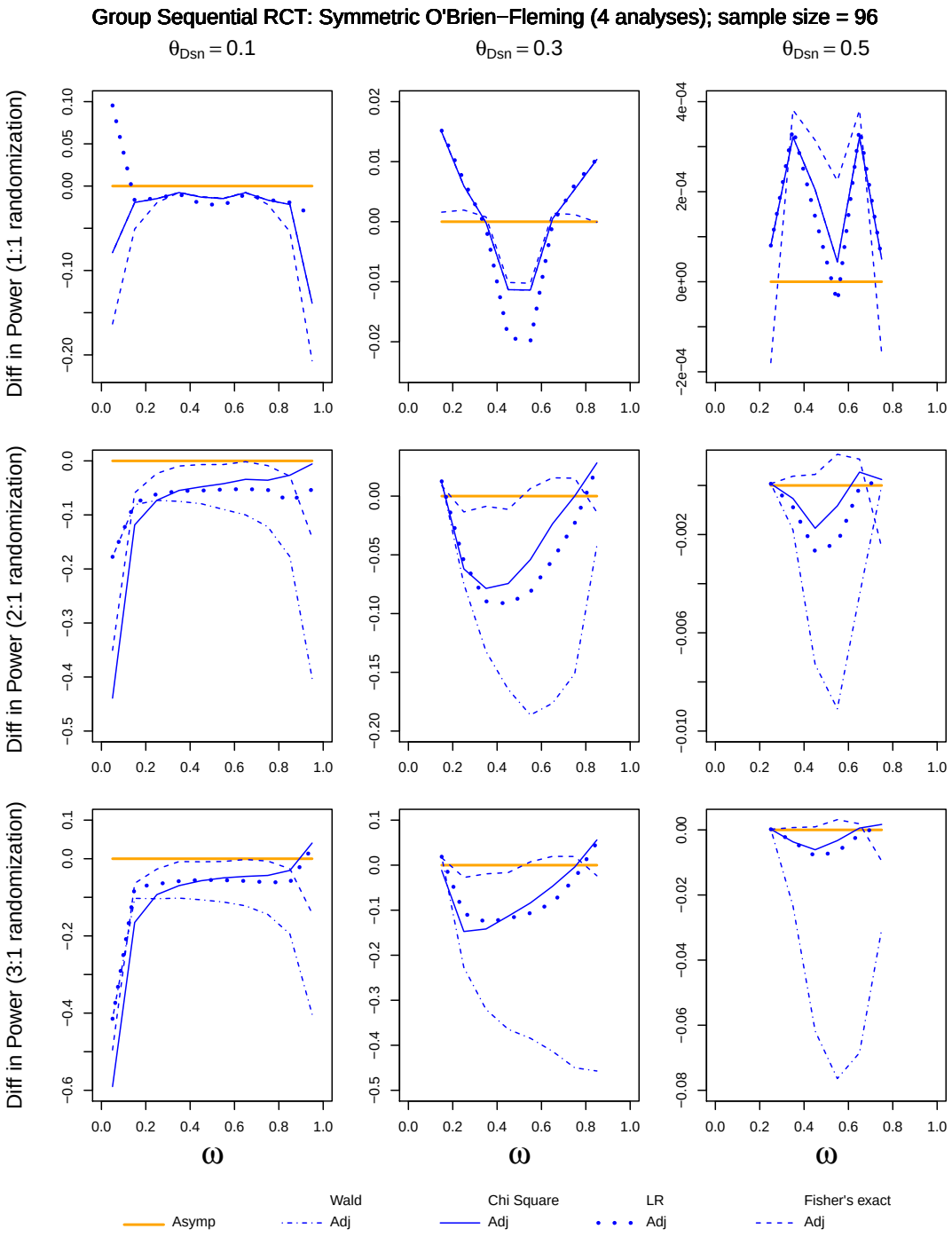


Figure 4-4: Symm OBF GSDs (up to 4 analyses, sample size of 96) for one-sided greater 0.025 nominal level for specified randomization ratio. Diff in power (simulated – asymp) is computed based on adjusted tests across a range of nuisance parameter ω from 10^5 sims.

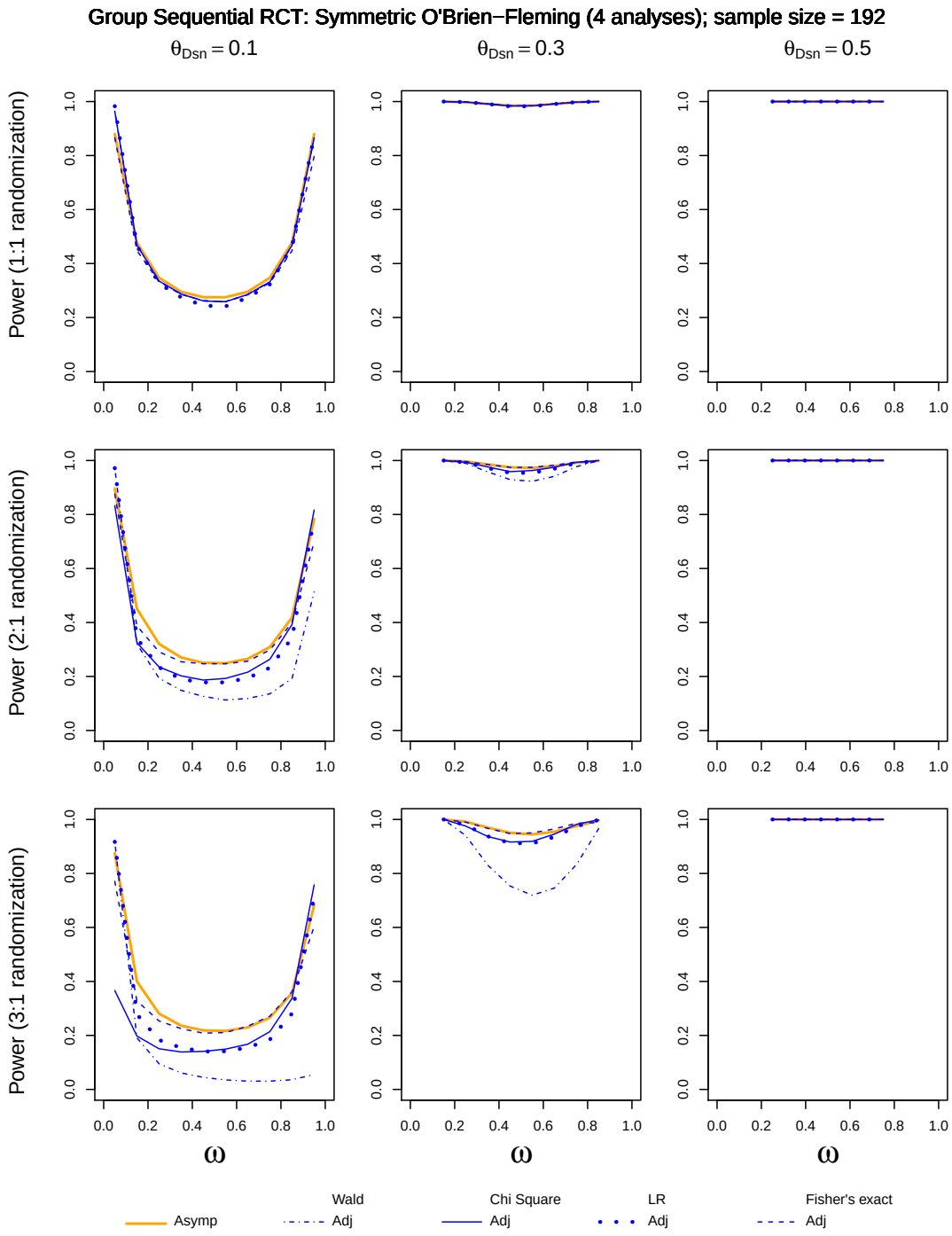


Figure 4-5: Symm OBF GSDs (up to 4 analyses, sample size of 192) for one-sided greater 0.025 nominal level for specified randomization ratio. Power is computed based on asymptotics and adjusted tests across a range of nuisance parameter ω from 10^5 sims.

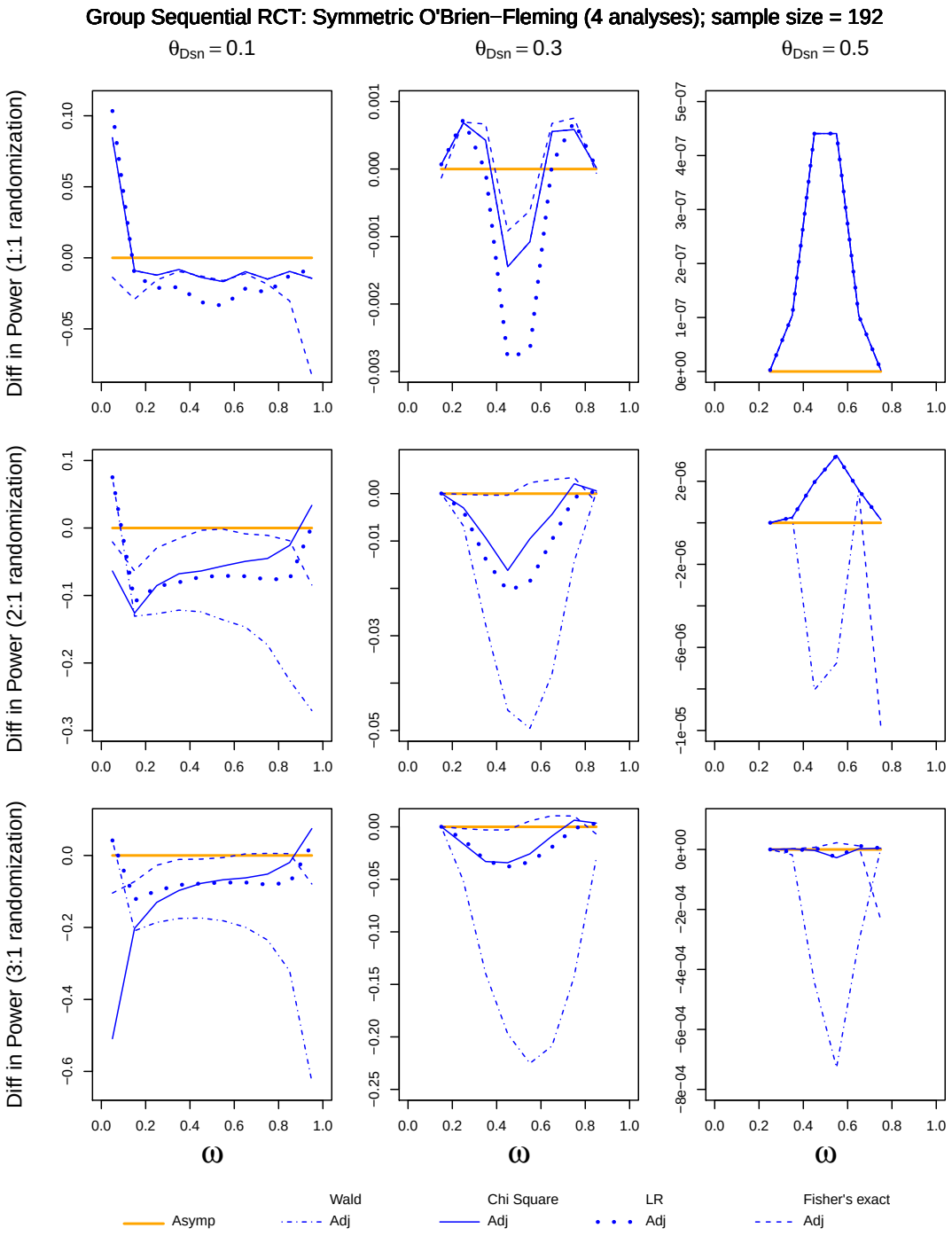


Figure 4-6: Symm OBF GSDs (up to 4 analyses, sample size of 192) for one-sided greater 0.025 nominal level for specified randomization ratio. Diff in power (simulated – asymp) is computed based on adjusted tests across a range of nuisance parameter ω from 10^5 sims.

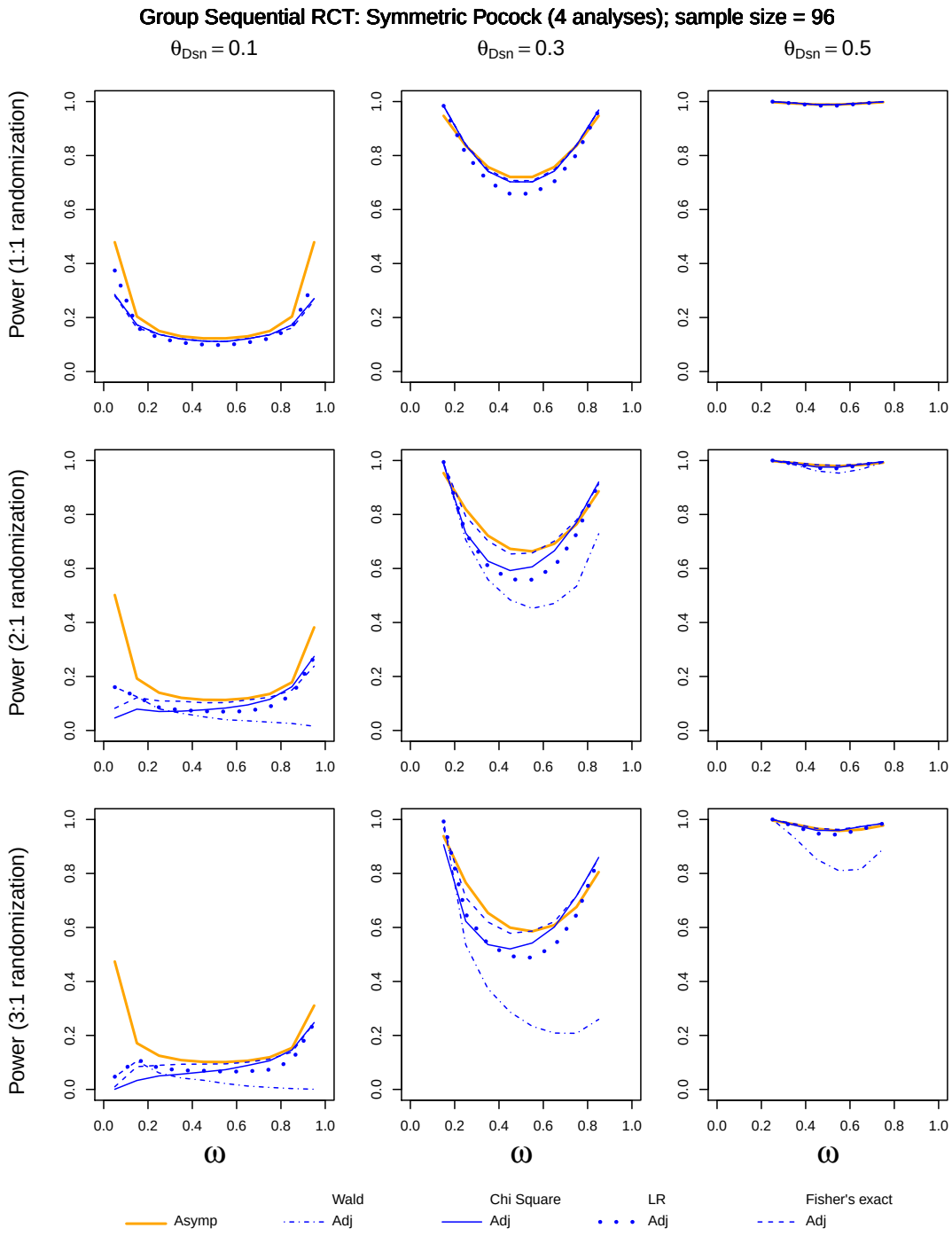


Figure 4-7: Symm Pocock GSDs (up to 4 analyses, sample size of 96) for one-sided greater 0.025 nominal level for specified randomization ratio. Power is computed based on asymptotics and adjusted tests across a range of nuisance parameter ω from 10^5 sims.

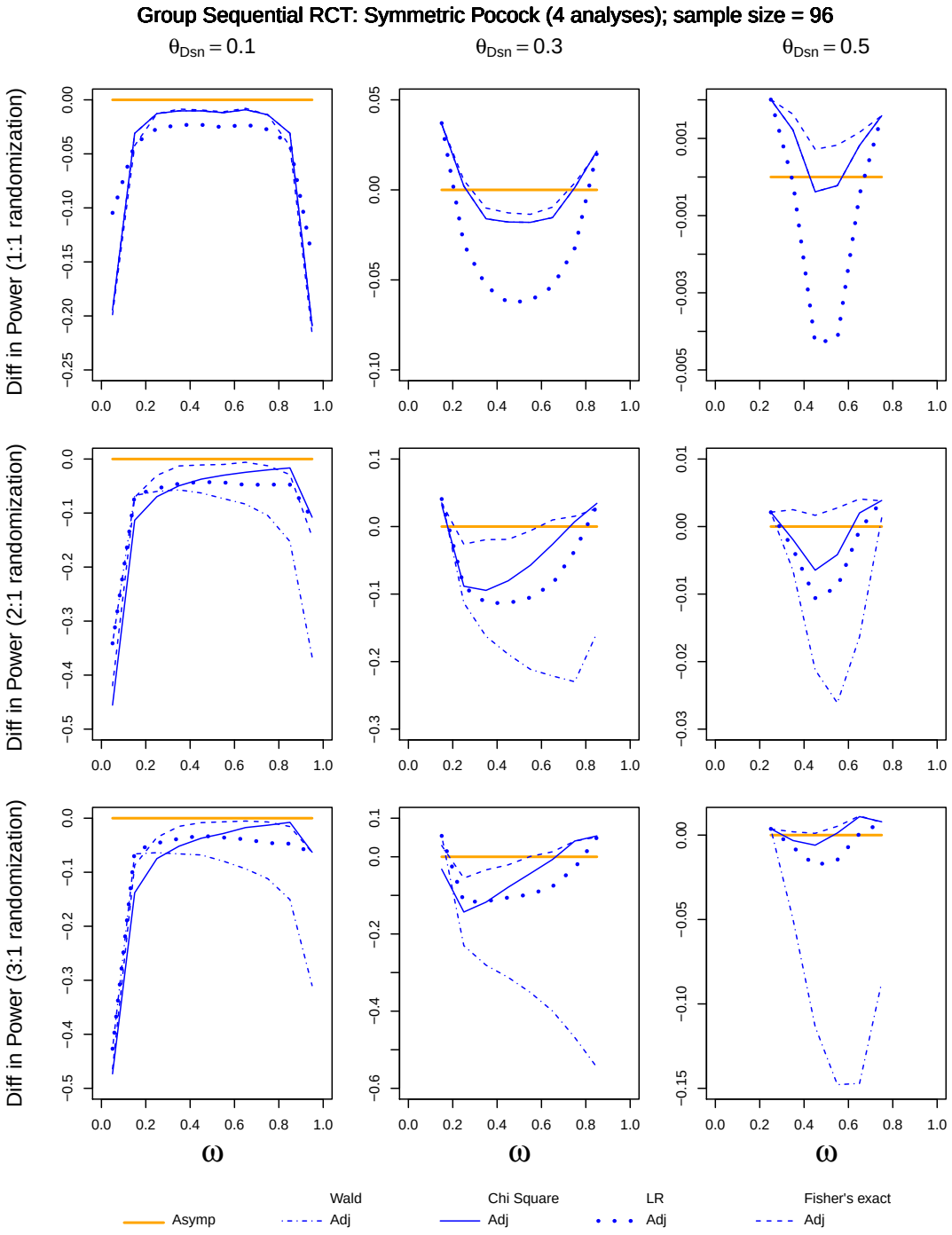


Figure 4-8: Symm Pocock GSDs (up to 4 analyses, sample size of 96) for one-sided greater 0.025 nominal level for specified randomization ratio. Diff in power (simulated – asymp) is computed based on adjusted tests across a range of nuisance parameter ω from 10^5 sims.

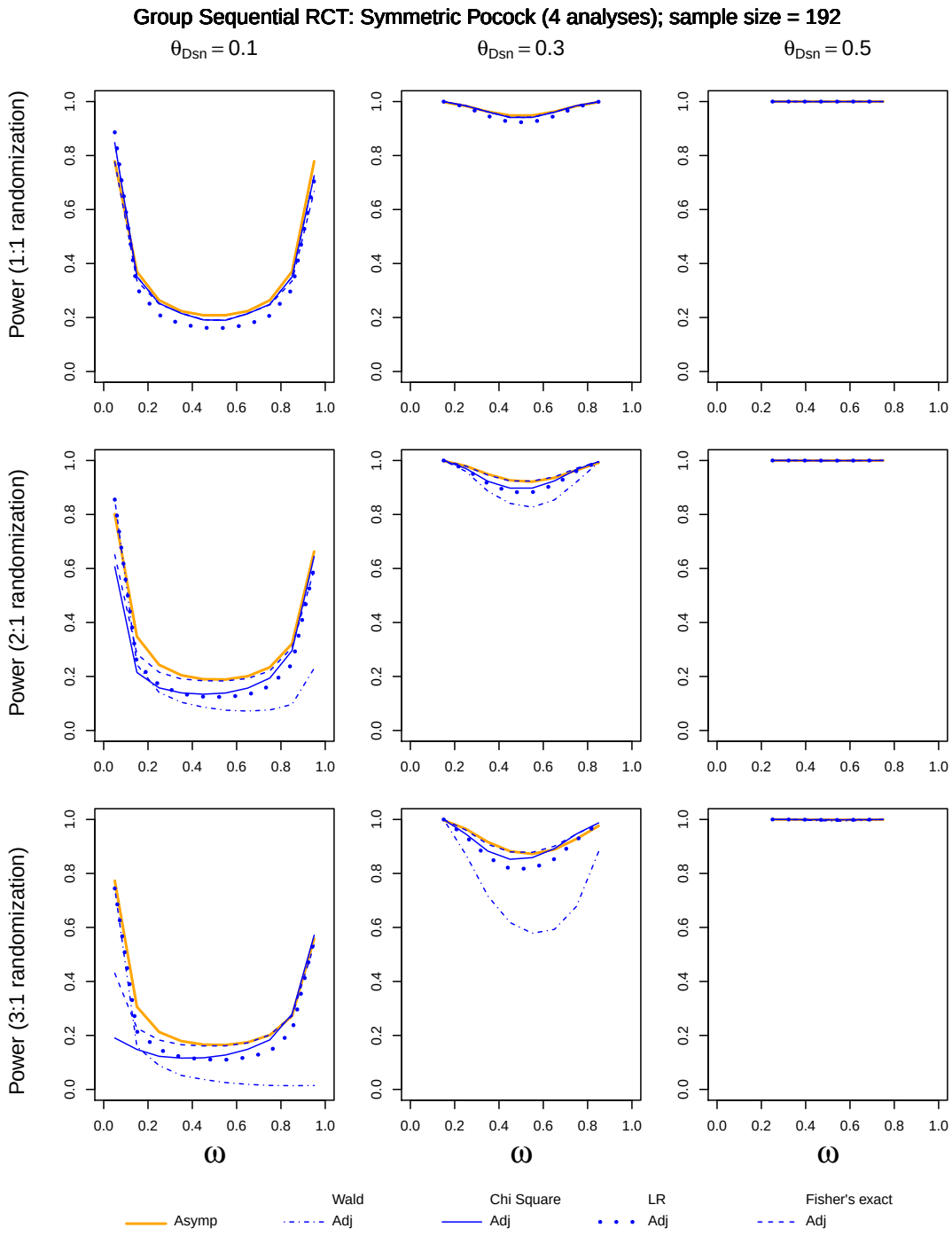


Figure 4-9: Symm Pocock GSDs (up to 4 analyses, sample size of 192) for one-sided greater 0.025 nominal level for specified randomization ratio. Power is computed based on asymptotics and adjusted tests across a range of nuisance parameter ω from 10^5 sims.

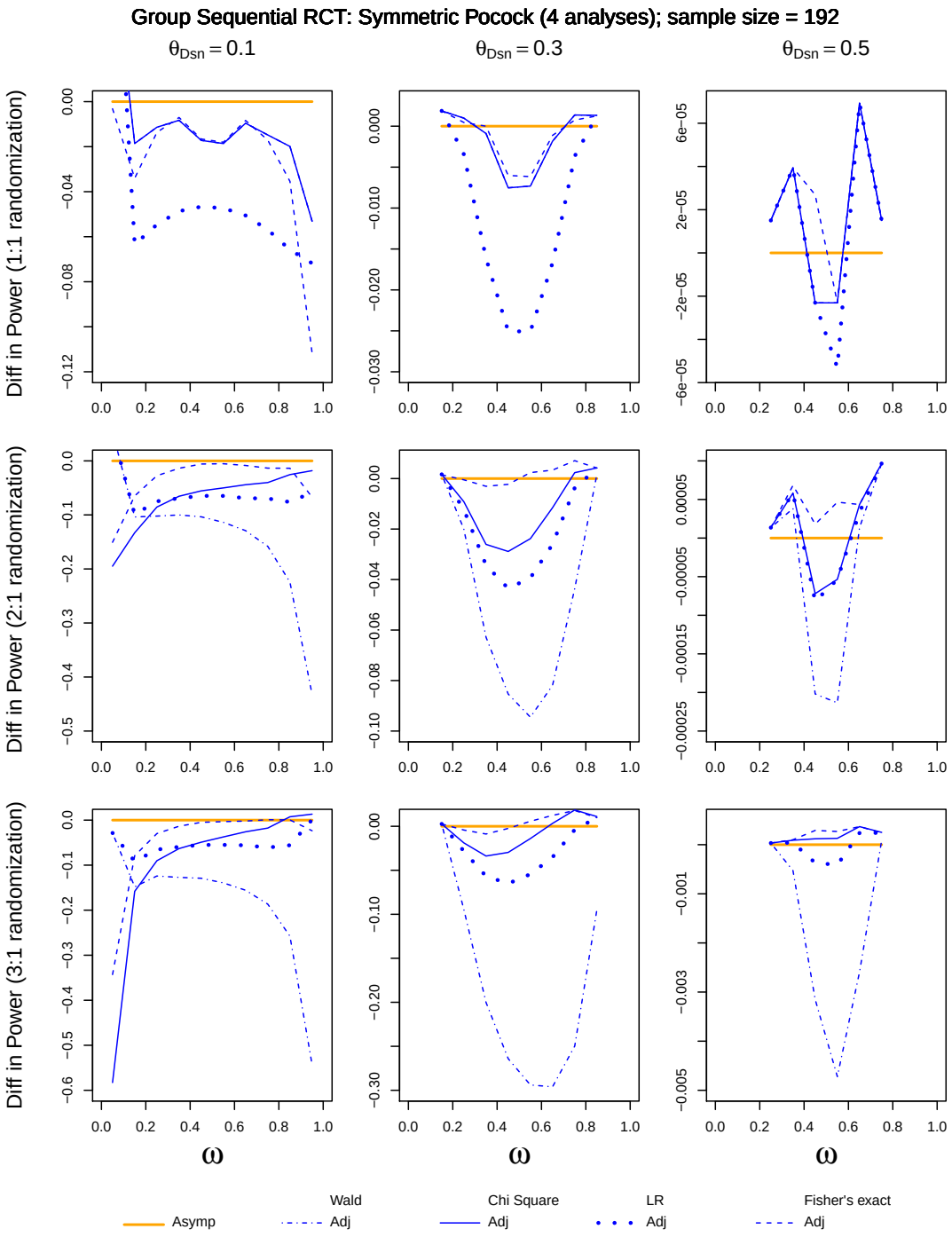


Figure 4-10: Symm Pocock GSDs (up to 4 analyses, sample size of 192) for one-sided greater 0.025 nominal level for specified randomization ratio. Diff in power (simulated – asymp) is computed based on adjusted tests across a range of nuisance parameter ω from 10^5 sims.

We further investigate power of the adjusted chi square and Fisher's exact tests by considering two symmetric group sequential designs, one using O'Brien-Fleming boundary and the other using Pocock boundary, with 1:1, 2:1, and 3:1 randomizations, design alternative $\theta_{Dsn} = 0.3$ and a specific design nuisance parameter $\omega_{Dsn} = 0.5$. Figures 4-11 and 4-12 contains plots of the fixed design power and corresponding asymptotic and simulated overall power according to 1) chi square and 2) Fisher's exact test statistics with adjusted critical values across a range of values of difference in proportions θ. Note that 100,000 simulations provided adequate precision to discriminate between curves. We find from these figures that:

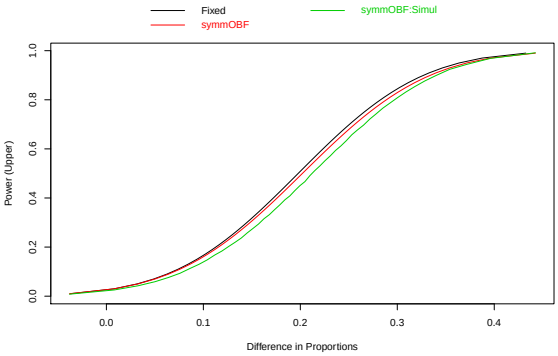
- The power based on simulations is lower than that based on asymptotic results: the degree of difference varies. The fixed sample design has the most power.

- Under 1:1 randomization, the agreement between power curves for asymptotic results and adjusted chi square is comparable, whereas there is more vertical separation between power curves for asymptotic results and adjusted Fisher's exact test.

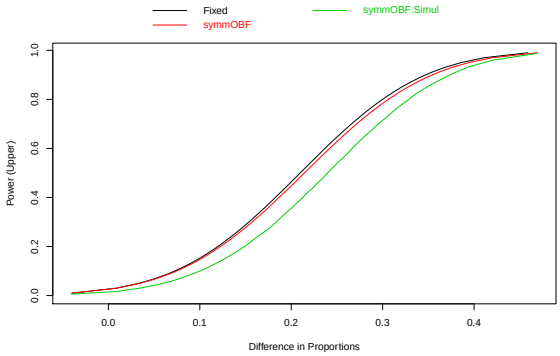
- Under 2:1 randomization, the agreement between power curves for asymptotic results and adjusted Fisher's exact test is comparable, whereas there is more vertical separation between power curves for asymptotic results and adjusted chi square.

- Under 3:1 randomization, the agreement between power curves for asymptotic results and adjusted test is comparable for both adjusted chi square and Fisher's exact tests.

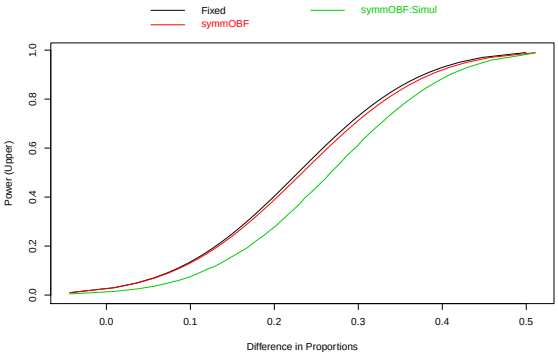
These findings appear consistent with what we have found so far with adjusted chi square and Fisher's exact tests for group sequential designs with different randomization ratios.



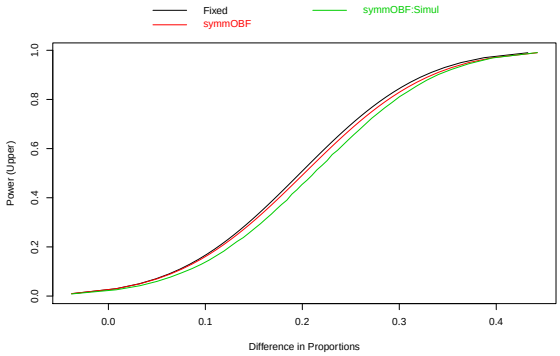
(a) Adjusted Chi Square (1:1 randomization)



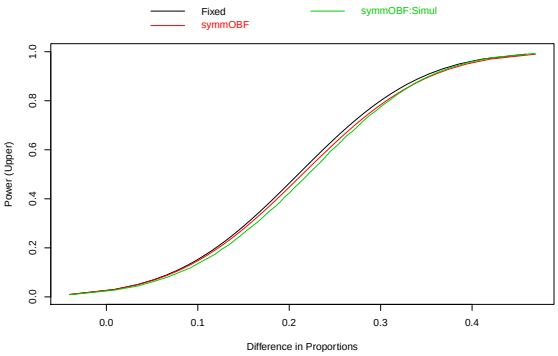
(b) Adjusted Fisher's exact (1:1 randomization)



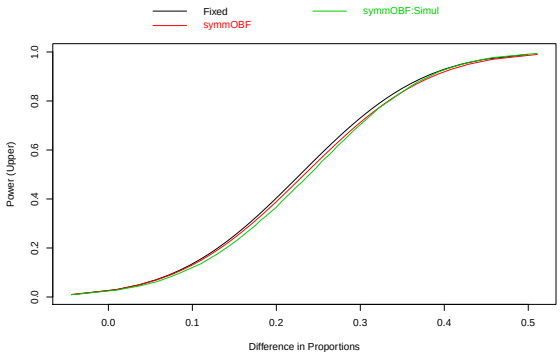
(c) Adjusted Chi Square (2:1 randomization)



(d) Adjusted Fisher's exact (2:1 randomization)

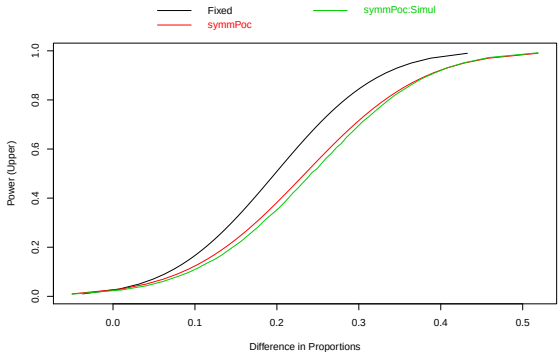


(e) Adjusted Chi Square (3:1 randomization)

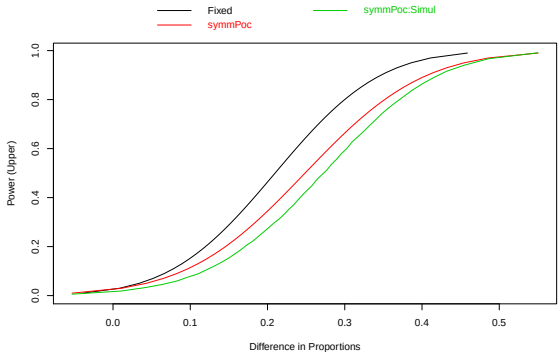


(f) Adjusted Fisher's exact (3:1 randomization)

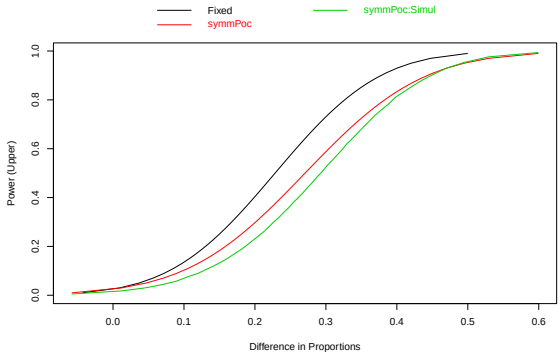
Figure 4-11: Symm OBF GSDs (up to 4 analyses, sample size of 96) with design alternative $\theta_{Dsn} = 0.3$ and nuisance parameter $\omega = 0.5$ for one-sided greater 0.025 nominal level for specified randomization ratio. Power is computed based on asymptotics and adjusted tests from 10^5 sims.



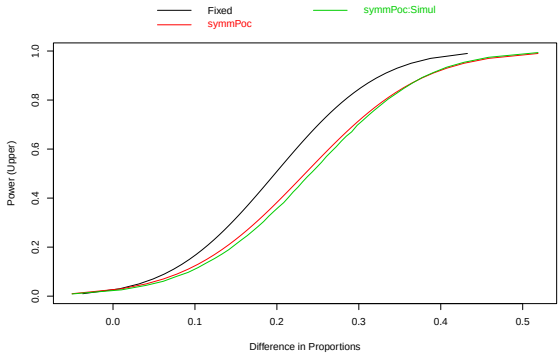
(a) Adjusted Chi Square (1:1 randomization)



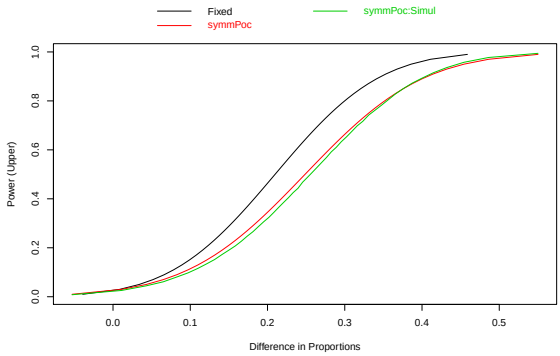
(b) Adjusted Fisher's exact (1:1 randomization)



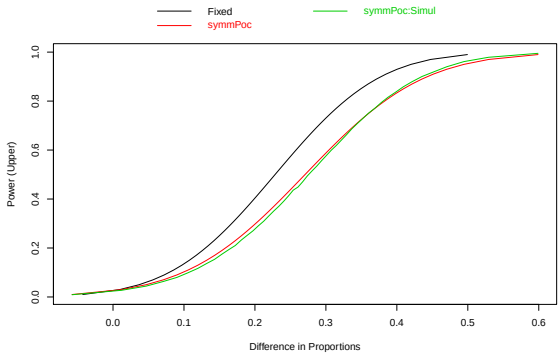
(c) Adjusted Chi Square (2:1 randomization)



(d) Adjusted Fisher's exact (2:1 randomization)



(e) Adjusted Chi Square (3:1 randomization)



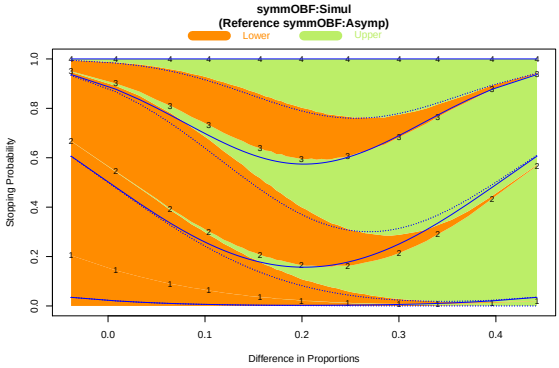
(f) Adjusted Fisher's exact (3:1 randomization)

Figure 4-12: Symm Pocock GSDs (up to 4 analyses, sample size of 96) with design alternative $\theta_{Dsn} = 0.3$ and nuisance parameter $\omega = 0.5$ for one-sided greater 0.025 nominal level for specified randomization ratio. Power is computed based on asymptotics and adjusted tests from 10^5 sims.

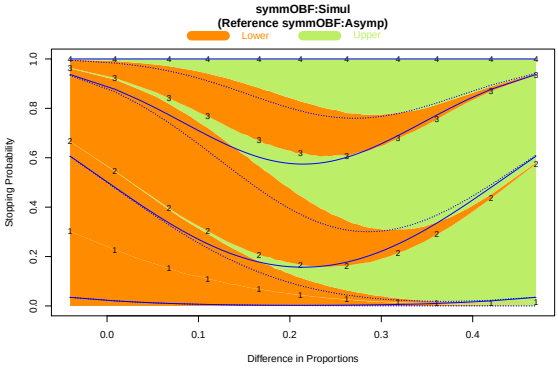
4.2.4 Evaluation: stopping probabilities and error spending function

Figures 4-13 and 4-14 contain plots of stopping probabilities, lower (futility) and upper (efficacy). From overlaying (reference) asymptotic stopping probabilities on top of the stopping probabilities computed based on simulations for a given unconditional exact test statistic, we find disagreement between the two. For O'Brien-Fleming designs, the magnitude of the disagreement is most noticeable at the second and third analyses: 1) under 2:1 randomization for adjusted chi square; and 2) under 1:1 randomization for adjusted Fisher's exact test. For Pocock designs, the magnitude of the disagreement is most noticeable at the second and third analyses under 1:1 randomization for adjusted chi square and Fisher's exact tests. However, simulated stopping probabilities are not markedly different from those based on asymptotic results.

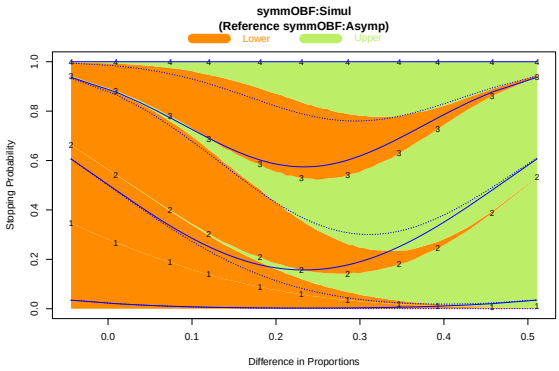
To further illustrate the behavior of the (upper) stopping probabilities, we consider the error spending function in Figures 4-15 to 4-20. The plots comparing error spending functions for adjusted chi square and Fisher's exact tests to each other, as well as the unadjusted tests and asymptotic results. For these we had to specify a particular value of nuisance parameter ω for each plot. These are more detailed versions of what was summarized in the overall type 1 error plots earlier. That is, the final analysis (time 4) upper stopping probability is the overall type 1 error for a given value of nuisance parameter ω and the specified design alternative θ_{Dsn}. But here we are interested in how well or poorly the error spending functions for the adjusted chi square and adjusted Fisher's exact test statistics agree with the asymptotic error spending functions. From examining Figures 4-15 to 4-20, we find that the adjusted chi square appears to have less departure from the asymptotic error spending function for smaller values of ω, which is consistent with what we saw in the Figures 4-13 and 4-14. For more extreme values of ω, adjusted chi square seemed to better handle those situations than adjusted Fisher's, perhaps because of adjusted chi square using restricted MLE whereas adjusted Fisher's gets rid of the nuisance parameter by conditioning on an ancillary statistic. In the case of 2:1 and 3:1 randomization assignments, adjusted Fisher's error spending function is closer to the asymptotic error spending function as seen when examining the relative error spent.



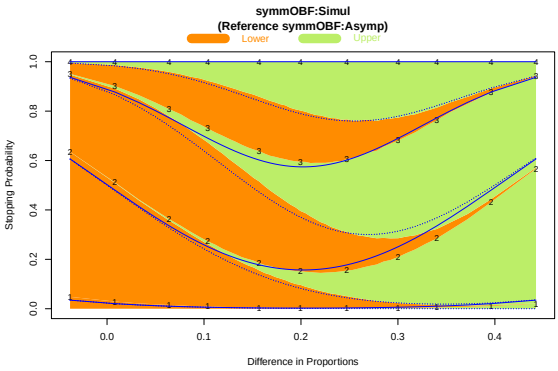
(a) Adjusted Chi Square (1:1 randomization)



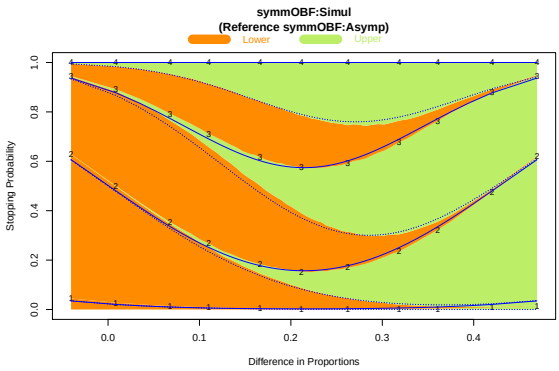
(b) Adjusted Fisher's exact (1:1 randomization)



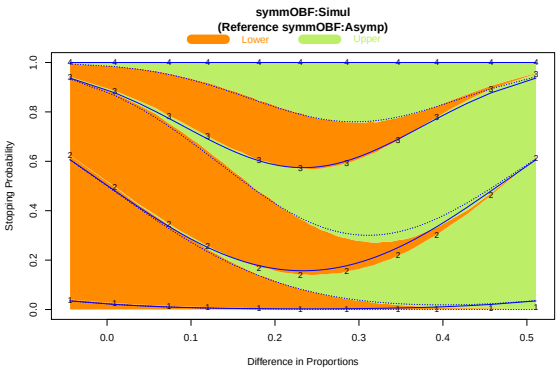
(c) Adjusted Chi Square (2:1 randomization)



(d) Adjusted Fisher's exact (2:1 randomization)

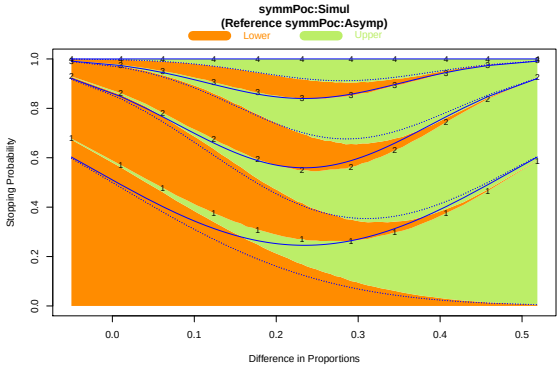


(e) Adjusted Chi Square (3:1 randomization)

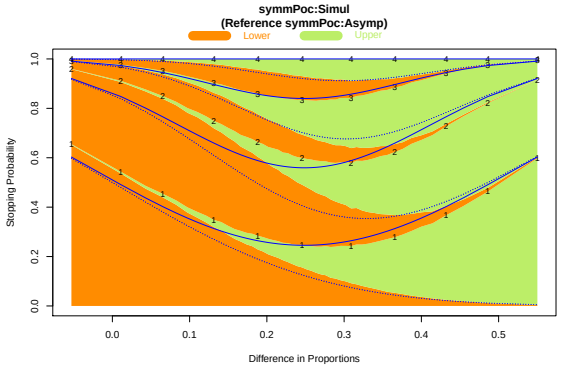


(f) Adjusted Fisher's exact (3:1 randomization)

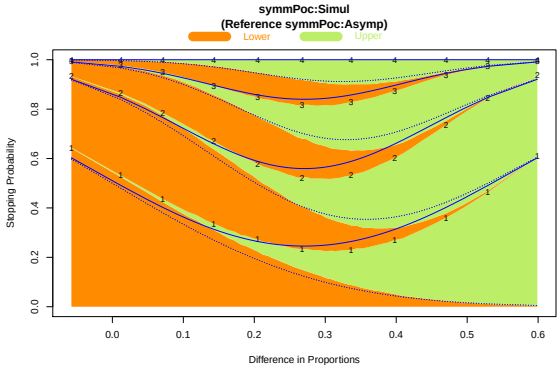
Figure 4-13: Symm OBF GSDs (up to 4 analyses, sample size of 96) with design alternative $\theta_{Dsn} = 0.3$ and nuisance parameter $\omega = 0.5$ for one-sided greater 0.025 nominal level for specified randomization ratio. Lower and upper stopping probabilities computed based on asymptotics (reference) and adjusted tests from 10^5 sims.



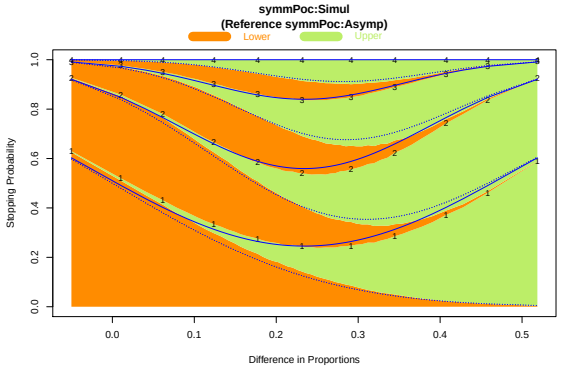
(a) Adjusted Chi Square (1:1 randomization)



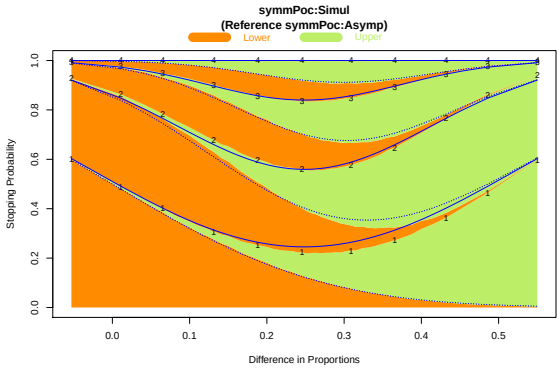
(b) Adjusted Fisher's exact (1:1 randomization)



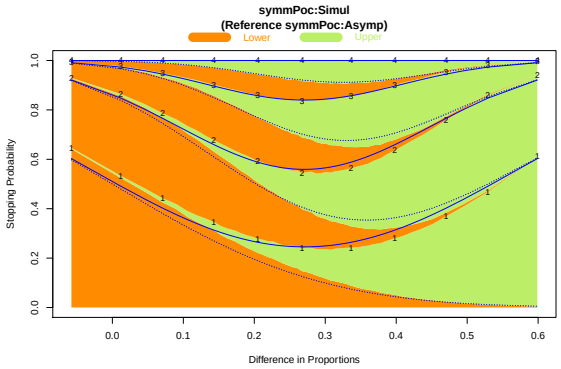
(c) Adjusted Chi Square (2:1 randomization)



(d) Adjusted Fisher's exact (2:1 randomization)



(e) Adjusted Chi Square (3:1 randomization)



(f) Adjusted Fisher's exact (3:1 randomization)

Figure 4-14: Symm Pocock GSDs (up to 4 analyses, sample size of 96) with design alternative $\theta_{Dsn} = 0.3$ and nuisance parameter $\omega = 0.5$ for one-sided greater 0.025 nominal level for specified randomization ratio. Lower and upper stopping probabilities computed based on asymptotics (reference) and adjusted tests from 10^5 sims.

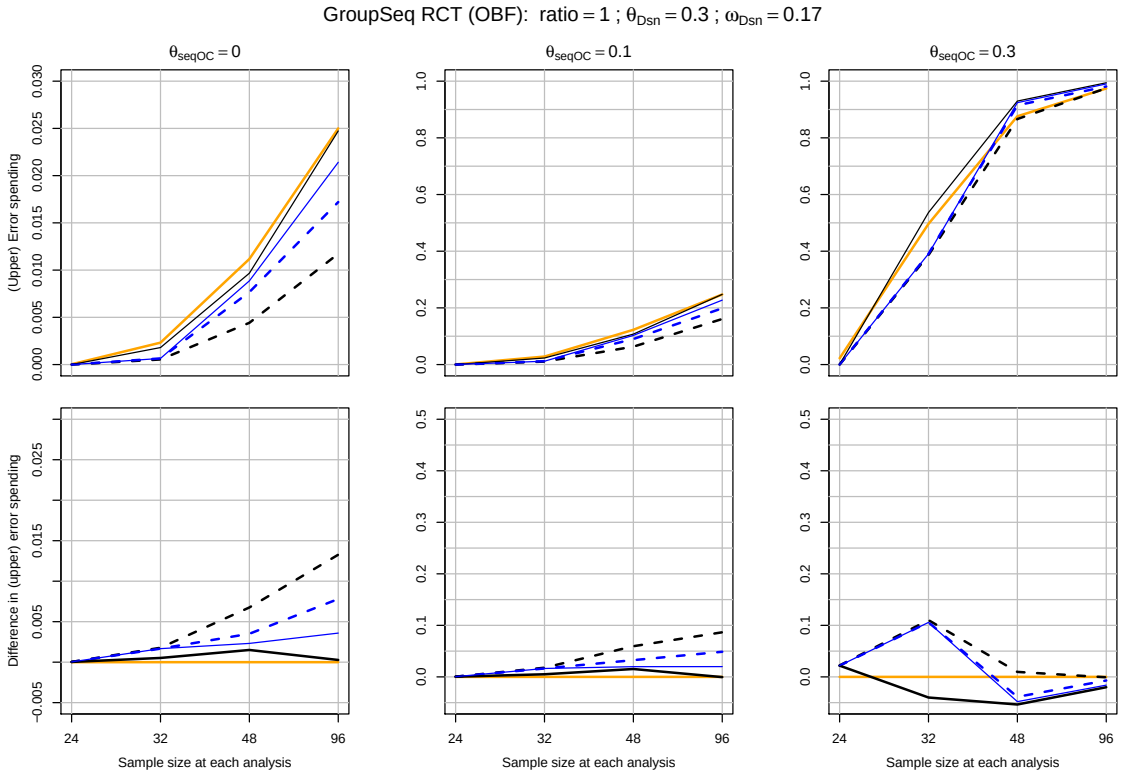


Figure 4-15: Symm OBF GSDs (up to 4 analyses, sample size of 96, 1:1 randomization) with design alternative $\theta_{Dsn} = 0.3$ and nuisance parameter $\omega = 0.17$ for one-sided greater 0.025 nominal level. Error spending (upper/efficacy stopping probability) and difference in error spending (asymp – simul) are computed for differences in proportions ($\theta_{seqOC} \in \{0, 0.1, 0.3\}$) based on asymptotics, unadjusted, and adjusted tests from 10^5 sims.

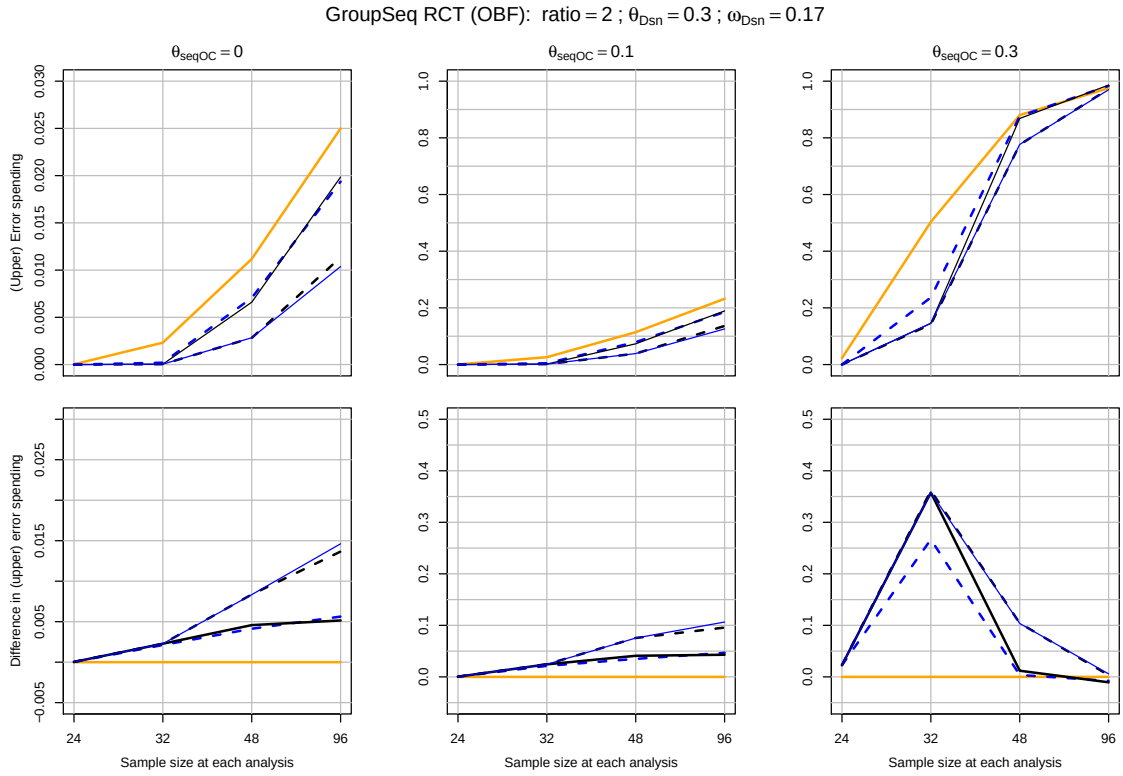


Figure 4-16: Symm OBF GSDs (up to 4 analyses, sample size of 96, 2:1 randomization) with design alternative $\theta_{Dsn} = 0.3$ and nuisance parameter $\omega = 0.17$ for one-sided greater 0.025 nominal level. Error spending (upper/efficacy stopping probability) and difference in error spending (asymp – simul) are computed for differences in proportions ($\theta_{seqOC} \in \{0, 0.1, 0.3\}$) based on asymptotics, unadjusted, and adjusted tests from 10^5 sims.

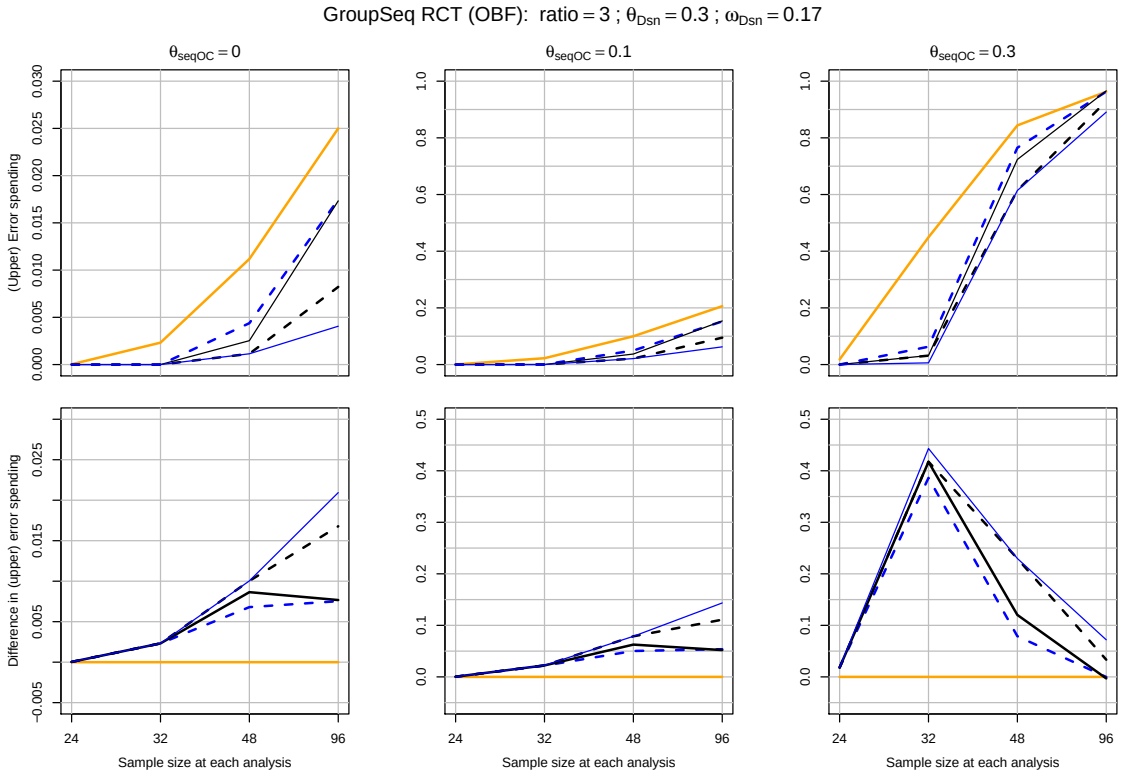


Figure 4-17: Symm OBF GSDs (up to 4 analyses, sample size of 96, 3:1 randomization) with design alternative $\theta_{Dsn} = 0.3$ and nuisance parameter $\omega = 0.17$ for one-sided greater 0.025 nominal level. Error spending (upper/efficacy stopping probability) and difference in error spending (asymp – simul) are computed for differences in proportions ($\theta_{seqOC} \in \{0, 0.1, 0.3\}$) based on asymptotics, unadjusted, and adjusted tests from 10^5 sims.

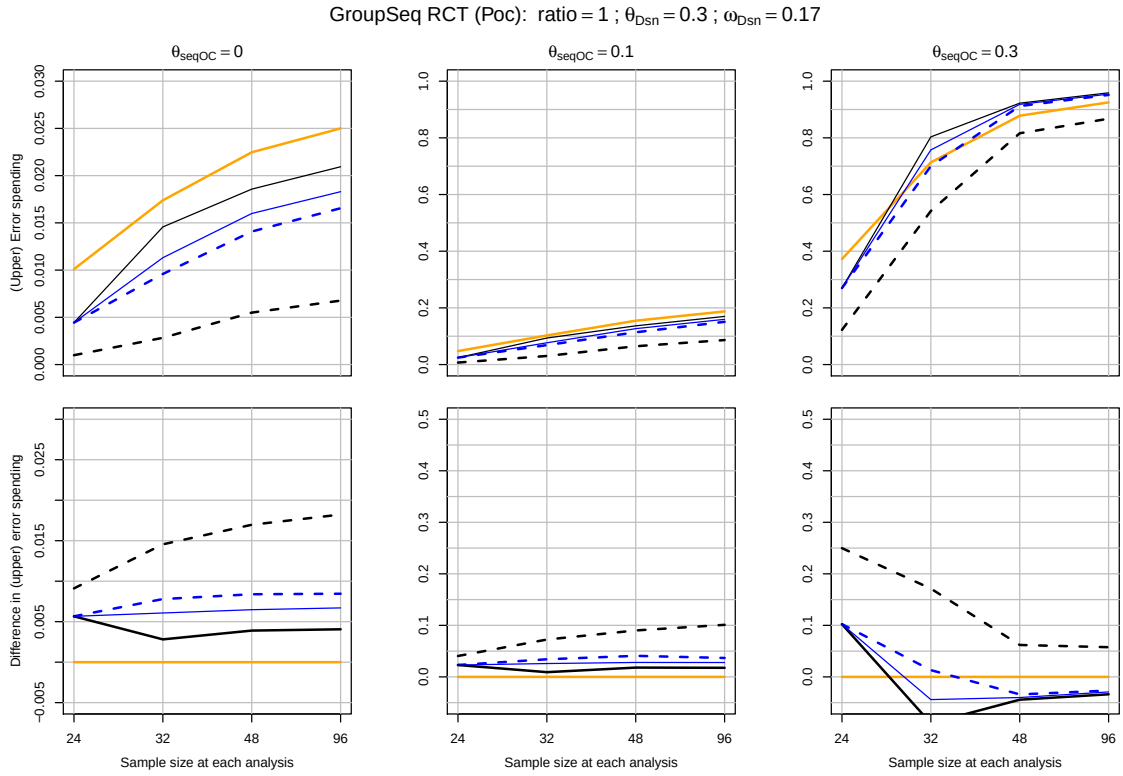


Figure 4-18: Symm Pocock GSDs (up to 4 analyses, sample size of 96, 1:1 randomization) with design alternative $\theta_{Dsn} = 0.3$ and nuisance parameter $\omega = 0.17$ for one-sided greater 0.025 nominal level. Error spending (upper/efficacy stopping probability) and difference in error spending (asymp – simul) are computed for differences in proportions ($\theta_{seqOC} \in \{0, 0.1, 0.3\}$) based on asymptotics, unadjusted, and adjusted tests from 10^5 sims.

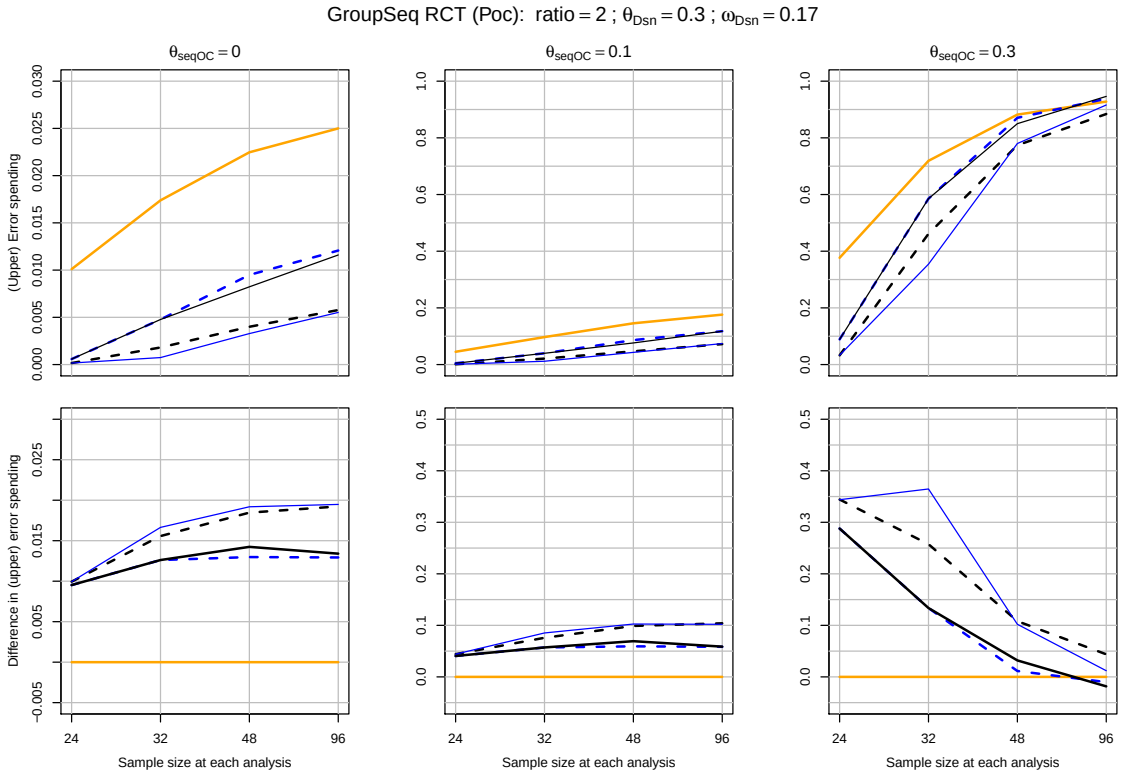


Figure 4-19: Symm Pocock GSDs (up to 4 analyses, sample size of 96, 2:1 randomization) with design alternative $\theta_{Dsn} = 0.3$ and nuisance parameter $\omega = 0.17$ for one-sided greater 0.025 nominal level. Error spending (upper/efficacy stopping probability) and difference in error spending (asymp – simul) are computed for differences in proportions ($\theta_{seqOC} \in \{0, 0.1, 0.3\}$) based on asymptotics, unadjusted, and adjusted tests from 10^5 sims.

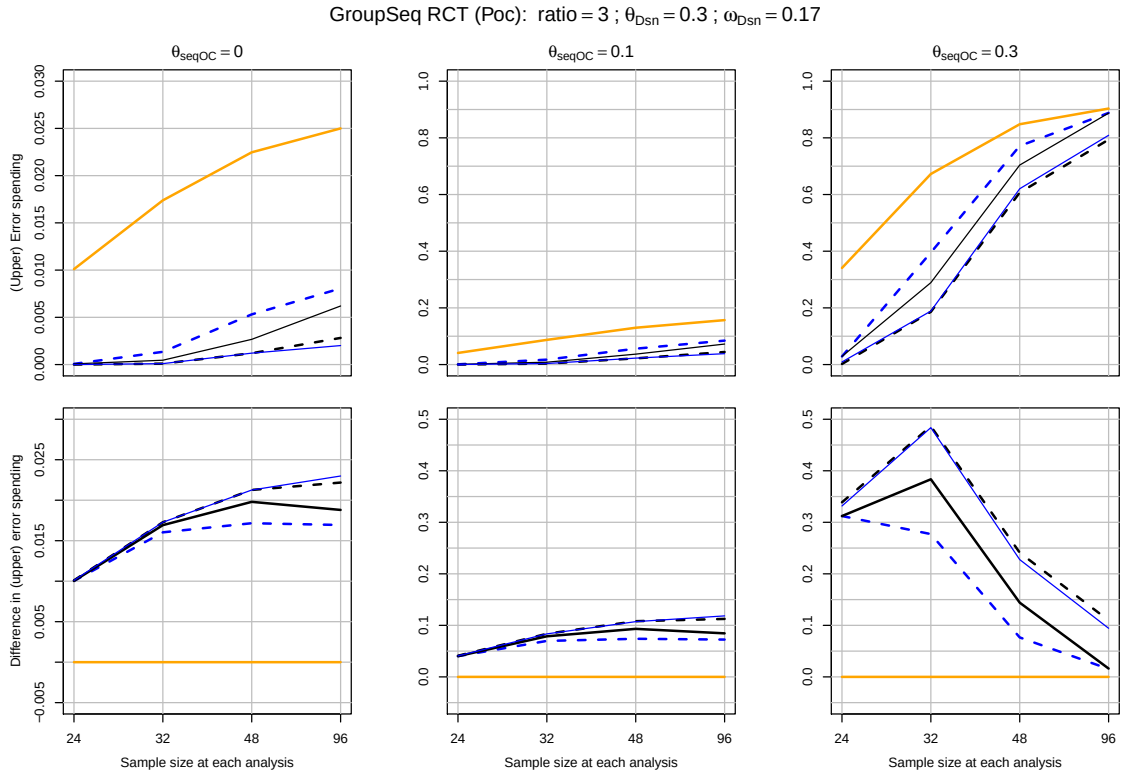
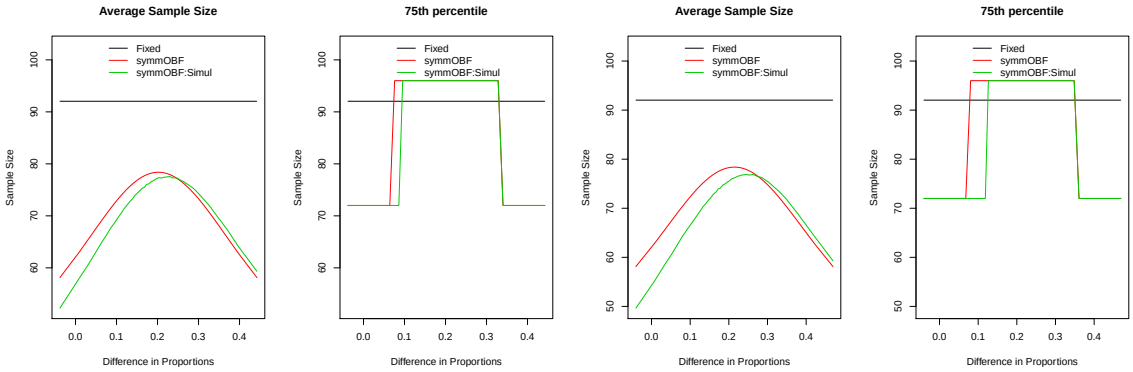


Figure 4-20: Symm Pocock GSDs (up to 4 analyses, sample size of 96, 3:1 randomization) with design alternative $\theta_{Dsn} = 0.3$ and nuisance parameter $\omega = 0.17$ for one-sided greater 0.025 nominal level. Error spending (upper/efficacy stopping probability) and difference in error spending (asymp – simul) are computed for differences in proportions ($\theta_{seqOC} \in \{0, 0.1, 0.3\}$) based on asymptotics, unadjusted, and adjusted tests from 10^5 sims.

4.2.5 Evaluation: sample size distribution

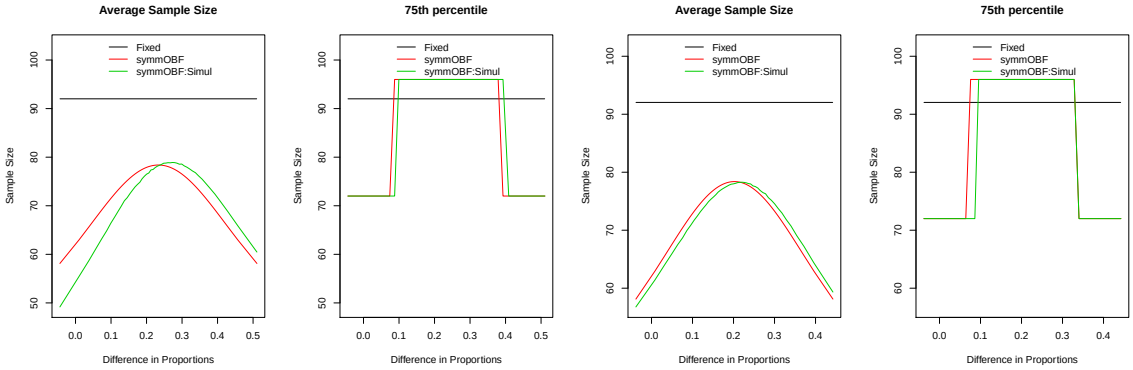
In the group sequential setting, the sample size is no longer fixed as in a fixed sample setting because studies may be terminated prior to the pre-planned final analysis time: a group sequential design having minimum average sample number (ASN) is termed efficient. From Figures 4-21 and 4-22 that contain plots of average sample number (ASN) and 75th percentile of the sample size distribution, we find that using either adjusted chi square or adjusted Fisher's exact test statistic tend to give similar results:

- ASN depends on value of difference in proportions θ. For smaller values of θ simulated ASN is smaller than asymptotic, and for larger values of θ simulated ASN is larger than asymptotic with slightly lower ASN for adjusted Fisher's exact test for these designs.



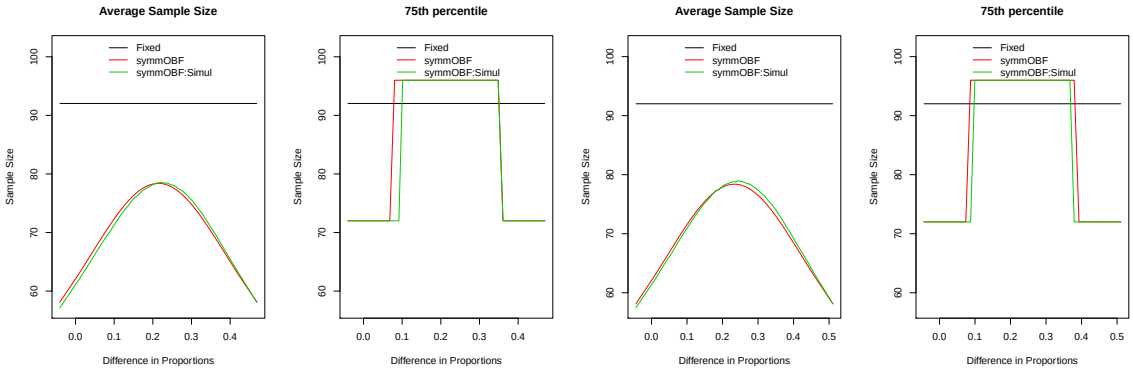
(a) Adjusted Chi Square (1:1 randomization)

(b) Adjusted Fisher's exact (1:1 randomization)



(c) Adjusted Chi Square (2:1 randomization)

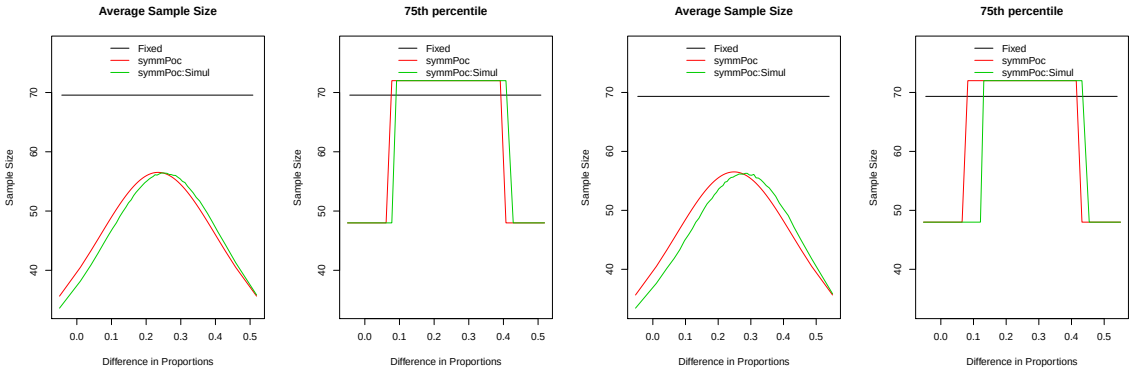
(d) Adjusted Fisher's exact (2:1 randomization)



(e) Adjusted Chi Square (3:1 randomization)

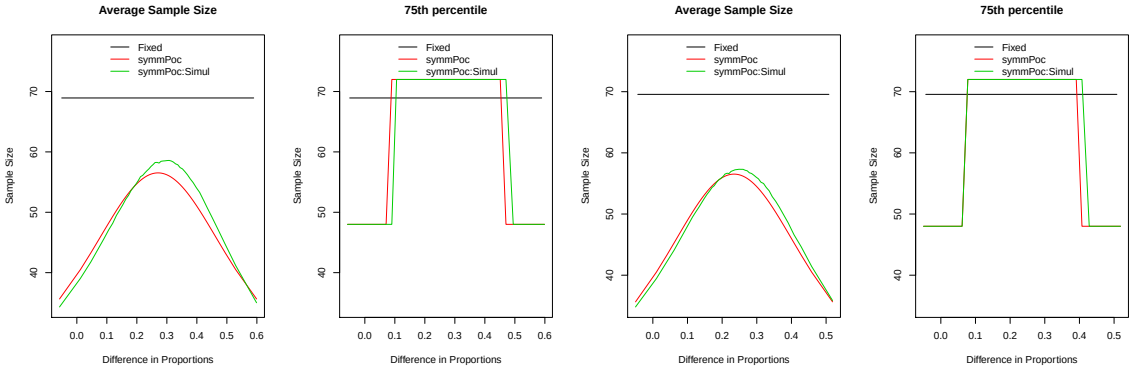
(f) Adjusted Fisher's exact (3:1 randomization)

Figure 4-21: Symm OBF GSDs (up to 4 analyses, sample size of 96) with design alternative $\theta_{Dsn} = 0.3$ and nuisance parameter $\omega = 0.5$ for one-sided greater 0.025 nominal level for specified randomization ratio. Average sample number (ASN) and 75th percentile of the sample size distribution computed based on FD, asymptotics, and adjusted tests from 10^5 sims.



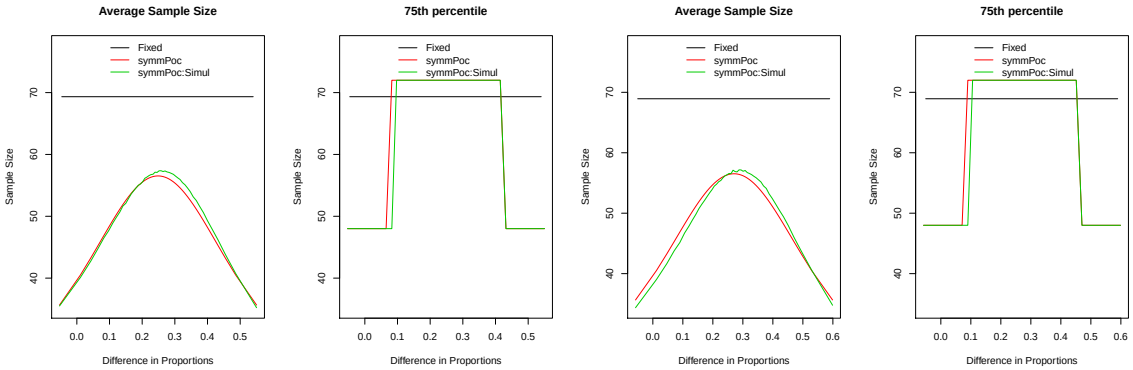
(a) Adjusted Chi Square (1:1 randomization)

(b) Adjusted Fisher's exact (1:1 randomization)



(c) Adjusted Chi Square (2:1 randomization)

(d) Adjusted Fisher's exact (2:1 randomization)



(e) Adjusted Chi Square (3:1 randomization)

(f) Adjusted Fisher's exact (3:1 randomization)

Figure 4-22: Symm Pocock GSDs (up to 4 analyses, sample size of 96) with design alternative $\theta_{Dsn} = 0.3$ and nuisance parameter $\omega = 0.5$ for one-sided greater 0.025 nominal level for specified randomization ratio. Average sample number (ASN) and 75th percentile of the sample size distribution computed based on FD, asymptotics, and adjusted tests from 10^5 sims.

4.2.6 Evaluation: efficiency of symmetric group sequential designs

From the evaluations thus far, we have found:

1. We need to use the adjusted P values at each analysis in order to avoid inflation of the type 1 error.

2. Unconditional exact tests based on Wald or LR tend to be more conservative in terms of type 1 error and have less power than tests based on chi square or Fisher's exact test.

3. With 1:1 randomization, neither chi square nor Fisher's exact test dominate the other with respect to power, though it does appear that basing adjustment on Fisher's exact test is more powerful over a broader range of settings. With 2:1 or 3:1 randomization, adjustment using Fisher's exact test seems to be preferable.

4. The use of adjusted P values results in operating characteristics (stopping probabilities and ASN) that differ somewhat from those estimated using asymptotic results.

Taken together, the above suggest that use of Fisher's exact test based on unconditional exact tests would tend to be best overall, and that explicit evaluation of operating characteristics through simulation will be needed to ensure that we understand the full operating characteristics of the tests.

It is thus of interest to additionally examine the efficiency of symmetric group sequential designs by finding the unified family boundary value shape parameter P that minimizes the ASN. Emerson and Fleming (1989) used asymptotic based approaches to identify "approximately" optimal group sequential designs. Here we can compare the similarity of their conclusions to the result obtained with adjusted tests and see whether the basic relationships observed in the asymptotic setting translate to the use of unconditional exact test P values.

Figures 4-23 to 4-30 examining ASN across a range of symmetric monitoring guidelines from the unified family of group sequential designs (Kittelson and Emerson, 1999). Tables 4-1 to 4-2 illustrating the optimal symmetric P boundary for several group sequential designs,

including 1:1 and 2:1 randomization ratios. Comparisons between ASN based on asymptotic results and ASN based on unconditional exact test using adjusted Fisher's exact test statistic show that ASN are nearly identical for either randomization ratio. Such a finding suggests that implementing the unconditional exact test using adjusted Fisher's exact test statistic performs well when considering ASN as an operating characteristic. Since such an approach is least conservative (based on type 1 error) across a variety of group sequential designs, including randomization ratios, we recommend implementing the unconditional exact test using adjusted Fisher's exact test statistic in settings such as those described in this thesis.

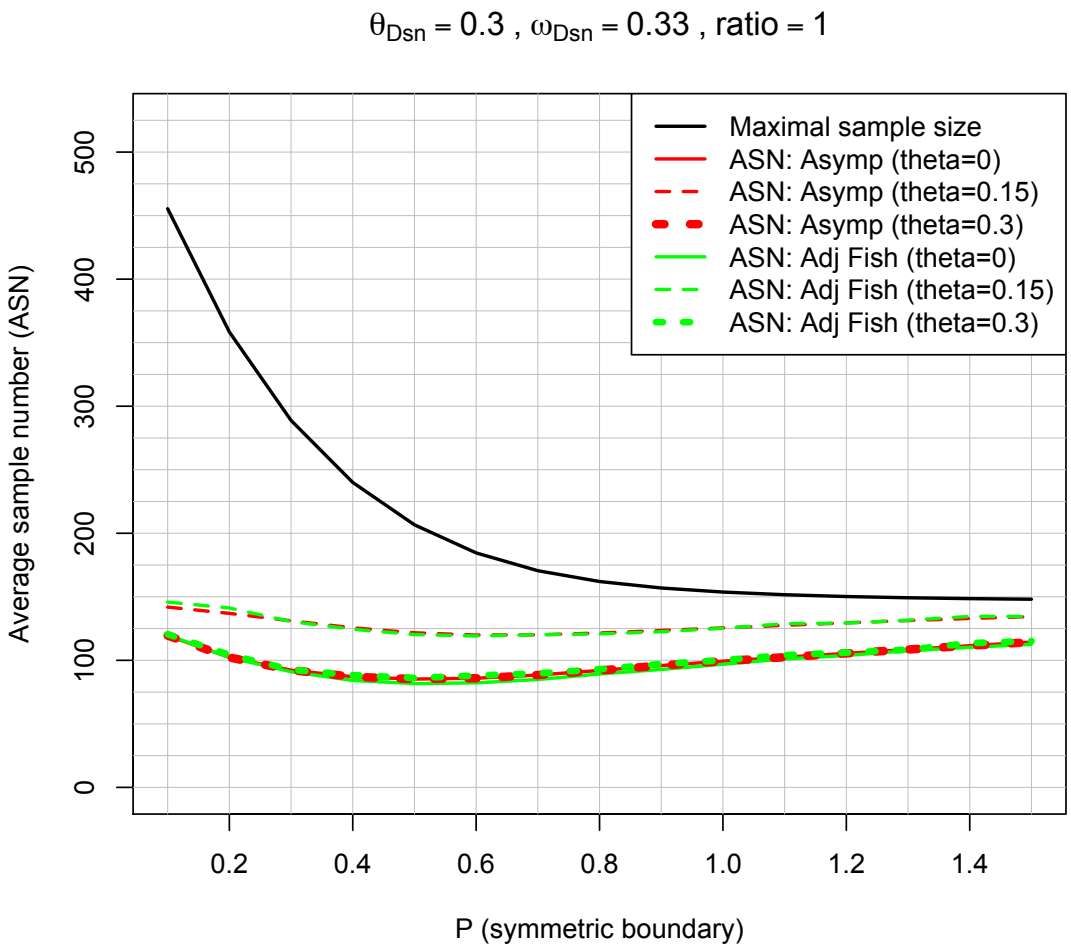


Figure 4-23: Simulated 10^4 RCTs with 1:1 randomization and 97.5% power at design alternative $\theta_{Dsn} = 0.3$ with $\omega = 0.33$ for one-sided greater level 0.025 hypothesis test. Expected sample size (or average sample number, ASN) is computed for range of symmetric P (unified family boundaries/monitoring guidelines). Maximal sample size is in black, asymptotic result is in red, and unconditional exact test using adjusted Fisher's exact test statistic is in green. Additionally, solid, dashed, and dotted lines correspond to $\theta = 0$, 0.15, and 0.3, respectively.

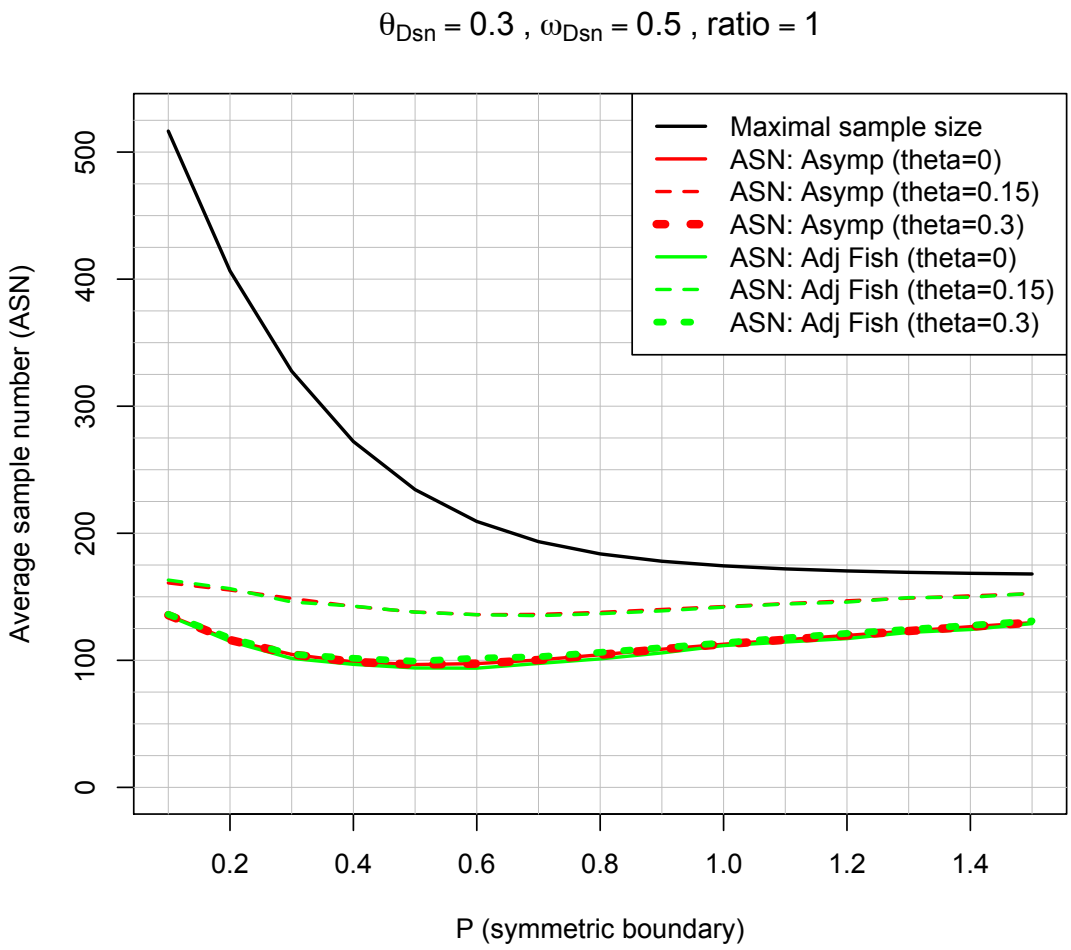


Figure 4-24: Simulated 10^4 RCTs with 1:1 randomization and 97.5% power at design alternative $\theta_{Dsn} = 0.3$ with $\omega = 0.50$ for one-sided greater level 0.025 hypothesis test. Expected sample size (or average sample number, ASN) is computed for range of symmetric P (unified family boundaries/monitoring guidelines). Maximal sample size is in black, asymptotic result is in red, and unconditional exact test using adjusted Fisher's exact test statistic is in green. Additionally, solid, dashed, and dotted lines correspond to $\theta = 0$, 0.15, and 0.3, respectively.

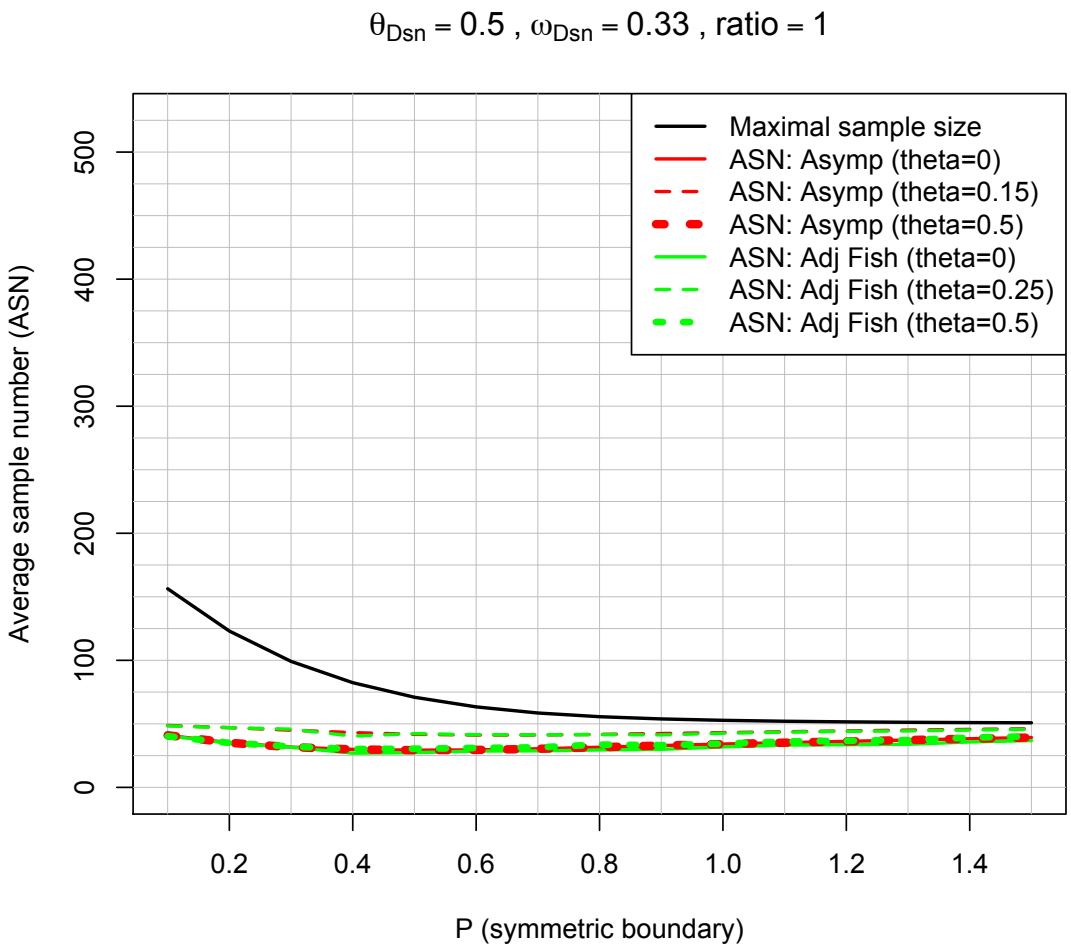


Figure 4-25: Simulated 10^4 RCTs with 1:1 randomization and 97.5% power at design alternative $\theta_{Dsn} = 0.5$ with $\omega = 0.33$ for one-sided greater level 0.025 hypothesis test. Expected sample size (or average sample number, ASN) is computed for range of symmetric P (unified family boundaries/monitoring guidelines). Maximal sample size is in black, asymptotic result is in red, and unconditional exact test using adjusted Fisher's exact test statistic is in green. Additionally, solid, dashed, and dotted lines correspond to $\theta = 0$, 0.25, and 0.5, respectively.

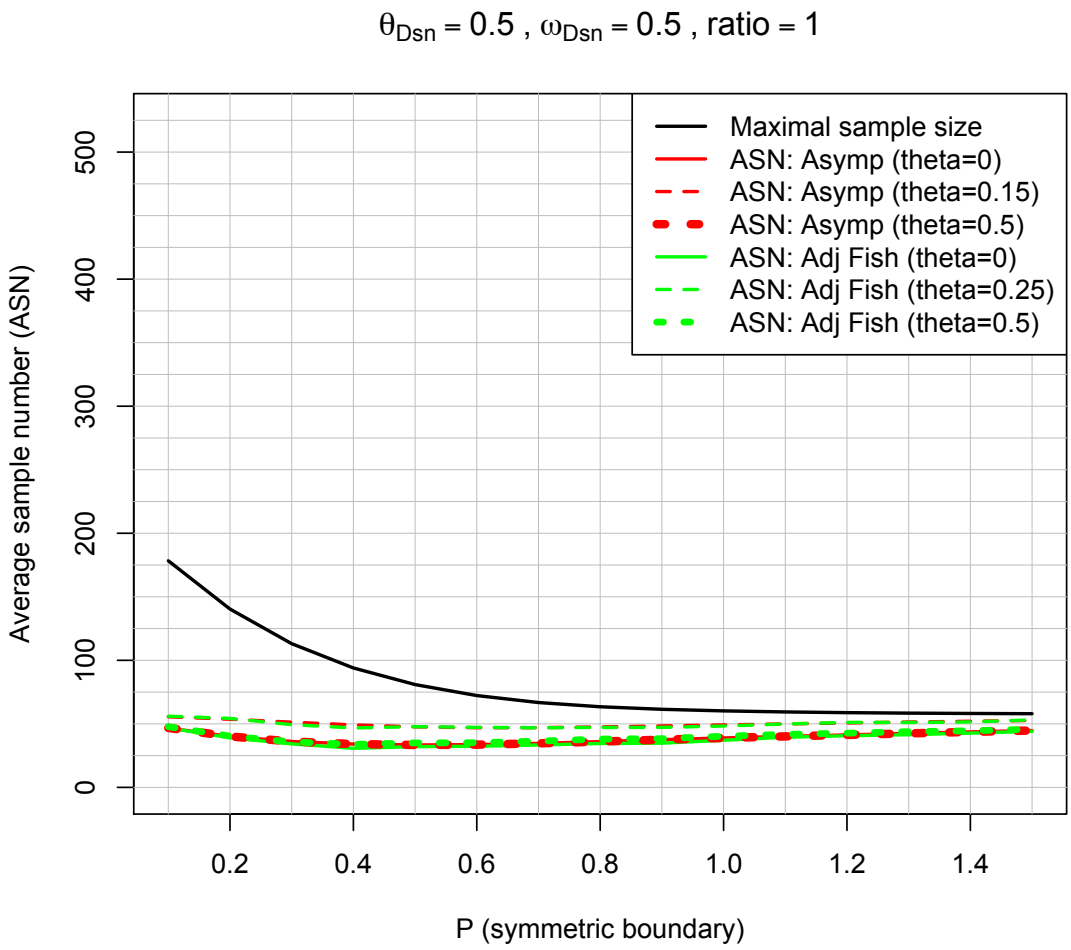


Figure 4-26: Simulated 10^4 RCTs with 1:1 randomization and 97.5% power at design alternative $\theta_{Dsn} = 0.5$ with $\omega = 0.50$ for one-sided greater level 0.025 hypothesis test. Expected sample size (or average sample number, ASN) is computed for range of symmetric P (unified family boundaries/monitoring guidelines). Maximal sample size is in black, asymptotic result is in red, and unconditional exact test using adjusted Fisher's exact test statistic is in green. Additionally, solid, dashed, and dotted lines correspond to $\theta = 0$, 0.25, and 0.5, respectively.

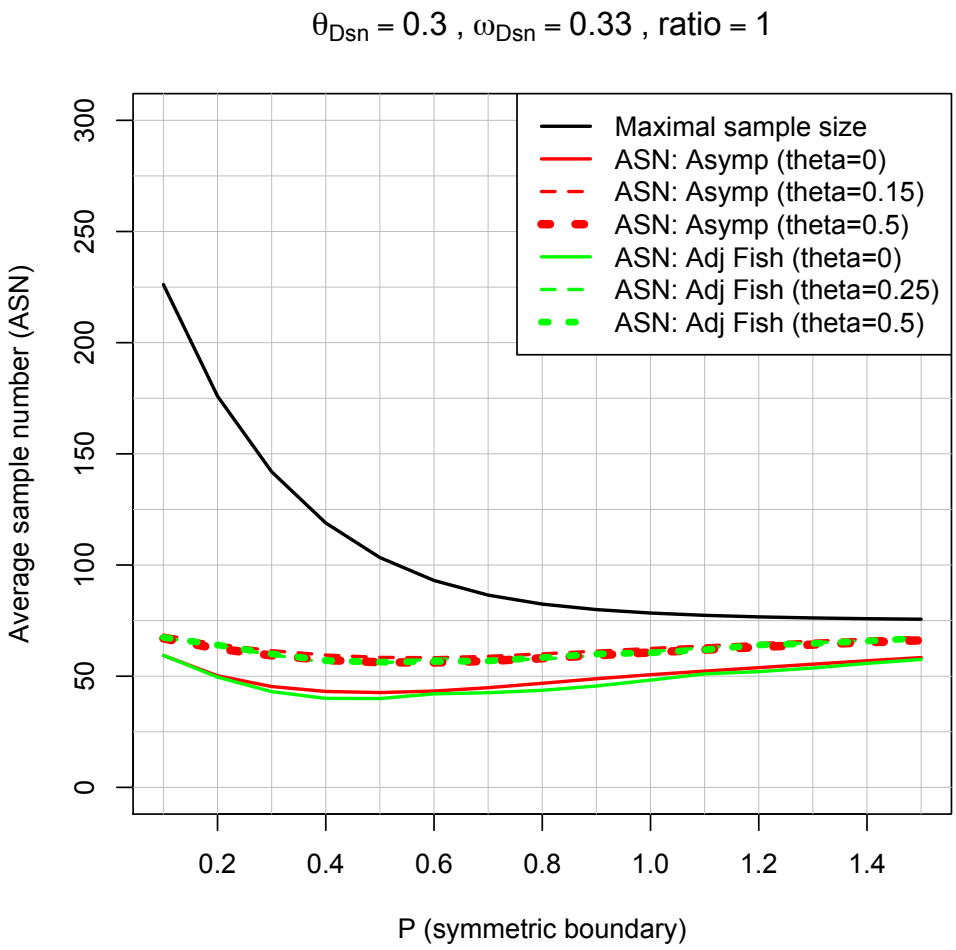


Figure 4-27: Simulated 10^4 RCTs with 1:1 randomization and 80% power at design alternative $\theta_{Dsn} = 0.3$ with $\omega = 0.33$ for one-sided greater level 0.025 hypothesis test. Expected sample size (or average sample number, ASN) is computed for range of symmetric P (unified family boundaries/monitoring guidelines). Maximal sample size is in black, asymptotic result is in red, and unconditional exact test using adjusted Fisher's exact test statistic is in green. Additionally, solid, dashed, and dotted lines correspond to $\theta = 0$, 0.15, and 0.3, respectively.

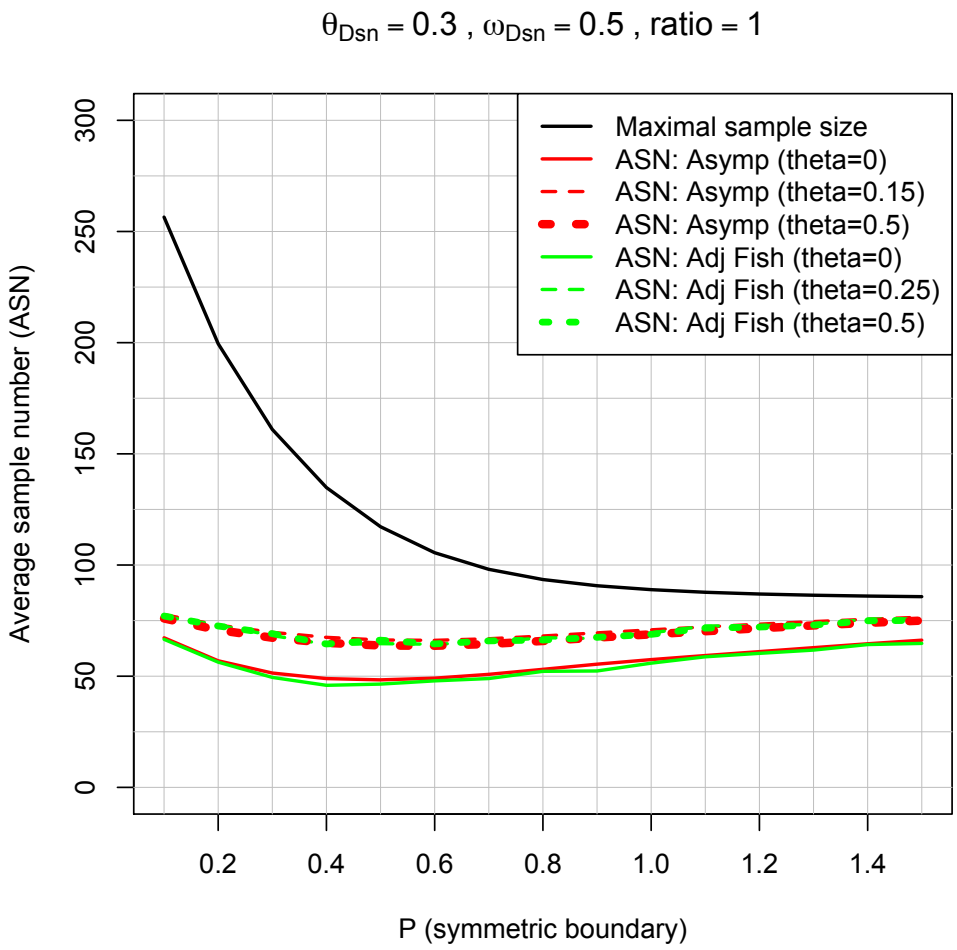


Figure 4-28: Simulated 10^4 RCTs with 1:1 randomization and 80% power at design alternative $\theta_{Dsn} = 0.3$ with $\omega = 0.50$ for one-sided greater level 0.025 hypothesis test. Expected sample size (or average sample number, ASN) is computed for range of symmetric P (unified family boundaries/monitoring guidelines). Maximal sample size is in black, asymptotic result is in red, and unconditional exact test using adjusted Fisher's exact test statistic is in green. Additionally, solid, dashed, and dotted lines correspond to $\theta = 0$, 0.15, and 0.3, respectively.

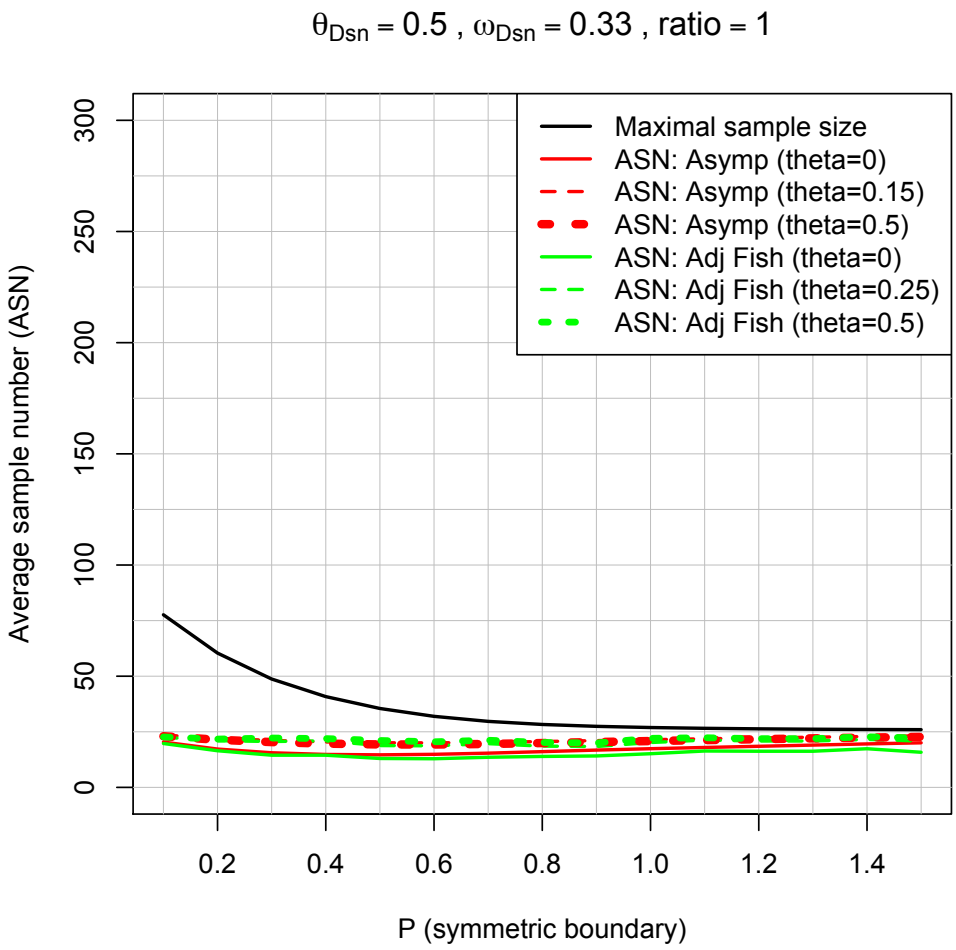


Figure 4-29: Simulated 10^4 RCTs with 1:1 randomization and 80% power at design alternative $\theta_{Dsn} = 0.5$ with $\omega = 0.33$ for one-sided greater level 0.025 hypothesis test. Expected sample size (or average sample number, ASN) is computed for range of symmetric P (unified family boundaries/monitoring guidelines). Maximal sample size is in black, asymptotic result is in red, and unconditional exact test using adjusted Fisher's exact test statistic is in green. Additionally, solid, dashed, and dotted lines correspond to $\theta = 0$, 0.25, and 0.5, respectively.

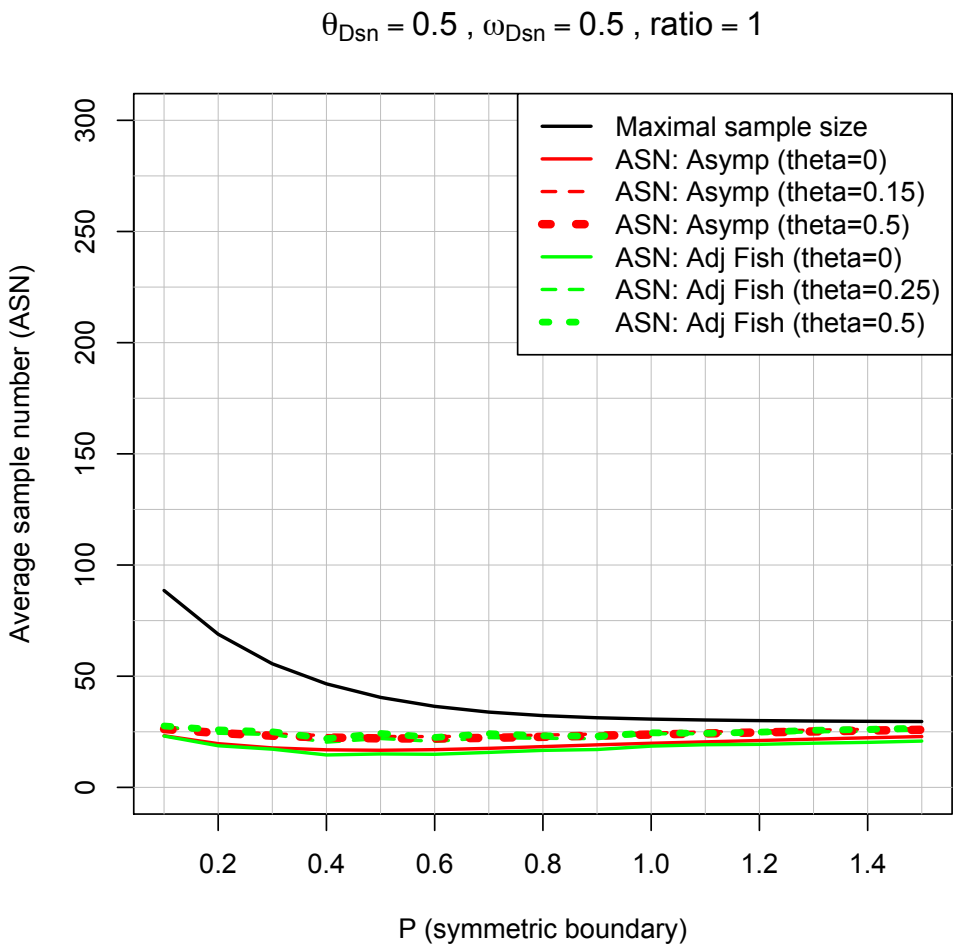


Figure 4-30: Simulated 10^4 RCTs with 1:1 randomization and 80% power at design alternative $\theta_{Dsn} = 0.5$ with $\omega = 0.50$ for one-sided greater level 0.025 hypothesis test. Expected sample size (or average sample number, ASN) is computed for range of symmetric P (unified family boundaries/monitoring guidelines). Maximal sample size is in black, asymptotic result is in red, and unconditional exact test using adjusted Fisher's exact test statistic is in green. Additionally, solid, dashed, and dotted lines correspond to $\theta = 0$, 0.25, and 0.5, respectively.

Table 4-1: Simulated 10^4 RCTs with 1:1 randomization and either 97.5% or 80% power at design alternative $\theta_{Dsn} \in \{0.1, 0.3, 0.5\}$ with $\omega = 0.50$ for one-sided greater level 0.025 hypothesis test. Expected sample size (or average sample number, ASN) is computed for optimal (i.e., P resulting in smallest/minimum ASN based on asymptotics), OBF, and Pocock symmetric P (unified family boundaries/monitoring guidelines). Comparisons are made between ASN computed using asymptotic results and the unconditional exact test with an adjusted Fisher's exact test statistic.

1:1 ratio; 4 analyses; 10,000 sims	null distn ($\theta = 0$, $\omega = 0.50$)				intermediate distn ($\theta = \theta_A / 2$, $\omega = 0.50$)				alternative distn ($\theta = \theta_A$, $\omega = 0.50$)			
	optimal		$P_{OBF} = 1.0$	$P_{Poc} = 0.5$	optimal		$P_{OBF} = 1.0$	$P_{Poc} = 0.5$	optimal		$P_{OBF} = 1.0$	$P_{Poc} = 0.5$
	P_{opt}	ASN_{opt}	ASN_{OBF}	ASN_{Poc}	P_{opt}	ASN_{opt}	ASN_{OBF}	ASN_{Poc}	P_{opt}	ASN_{opt}	ASN_{OBF}	ASN_{Poc}
Pwr = 97.5%												
$\theta_A = 0.1$												
Asymp	0.520	887	1035	888	0.640	1247	1307	1267	0.520	887	1035	888
Adj Fish		883	1027	881		1284	1305	1267		898	1045	899
$\theta_A = 0.3$												
Asymp	0.520	97	113	97	0.640	136	142	138	0.520	97	113	97
Adj Fish		94	112	94		137	142	138		98	114	99
$\theta_A = 0.5$												
Asymp	0.520	33	39	33	0.640	47	49	48	0.520	33	39	33
Adj Fish		32	37	32		46	48	48		34	40	35
Pwr = 80%												
$\theta_A = 0.1$												
Asymp	0.488	444	527.77916	444	0.565	607	651	609	0.560	584	632	586
Adj Fish		443	525	439		604	649	609		589	631	597
$\theta_A = 0.3$												
Asymp	0.488	48	57.46635	48	0.565	66	71	66	0.560	64	69	64
Adj Fish		46	56	46		65	70	65		65	69	66
$\theta_A = 0.5$												
Asymp	0.488	17	19.84132	17	0.565	23	24	23	0.560	22	24	22
Adj Fish		15	19	15		21	24	22		23	25	24

Table 4-2: Simulated 10^4 RCTs with 2:1 randomization and either 97.5% or 80% power at design alternative $\theta_{Dsn} \in \{0.1, 0.3, 0.5\}$ with $\omega = 0.50$ for one-sided greater level 0.025 hypothesis test. Expected sample size (or average sample number, ASN) is computed for optimal (i.e., P resulting in smallest/minimum ASN based on asymptotics), OBF, and Pocock symmetric P (unified family boundaries/monitoring guidelines). Comparisons are made between ASN computed using asymptotic results and the unconditional exact test with an adjusted Fisher's exact test statistic.

2:1 ratio; 4 analyses; 10,000 sims	null distn ($\theta = 0$, $\omega = 0.50$)				intermediate distn ($\theta = \theta_A / 2$, $\omega = 0.50$)				alternative distn ($\theta = \theta_A$, $\omega = 0.50$)			
	optimal		$P_{OBF} = 1.0$	$P_{Poc} = 0.5$	optimal		$P_{OBF} = 1.0$	$P_{Poc} = 0.5$	optimal		$P_{OBF} = 1.0$	$P_{Poc} = 0.5$
	P_{opt}	ASN_{opt}	ASN_{OBF}	ASN_{Poc}	P_{opt}	ASN_{opt}	ASN_{OBF}	ASN_{Poc}	P_{opt}	ASN_{opt}	ASN_{OBF}	ASN_{Poc}
Pwr = 97.5%												
$\theta_A = 0.1$												
Asymp	0.520	998	1164	999	0.640	1403	1471	1426	0.520	998	1164	999
Adj Fish		986	1158	999		1402	1469	1435		1006	1166	1010
$\theta_A = 0.3$												
Asymp	0.520	109	127	109	0.639	153	160	155	0.520	109	127	109
Adj Fish		107	126	106		152	160	155		111	128	110
$\theta_A = 0.5$												
Asymp	0.520	38	44	38	0.640	53	55	54	0.520	38	44	38
Adj Fish		36	43	37		52	55	54		38	45	38
Pwr = 80%												
$\theta_A = 0.1$												
Asymp	0.488	499	594	500	0.565	683	732	685	0.560	657	711	659
Adj Fish		486	589	490		686	729	676		661	713	656
$\theta_A = 0.3$												
Asymp	0.488	54	65	54	0.565	74	80	75	0.560	72	77	72
Adj Fish		54	64	54		73	80	74		72	78	73
$\theta_A = 0.5$												
Asymp	0.488	19	22	19	0.566	26	28	26	0.560	25	27	25
Adj Fish		18	21	18		25	27	26		26	27	26

Chapter 5

DISCUSSION

In this thesis, we described the issues that complicate exact inference for independent binomial outcomes in small to moderate sample sizes: (1) the presence of a mean-variance relationship that depends on nuisance parameters; (2) discreteness of the outcome space; and (3) departures from normality. Although large sample theory based on Wald, score, and likelihood ratio (LR) tests are well developed, suitable small sample methods are necessary when "large" samples are not feasible. Fisher's exact test, which conditions on an ancillary statistic to eliminate nuisance parameters, is common for small samples, however it is "exact" only when a user is willing to base decisions on flipping a biased coin for some outcomes. However, in practice randomized tests are typically not used because they allow for the possibility of different conclusions in instances when the observed outcome is the same. The nonrandomized version of Fisher's exact test tends to be conservative due to the discreteness of the outcome space.

To address the various issues that arise with the asymptotic and/or small sample tests, Barnard (1945, 1947) introduced the concept of unconditional exact tests that use exact distributions of a test statistic evaluated over all possible values of the nuisance parameter. For test statistics derived based on asymptotic approximations, these "unconditional exact tests" ensure that the realized type 1 error is less than or equal to the nominal level. On the other hand, an unconditional test based on the conservative Fisher's exact test statistic can better achieve the nominal type 1 error. In fixed sample settings, it has been found that unconditional exact tests are preferred to conditional exact tests (Mehta and Senchaudhuri, 2003).

In this thesis we first illustrated the behavior of candidate unconditional exact tests in the fixed sample setting, and then extended the comparisons to the group sequential setting. Adjustment of the fixed sample tests was defined as choosing the critical value that

minimizes the conservativeness of the actual type 1 error without exceeding the nominal level of significance. We suggested three methods of using critical values derived from adjusted fixed sample tests to determine the rejection region of the outcome space when testing binomial proportions in a group sequential setting: (1) at the final analysis time only; (2) at analysis times after accrual of more than 50% of the maximal sample size; and (3) at every analysis time.

We considered frequentist operating characteristics when evaluating group sequential designs: overall type 1 error; overall statistical power; stopping probabilities and error spending function; and average sample number (ASN). We found that using the fixed sample critical values is adequate provided they are used at each of the interim analyses. We found relative behavior of Wald, chi square, LR, and Fisher's exact test statistics all depend on the sample size and randomization ratio, as well as the boundary shape function used in the group sequential stopping rule. Owing to its tendency to behave well across a wide variety of settings, we recommend implementation of the unconditional exact test using the adjusted Fisher's exact test statistic at every analysis. Because the absolute behavior of that test varies according to the desired type 1 error, the randomization ratio, and the sample size, we recommend that the operating characteristics for each candidate stopping rule, which depend on the number of interim analyses, be evaluated explicitly for the chosen unconditional exact test.

Extensions from this thesis for evaluating group sequential designs include: (1) performing two-sided hypothesis tests; (2) considering alternative approaches to specifying the rejection region; and (3) computing confidence intervals for full estimation, since we focused on testing.

BIBLIOGRAPHY

[1] Alan Agresti. *Categorical data analysis*. John Wiley & Sons, 2002.

[2] Peter Armitage, CK McPherson, and BC Rowe. Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A (General)*, pages 235–244, 1969.

[3] GA Barnard. A new test for 2×2 contingency tables. *Nature*, 156:177, 1945.

[4] GA Barnard. Significance tests for 2× 2 tables. *Biometrika*, pages 123–138, 1947.

[5] George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

[6] Scott S Emerson and Thomas R Fleming. Symmetric group sequential test designs. *Biometrics*, pages 905–923, 1989.

[7] Scott S Emerson and Thomas R Fleming. Parameter estimation following group sequential hypothesis testing. *Biometrika*, 77(4):875–892, 1990.

[8] Scott S Emerson, John M Kittelson, and Daniel L Gillen. Frequentist evaluation of group sequential clinical trial designs. *Statistics in medicine*, 26(28):5047–5080, 2007.

[9] SS Emerson and PLC Banks. *Case Studies in Biometry (N. Lange, et al., eds.)*. Wiley New York, 1994.

[10] David J Finney. The fisher-yates test of significance in 2× 2 contingency tables. *Biometrika*, pages 145–156, 1948.

[11] Ronald A Fisher. On the interpretation of $\chi 2$ from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, pages 87–94, 1922.

[12] Lawrence M Friedman, Curt Furberg, and David L DeMets. *Fundamentals of clinical trials*, volume 4. Springer, 2010.

[13] Daniel L Gillen and Scott S Emerson. A note on p-values under group sequential testing and nonproportional hazards. *Biometrics*, 61(2):546–551, 2005.

[14] Daniel L Gillen and Scott S Emerson. Designing, monitoring, and analyzing group sequential clinical trials using the rctdesign package for r. In *Proceedings of the Fourth Seattle Symposium in Biostatistics: Clinical Trials*, pages 177–208. Springer, 2013.

[15] Christopher Jennison and Bruce W Turnbull. *Group sequential methods with applications to clinical trials*. CRC Press, 2010.

[16] John M Kittelson and Scott S Emerson. A unifying family of group sequential test designs. *Biometrics*, 55(3):874–882, 1999.

[17] E.L. Lehmann. *Elements of Large-Sample Theory*. Springer Texts in Statistics. Springer, 1999.

[18] Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer, 2006.

[19] Erich Leo Lehmann and George Casella. *Theory of point estimation*, volume 31. Springer, 1998.

[20] Cyrus R Mehta and Pralay Senchaudhuri. Conditional versus unconditional exact tests for comparing two binomials. *Cytel Software corporation*, 675, 2003.

[21] Peter C O'Brien and Thomas R Fleming. A multiple testing procedure for clinical trials. *Biometrics*, pages 549–556, 1979.

[22] Stuart J Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199, 1977.

[23] Stuart J Pocock. *Clinical trials: a practical approach*, volume 23. Wiley Chichester, 1983.

[24] John Whitehead. *The design and analysis of sequential clinical trials*. John Wiley & Sons, 1997.

[25] Frank Yates. Contingency tables involving small numbers and the χ 2 test. *Supplement to the Journal of the Royal Statistical Society*, pages 217–235, 1934.

Appendix A

DERIVATION OF THE FIXED SAMPLE DISTRIBUTION OF $\hat{\theta}$ VIA TRANSFORMATION (JACOBIAN) METHOD

Let $A_i \sim \mathcal{B}(n_i, p_i)$ for independent groups $i = 0, 1$. The maximum likelihood estimator of p_i is $\hat{p}_i = A_i/n_i$ for groups $i = 0, 1$. Suppose the target of inference is $\theta = p_1 - p_0$, with MLE $\hat{\theta} = \hat{p}_1 - \hat{p}_0$. To derive the distribution of $\hat{\theta}$, we consider the transformation (Jacobian) method. Let

$$\begin{cases} V = \frac{A_1}{n_1} - \frac{A_0}{n_0} & \in (-1, 1) \\ W = \frac{A_0}{n_0} & \in (0, 1) \end{cases} \Longrightarrow \begin{cases} A_1 = n_1(V + W) & \in \{0, 1, \ldots, n_1\} \\ A_0 = n_0 W & \in \{0, 1, \ldots, n_0\} \end{cases}$$

The Jacobian is

$$|J|^+ = \begin{vmatrix} n_1 & n_1 \\ 0 & n_0 \end{vmatrix}^+ = n_1 n_0.$$

The joint distribution of (V, W) is

$$\begin{aligned} f_{V,W}(v, w) &= f_{A_0,A_1}(n_0 w, n_1(v + w)) \cdot n_1 n_0 \\ &\overset{A_1, A_0 \, indep}{=} f_{A_1}(n_1(v + w)) \cdot f_{A_0}(n_0 w) \cdot n_1 n_0 \\ &= \binom{n_1}{n_1(v + w)} p_1^{n_1(v+w)} (1 - p_1)^{n_1(1-v-w)} \cdot \binom{n_0}{n_0 w} p_0^{n_0 w} (1 - p_0)^{n_0(1-w)} \cdot n_1 n_0. \end{aligned}$$

Integrating out w from the joint density leads to the marginal distribution of $V \equiv \hat{\theta}$)

$$\begin{aligned} f_{\hat{\theta}} \equiv f_V(v) &= \int_0^1 f_{V,W}(v, w) dw \\ &\propto_v \left(\frac{p_1}{1 - p_1} \right)^{n_1 v} \cdot \int_0^1 \binom{n_1}{n_1(v + w)} \binom{n_0}{n_0 w} \left[\left(\frac{p_1}{1 - p_1} \right)^{n_1} \left(\frac{p_0}{1 - p_0} \right)^{n_0} \right]^w dw \end{aligned}$$

where $\binom{n_1}{n_1(v+w)} \binom{n_0}{n_0 w}$ can be written as

$$\Gamma(n_0 w + 1)\, \Gamma(n_0 - n_0 w + 1)\, \Gamma(n_1 v + n_1 w + 1)\, \Gamma(n_1 - n_1 v - n_1 w + 1).$$

However, obtaining an explicit form for the marginal distribution (density) of $\hat{\theta}$ is difficult due to w's in the binomial coefficients. Therefore, obtaining an explicit form for the cumulative distribution function of $\hat{\theta}$

$$F_{\hat{\theta}} \equiv F_V(v) = \int_{-1}^{v} f_V(v')dv'$$

is also difficult.

Appendix B

COMPARISON OF VARIANCES ACCORDING TO PARAMETRIZATION OF NUISANCE PARAMETER

Let $\hat{\theta}$ denote the estimator of the true value of the difference in binomial proportions, θ. The variance of the estimator $\hat{\theta}$ for two parameterizations of the nuisance parameter (p_0 and ω) are

$$\begin{aligned} Var_{p_0}(\hat{\theta}) &= \frac{p_0(1-p_0)}{n_0} + \frac{p_1(1-p_1)}{n_1} \\ &= \frac{p_0(1-p_0)}{n_0} + \frac{(p_0+\theta)(1-p_0-\theta)}{n_1} \\ &\overset{\text{(if } n_1=n_0=n\text{)}}{=} \frac{2\left(p_0(1-p_0) + (1-2p_0)\frac{\theta}{2} - \frac{\theta^2}{2}\right)}{n} \end{aligned}$$

and, when $p_0 = \omega - \theta/2$ and $p_1 = \omega + \theta/2$,

$$\begin{aligned} Var_{\omega}(\hat{\theta}) &= \frac{p_0(1-p_0)}{n_0} + \frac{p_1(1-p_1)}{n_1} \\ &= \frac{\left(\omega - \frac{\theta}{2}\right)\left(1-\omega+\frac{\theta}{2}\right)}{n_0} + \frac{\left(\omega + \frac{\theta}{2}\right)\left(1-\omega-\frac{\theta}{2}\right)}{n_1} \\ &= \frac{\left(\omega - \omega^2 + \frac{\theta\omega}{2} - \frac{\theta}{2} + \frac{\theta\omega}{2} - \frac{\theta^2}{4}\right)}{n_0} + \frac{\left(\omega - \omega^2 - \frac{\theta\omega}{2} + \frac{\theta}{2} - \frac{\theta\omega}{2} - \frac{\theta^2}{4}\right)}{n_1} \\ &\overset{\text{(if } n_1=n_0=n\text{)}}{=} \frac{2\left(\omega - \omega^2 - \frac{\theta^2}{2}\right)}{n}. \end{aligned}$$