A Comparison of Approaches for Unplanned Sample Size Changes in Phase II Clinical

Trials


By

Molly Olson


Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biostatistics

June 30, 2017

Nashville, Tennessee



Approved:                                                                Date:

_____          _____

Tatsuki Koyama, Ph.D.

_____          _____

Jeffrey Blume, Ph.D.

# ACKNOWLEDGMENTS

I would like to express my deepest appreciation to my advisor, Professor Tatsuki Koyama, for his constant enthusiasm, encouragement, and willingness to teach over the course of my thesis work and my time at Vanderbilt University. Without his continual guidance and assistance, this thesis would not have been possible. I would also like to thank the Director of Graduate Studies and committee member, Professor Jeffrey Blume, for providing me ample opportunities for advancing my education and for being supportive throughout my graduate career.

I would like to thank my family for being unconditionally supportive throughout my endeavors and encouraging me to pursue my ambition. I'm especially thankful to my parents who supported me in so many ways through the years and always believed in me. I would also like to thank my sister for always listening and being there for me when I needed it.

Lastly, my deepest gratitude is expressed to my fellow graduate students. The last two years wouldn't have been possible without their support, help, and encouragement. My time at Vanderbilt has been a wonderful journey, and I'm honored to have shared my time with them.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# ABSTRACT

Oncology phase II clinical trials are used to evaluate the initial effect of a new regimen to determine if there warrants further study in a phase III clinical trial. Two-stage designs with an early futility stop are commonly used in these phase II trials. It is common for attained sample sizes in these trials to be different from the designed sample sizes due to over- and under- enrollment. Currently, when the attained sample size differs from that planned, common practice is to treat the attained sample sizes as planned, and this practice leads to invalid inference. In this thesis, we examine the problems and solutions in hypothesis testing for two-stage phase II clinical trial designs when attained sample sizes differ from the planned sample sizes. We describe existing methods for redesigning trials when there is over- or under-enrollment in either the first or second stage and introduce a new method for redesigning a two-stage clinical trial when the first stage sample size deviates from planned. We focus our investigation when there is over- or under-enrollment in the first stage. We compare the frequentist methods of Chang *et al.*, Olson and Koyama, and the Likelihood two-stage design by applying these methods to two-stage designs with deviations in the first stage of $\pm 10$. We examine type I error rates, power, probability of early termination and expected sample size under the null hypothesis in a number of two-stage designs. We also compare error rates in these methods using a Monte Carlo simulation.

# Chapter 1

# Introduction

Oncology phase II clinical trials are often used to evaluate the initial effect of a new regimen to determine if further study is warranted in a phase III clinical trial [1, 2, 3]. Simon's two-stage design [2] is a commonly used design in phase II oncology clinical trials. Koyama and Chen [3] point out that it is common for actual sample sizes of these phase II trials to differ from the planned, pre-specified sample sizes. Under-enrollment happens due to unanticipated accruement speed or drop-out rates, while multi-center trials can be delayed in communication of enrollment and response information causing over-enrollment. Currently, when the attained sample size differs from planned, common practice is to treat the attained sample sizes as planned. Though, this practice leads to invalid inference, and controlling error rates in hypothesis testing in these cases is not straightforward [1, 3]. Therefore, extensions of two-stage designs for hypothesis testing with unplanned sample size changes is essential.

Many traditional frequentist methods have been proposed to handle unplanned sample sizes in the second stage while using the planned stage I sample size; however, our literature review found that only a few traditional frequentist methods handle unplanned sample sizes in stage I. Moreover, when focusing on deviations in sample sizes in the second stage, many proposed methods are adjusting inference procedures rather than proposing a redesign, where a redesign is a new two-stage design based on the original design and attained sample sizes that is used to carry out hypothesis testing. Likelihood based designs, that can be used to extend Simon's design, offer a nice solution to handling sample size deviations because these designs offer flexibility in sample size without inflation of the type I error rate. Because calculations of p-values are complicated when attained sample sizes are different from planned [3], we focus on methods that offer redesigns of a planned

two-stage design that will be prespecified along with the planned design.

In this paper, we discuss the different approaches for two-stage designs when the attained stage II sample size is different from planned and when attained sample sizes in the first stage is different from planned. We review common two-stage designs in chapter 2 and illustrate redesign methods in chapters 3 and 4. In chapter 5, we review a concrete example from a Likelihood-based clinical trial, and in chapter 6, we present results of a numerical and theoretical study comparing traditional frequentist properties of approaches in the setting where stage I sample size differs from planned.

# Chapter 2

# Background

Two-stage designs for clinical trials are common designs for phase II oncology clinical trials [2]. In two-stage designs, the null hypothesis, $H_0 : p \leq p_0$, is tested against the alternative, $H_1 : p > p_0$, where $p$ is the true response probability and $p_0$ is the highest probability of response that would indicate that the research regimen is uninteresting. The power is set at $p_1$ ($p_1 > p_0$), the lowest probability of response that would indicate that the research regimen warrants further investigation. Under these hypotheses, it is required that the type I error rate be less than $\alpha$ and power be greater than $1 - \beta$. The general framework of a two-stage design includes a sample size and critical value in each of the two-stages. Let $n_1$ denote the first stage sample size, $n_t$ the sample size at the end of the second stage, and let $n_2$ be the sample size for the second stage; $n_2 = n_t - n_1$. Let $X_1$ be the number of successes observed in the first stage and $X_2$ be the number of additional successes in the second stage so that $X_1 \sim \text{Binomial}(n_1, p)$ and $X_2 \sim \text{Binomial}(n_2, p)$. Also, let $X_t = X_1 + X_2$. In the first stage, $n_1$ subjects are enrolled. Let $r_1$ be the first stage critical value. If $r_1$ or fewer subjects ($X_1 \leq r_1$) are successes, then the trial is stopped for futility. If $r_1 + 1$ or more subjects are successful, then the trial continues to the second stage by enrolling $n_2$ additional subjects. Let $r_t$ the critical value for the end of the second stage. If $r_t$ or fewer out of the $n_t$ subjects are successful ($X_t = X_1 + X_2 \leq r_t$), the treatment is considered to be futile, otherwise if $X_t \geq r_t + 1$ subjects succeed, the treatment is considered to be effective and will warrant further study.

Let $b$ denote the binomial probability mass function,

$$b(x, n, p) = \binom{n}{x} p^x (1-p)^{n-x}, x = 1, 2, .., n \tag{2.1}$$

and $B$ denote the cumulative binomial distribution function,

$$B(x,n,p) = \sum_{i=0}^{x} \binom{n}{i} p^i (1-p)^{n-i}, x = 1, 2, ..., n \tag{2.2}$$

The probability of early termination (PET) with a given probability $p$ in two-stage designs is

$$\text{PET} = B(r_1, n_1, p) = P[X_1 \leq r_1 | n_1, p] \tag{2.3}$$

The expected sample size for a given $p$ is then

$$\text{EN} = n_1 + (1 - \text{PET})n_2 \tag{2.4}$$

and the conditional power is then

$$CP(x_1, n_2, p) = \sum_{x_2 = r_t - x_1 + 1}^{n_2} b(x_2, n_2, p) \tag{2.5}$$

It then follows that unconditional power, $UCP(p)$, given probability $p$, is given by

$$UCP(p) = 1 - \left( B(r_1, n_1, p) + \sum_{x=r_1+1}^{min[n_1, r_t]} b(x, p, n_1) B(r_t - x, n_2, p) \right)$$
$$= \sum_{x_1 = r_1 + 1}^{n_1} \left\{ \sum_{x_2 = r_t - x_1 + 1}^{n_2} b(x_2, n_2, p) \right\} b(x_1, n_1, p) \tag{2.6}$$

We require $UCP(p_1) \geq 1 - \beta$ and $UCP(p_0) \leq \alpha$

Simon introduced Optimal and Minimax criteria for selecting good designs [2]. An Optimal two-stage design is a two-stage design which minimizes the expected sample size under the null hypothesis ($\text{EN}_0$) while still satisfying the type I and type II error probability restrictions. The Minimax design minimizes the maximum sample size ($n_t = n_1 + n_2$) while satisfying error constraints. Jung *et al.* [4] introduced an extension of Simon's designs called Admissible designs that are considered a compromise between Optimal and

Minimax. Admissible designs minimize $q \times n_t + (1-q) \times \mathrm{EN}$ for a given $q \in [0,1]$ [4]. Admissible designs satisfy $(\alpha, \beta)$ constraints and obtain an expected sample size and maximum sample size somewhere between those of Optimal and Minimax designs. Admissible designs may be attractive because they have agreeable properties of both the Minimax and Optimal design. Simon's designs do not allow for early termination of the trial for efficacy [2], and we do not consider such designs here. We focus this paper on two-stage designs that are either Optimal, Minimax, or Admissible.

# Chapter 3

# Deviation from Planned Sample Sizes In Second Stage

When over-enrollment occurs in the first stage, a straightforward solution is to perform an interim analysis on the planned number of first stage subjects, and adjust the testing procedure for a sample size in the second stage that may be different than planned. Likewise, it is also straightforward to simply wait for the appropriate enrollment for the first stage when under-enrollment occurs in the first stage. This is why much of the literature that handles sample size deviations only consider deviations in the second stage. Literature exists describing point estimation of the response rate and p-values for hypothesis testing when stage two sample size is modified [3][5][6][7][8][9][10]. Porcher *et al.* [1] conducted a comprehensive review of these methods. Among them is Koyama and Chen who have shown that the p-value in two-stage trials will depend on the planned design in addition to the attained data and is complicated in the setting of unplanned sample size [3]. We only focus on methods that recalculate critical values for hypothesis testing or decision-making, called redesigns, and will not focus on p-value calculations.

In the cases where we evaluate stage I as planned and over- or under-enrollment occurs in the second stage, it is possible to adjust the testing procedure for the attained enrollment in the second stage. This is possible under the assumption of non-informative dropouts; stage I is concluded when the number of non-missing subjects is equal to the planned stage I sample size, and if over-enrollment occurs in the first stage, those subjects will only be considered for the second stage analysis [3]. Koyama and Chen propose a method for inference when stage II sample sizes deviate from the planned stage II sample size [3]. Let

$n_1, n_t, r_1, r_t, \alpha$ and $\beta$ be the original design parameters as defined earlier. The authors let the first stage remain as planned and propose a redesign of the second stage. Using conditional power evaluated at $p_0$, they calculate a new critical value, $r_t^*$, by finding the maximum integer such that

$$CP(x_1, n_2^*, p_0) \leq CP(x_1, n_2, p_0) \equiv P[X_2^* \geq r_t^* | X_1 = x_1, p_0] \leq P[X_2 \geq r_t | X_1 = x_1, p_0] \quad (3.1)$$

where $X_2^* \sim \text{Binomial}(n_2^*, p_0)$ and $n_2^*$ is the attained stage II sample size. This method will result in a controlled unconditional type I error rate because the new critical value gives a conditional type I error rate that is more conservative than the original conditional type I error rate, regardless of the observed stage II sample size. The authors comment that with the new critical value, $r_t^*$, the total number of positive responses required to reject the null hypothesis at the end of the study may be different depending on $x_1$ because it is conditional on the result of the first stage.

Zeng *et al.* [11] propose methodology that attempts to maximize the unconditional power while controlling for the type I error to calculate the stage II critical value for the attained second stage sample size. The authors define $r_2^*$ to be the new second stage critical value when $x_1 \geq r_1$ and $r_t^* \equiv r_2^* + x_1$. The second stage critical value will be the integer that maximizes

$$\text{Power} = \sum_{x_1=r_1}^{n_1} \binom{n_1}{x_1} p_1^{x_1} (1-p_1)^{n_1-x_1} \sum_{x_2=r_2^*}^{n_2^*} \binom{n_2^*}{x_2} p_1^{x_2} (1-p_1)^{n_2^*-x_2} \quad (3.2)$$

while subject to

$$\text{Type I error} = \sum_{x_1=r_1}^{n_1} \binom{n_1}{x_1} p_0^{x_1} (1-p_0)^{n_1-x_1} \sum_{x_2=r_2^*}^{n_2^*} \binom{n_2^*}{x_2} p_0^{x_2} (1-p_0)^{n_2^*-x_2} \leq \alpha \quad (3.3)$$

Though it is theoretically possible to find $r_2^*$, it does not have a closed form solution, and the computation is exhaustive. Instead, the authors propose a normal approximation for the

binomial random variable to ease the computation of power. That is,

$$\sum_{x_2=r_2^*}^{n_2^*} \binom{n_2^*}{x_2} p^{x_2}(1-p)^{n_2^*-x_2} \approx 1 - \Phi\left(\frac{r_2^* - n_2^* p}{\sqrt{n_2^* p(1-p)}}\right) \tag{3.4}$$

Substituting the above equation in for power and type I error, and using Lagrange multipliers and differentiating with respect to $r_2^*$, the problem is then equivalent to solving the equation

$$\left(\frac{1}{p_0(1-p_0)} - \frac{1}{p_1(1-p_1)}\right) r_2^{*2} - \frac{2n_2^*(p_0-p_1)}{(1-p_0)(1-p_1)} r_2^* + \frac{n_2^{*2}(p_0-p_1)}{(1-p_0)(1-p_1)} - 2n_2^* log\left(\frac{\lambda a(x_1)}{b(x_1)}\right) = 0 \tag{3.5}$$

where $a(x_1) = \binom{n_1}{x_1} p_0^{x_1}(1-p_0)^{n_1-x_1}$, $b(x_1) = \binom{n_1}{x_1} p_1^{x_1}(1-p_1)^{n_1-x_1}$, and $\lambda$ is the Lagrange multiplier. The new critical value, $r_2^*$, is then $\max(0, min(r_2^*, n_2^*))$. The authors suggest searching over a reasonable range of $\lambda$ to find a $\lambda$ such that the type I error is as closed to $\alpha$ as possible.

The authors performed a numerical study to compare and found that, in almost all scenarios that were considered, Zeng *et al.*'s method had higher power than Koyama and Chen, and this is mostly because Koyama and Chen's method most often results in a lower type I error rate due to conservative controlling of conditional type I error.

# Chapter 4

# Deviation from Planned Sample Sizes in First Stage

A straightforward solution to under-enrollment in the first stage is to simply wait until the appropriate enrollment has been reached. Because accruement of subjects can be unexpected in the first stage, and some situations require early evaluation of the first stage, it is still imperative that methods are available to handle situations where attained sample sizes differ from the planned sample size in stage I. Green and Dahlberg [12] and Chen and Ng [13] propose methods for inference when first stage sample sizes differ from those planned. Recall that $p_0$ is the highest probability of response that would indicate that the research regimen is uninteresting and $p_1$ is the lowest probability of response that would indicate that the research regimen warrants further investigation. Green and Dahlberg propose a method for sample size deviations by extending the Southwest Oncology Group (SWOG) standard approach to phase II trial design, where the SWOG's standard approach is to use only two-stage designs with a type I error rate of 5% and power of 90%. Southwest Oncology Group's inference method suggestion is to perform a hypothesis test on $H_0 : p = p_1$ versus $H_1 : p < p_1$ in the first stage with type I error rate of 2% and concluding futility if the p-value for this test is $\leq 0.02$. They then suggest testing $H_0 : p = p_0$ versus $H_1 : p > p_0$ in the second stage at the 0.05 level. The type I error rate of 2% corresponds to intuition regarding what constitutes evidence in favor of a hypothesis when the sample size is half of the planned total and we test $p_1$ in the first stage since one is more concerned with missing an active agent [12]. Green and Dahlberg extend the SWOG approach by applying this testing method on the attained design, but by performing an unadjusted 0.055 level test

of $H_0$ based on the attained total sample size at the second stage. This means that they perform a hypothesis test on $H_0 : p = p_1$ versus $H_1 : p < p_1$ at the first stage at the 0.02 level based on the attained sample size ($n_1^{**}$) and then perform the unadjusted 0.055 level test of $H_0$ based on the attained total sample size. [12] The 0.055 level was chosen because of the discreteness of the binomial distribution and to achieve a type I error rate closer to 0.05. The authors demonstrated that this approach controls type I error and achieves desired power only in the limited situation when an overall $\alpha$-level is 0.05, and it is unclear how this method would generalize to any $\alpha$-level [14]. Li *et al.* also indicate that this limited approach, and particularly testing a hypothesis in the first stage with a type I error rate of 2%, is arbitrary and lacks theoretical justification.

Chen and Ng [13] suggest an approach to unplanned sample sizes by considering a range of sample sizes in both the first and second stages. They search these ranges for the Minimax and Optimal designs that satisfy error constraints using the average probability of termination for all possible first stage sample sizes and average expected sample size for all possible stage I and stage II sample size combinations that they consider [13]. Some limitations of this approach are that attained sample sizes may fall outside of the specified ranges, and only the average error probabilities are controlled rather than the actual error probabilities corresponding to the attained sample sizes.

## 4.1 *Chang et al.* Alternative Designs and Adaptation

Chang *et al.* [15] proposed an alternative design that is an extension of two-stage designs in order to handle unplanned sample sizes in both the first and second stages, though we only consider this extension for over- and under-enrollment in the first stage. Their method calculates new critical values for attained sample sizes *a priori*, and thus one is able to create and pre-specify a new design based on a preferred Simon or Admissible design in defense of the events of unplanned sample sizes.

We use this method to pre-specify new designs; that is, we calculate new critical values

for different combinations of possible deviations in sample sizes pre-attainment. Again, let $n_1$, $n_t$, $r_1$, $r_t$, $p_0$, $p_1$, $\alpha$, and $\beta$ be the original, planned design parameters. Now, let $n_1^{**}$ be the attained sample size for the first stage and let $n_2^{**}$ be the attained sample size for the second stage.

Chang *et al.* propose a method for updating the stage I critical value based on the following $\beta$-spending function, where $m$ is the attained sample size in the first stage.

$$\beta(m) = \begin{cases} \beta_1 m/n_1 & \text{if } m \leq n_1 \\ \beta_1 + (\beta - \beta_1)(m - n_1)/n_2 & \text{if } m > n_1 \end{cases} \tag{4.1}$$

Where $\beta_1 = P(X_1 \leq r_1 | n_1, p = p_1)$ is the stage I type II error probability. They then find a new stage one critical value, $s_1$, using this probability spending function such that $P(X_1 \leq s_1 | n_1^{**}) \approx \beta(n_1^{**})$, where $\approx$ means "closest to." After $s_1$ is selected, they then search for an integer for the second stage critical value, $s_t$, that satisfies

$$P(X_1 > s_1, X_t > s_t | n_1^{**}, n_2^{**}, p_0)$$
$$= \sum_{s_1}^{n_1^{**}} P(X_2 > s_t - X_1 | X_1 = x_1) P(X_1 > s_1) \tag{4.2}$$
$$\leq \alpha$$

Chang *et al.*'s design can be used for any $\alpha$-level and is flexible, close to the original design, and preserves the desired traditional frequentist type I error rate.

We modify Chang *et al.*'s method because we prefer to be conservative when straying from a desired Simon's or Admissible design. We modify the Chang *et al.*'s approach by selecting $s_1$ such that the new design's probability of early termination under the null (PET$_0^{**}$) is closest to the planned PET under the null (PET$_0$). This is conservative because when the attained stage I sample size deviates further from the planned sample size, the PET$_0^{**}$ in Chang *et al.*'s designs can get further from the original design's. By selecting the

closest integer such that the $\text{PET}_0^{**}$ is closest to planned, the probability of early termination is greater under large deviations. We select $s_1$ such that

$$P(X_1 \leq s_1 | n_1^{**}, p_0) \approx P(X_1 \leq r_1 | n_1, p_0) \tag{4.3}$$

We then select the stage two critical value, $s_t$, in the same fashion as Chang *et al.*'s design. Another conservative option would be to be choose $s_1$ such that the probability of early termination under the null with the redesign is greater than or equal to the original design. In either case, these conservative designs tend to be close when the attained sample size is close to the original, so we consider the case where the probability of early termination is closest to the original. We call this adaptation to Chang *et al.*'s design "Olson and Koyama's design (OK design)."

## 4.2 Likelihood Design

Briefly, likelihood methods in phase II designs use the likelihood ratio as a measure of evidence [16]. The Law of Likelihood states that "if the first hypothesis, $H_1$, implies that the probability that a random variable $X$ takes the value $x$ is $P_1(X)$, while the second hypothesis, $H_2$, implies that the probability is $P_2(x)$, then the observation $X = x$ is evidence supporting $H_1$ over $H_2$ if and only if $P_1(x) > P_2(x)$, and the likelihood ratio measures the strength of evidence." [17] Define the likelihood function $\text{L}_n(p)$ in phase II trials to be

$$\begin{aligned}
\text{L}_n(p) &= P(X|p,n) \\
&= \binom{n}{x} p^x (1-p)^{n-x} \\
&\propto p^x (1-p)^{n-x}
\end{aligned} \tag{4.4}$$

Here, the likelihood ratio is

$$\begin{aligned} \text{LR}_n &= \frac{\text{L}_n(p_1)}{\text{L}_n(p_2)} \\ &= \frac{p_1^x(1-p_1)^{n-x}}{p_2^x(1-p_2)^{n-x}} \end{aligned} \qquad (4.5)$$

The hypothesis that is better supported is the hypothesis that assigns a higher probability to the observed events [17]. If the likelihood ratio is greater than 1, the evidence favors $H_1$ over $H_2$, and if the likelihood ratio is less than 1, the evidence favors $H_2$ over $H_1$. The likelihood ratio is continuous on the scale of $(0,\infty)$, and this scale can be broken up into categories such as 'weak' and 'strong' evidence [17]. We make the following decision at the conclusion of the study. If the $\text{LR}_n \in [0, 1/k]$, there is evidence for the null hypothesis, if $\text{LR}_n \in [1/k, k]$, there is weak evidence for either hypothesis, and if $\text{LR}_n \in [k, \infty]$, there is evidence for the alternative hypothesis, where $k$ is a value that is a benchmark for distinguishing strength of evidence.

To illustrate the use of the likelihood ratio as a measure of evidence, consider a study with the response rate of patients. Suppose a researcher is interested in looking at the response (yes/no) of 50 patients, while testing the null hypothesis, $H_0 : p = 0.3$, versus $H_1 : p = 0.45$. Because the response of each patient is independent and binary, the probability model is binomial, as in equation (4.4), where $p$ is the unknown probability of response. Suppose 17 responses were observed. Then equation (4.5) gives us the likelihood ratio

$$\frac{0.3^{17}(1-0.3)^{50-17}}{0.45^{17}(1-0.45)^{50-17}} = 2.90 \qquad (4.6)$$

This likelihood ratio means that the data support $H_0$ over $H_1$ by a factor of 2.90. If we used a benchmark of $k = 8$, this would mean the evidence in favor of $H_0$ over $H_1$ is weak because $2.90 < 8$. Figure 4.1 is the likelihood function standardized by the maximum value (a constant) and gives a visual representation of the evidence about $p$ over the parameter space. We see that the null hypothesis value is represented by the circle on the likelihood function $x$-axis where the alternative hypothesis is represented by the square. The ratio of

Figure 4.1: Likelihood function for probability of response
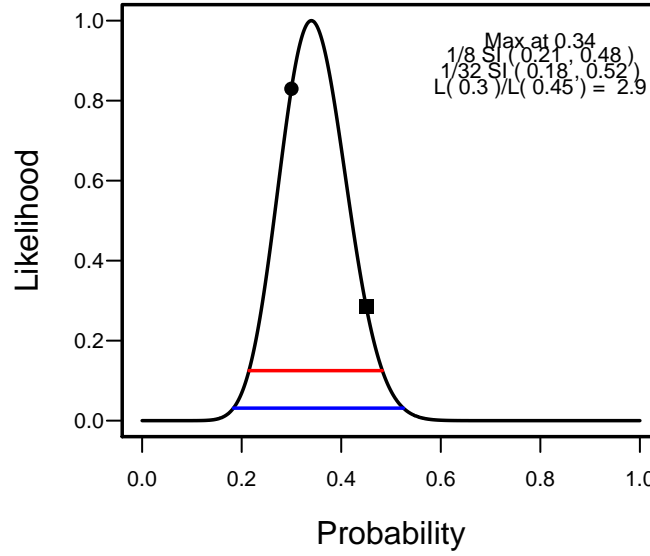


SI means support interval. Red line represents the 1/8 support interval where blue
line represents the 1/32 support interval. Circle represents the value of the null
hypothesis where square represents the value of the alternative hypothesis.

their *y*-axis values is the likelihood ratio. The maximum occurs at the peak of the likelihood

function with a value of 0.34, and this is called the maximum likelihood estimator (MLE).

A 1/8 support interval, shown by the red line, is (0.21, 0.48), and a 1/32 support interval,

shown by the blue line, is (0.18,0.52). A support interval identifies all parameter values

for *p* that are consistent with the data at a certain level (*k*), and the values that are most

consistent with the data occur at the crest of the likelihood function – near the MLE [17].

Furthermore, the probability of observing weak evidence is

$$\gamma_i = P(k_a \leq LR_n \leq k_b | H_i), k_a \leq 1 \leq k_b \tag{4.7}$$

*i*=0 for null hypothesis and *i*=1 for alternative hypothesis, where $k_a$ and $k_b$ are lower and

upper benchmarks, respectively, for description of evidence. Usually, $k_b = k$ and $k_a = 1/k$.

The probability of observing strong evidence is

$$
\tau_i = \begin{cases} P(LR_n > k_b | H_i) & \text{if } i = 1 \\ P(LR_n < k_b | H_i) & \text{if } i = 0 \end{cases} \tag{4.8}
$$

and the probability of observing misleading evidence is

$$
\lambda_i = \begin{cases} P(LR_n > k_b | H_i) & \text{if } i = 0 \\ P(LR_n < k_b | H_i) & \text{if } i = 1 \end{cases} \tag{4.9}
$$

Misleading evidence is observing strong evidence for the incorrect hypothesis over the correct hypothesis, similar to a type I or a type II error. One advantage of a likelihood sequential design is that the universal bound of misleading evidence under the null hypothesis is $P(LR_n > k_b | H_0) \leq \frac{1}{k_b}$ for any n $\geq 1$ when $\frac{1}{k_a} = k_b = k > 1$. This is advantageous because the probability that the trial is stopped with misleading evidence under the null hypothesis at any point in time is less than or equal to $\frac{1}{k_b}$. As data accumulate, the probability of misleading evidence converges to 0, and this probability is often much less than $\frac{1}{k_b}$ [18] [19].

Ayers and Blume [16] consider a phase II two-stage design based on the likelihood. Their likelihood two-stage design will enroll $n_1$ subjects into the first stage. If we observe a likelihood ratio that is $k_{a_1} < LR_{n_1} < k_{b_1}$, where $k_{a_1}$ and $k_{b_1}$ are benchmarks for description of evidence in the first stage, we continue to the second stage. If we observe $LR_{n_1} \leq k_{a_1}$, the study will stop for futility and if we observe $LR_{n_1} \geq k_{b_1}$, the study will stop for efficacy. In stage II, $n_2$ subjects are enrolled, and the decision is based on $LR_{n_t} = LR_{n_1} LR_{n_2}$. If $LR_{n_t}$ is $k_{a_t} < LR_{n_t} < k_{b_t}$, where $k_{a_t}$ and $k_{b_t}$ are benchmarks at the end of stage II, then the study will conclude with weak evidence. The study will conclude with evidence for the alternative hypothesis if $LR_{n_t} \geq k_{b_t}$ and evidence for the null hypothesis if $LR_{n_t} \leq k_{a_t}$. Because these designs are not restricted by error rates, this method offers favorable flexibility for unplanned sample sizes in the first stage. Likewise, one is able to add cohorts and the end

of the second stage when there proves to be weak evidence without penalization because the strength of evidence is unaffected by the number of looks at the data [17].

We compare traditional frequentist and Likelihood two-stage designs by adapting the Likelihood two-stage design to emulate conventional two-stage designs such as Optimal, Minimax, or Admissible designs with binary evidential zones: reject the null or fail to reject the null. In order to do this, one can start with a Simon-like two-stage design and redesign with a likelihood ratio approach by setting

$$
\begin{aligned}
k_{a_1} &= \left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)^{r_1} \left(\frac{1-p_1}{1-p_0}\right)^{n_1} = \left(\frac{1-p_0}{1-p_1}\right)^{r_1-n_1} \left(\frac{p_1}{p_0}\right)^{r_1}, \\
k_{a_t} &= \left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)^{r_t} \left(\frac{1-p_1}{1-p_0}\right)^{n_t} = \left(\frac{1-p_0}{1-p_1}\right)^{r_t-n_t} \left(\frac{p_1}{p_0}\right)^{r_t}, \\
k_{b_1} &= \infty, \\
k_{b_t} &= \infty,
\end{aligned}
\tag{4.10}
$$

where $n_1, n_t, r_1, r_2$ are two-stage design parameters. Then, using $k_{a_j}$ and $k_{b_j}$, we recalculate the critical values, $s_1$ and $s_t$, using

$$
\begin{aligned}
s_1 &= \frac{\log(k_{a_1}) - n_1^{**}\log\left(\frac{1-p_1}{1-p_0}\right)}{\log\left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)} \\
s_t &= \frac{\log(k_{a_t}) - n_t^{**}\log\left(\frac{1-p_1}{1-p_0}\right)}{\log\left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)}
\end{aligned}
\tag{4.11}
$$

where $n_t^{**}$ is the total attained sample size at the end of stage II. If $s_1 < 0$ or $s_t < 0$, they are set equal to zero. It is possible for these critical values to be less than 0 when the study design has small sample sizes and deviation from the planned sample size is extreme. Under the traditional frequentist two-stage design conditions and using these critical values, we can calculate design characteristics for any attained $n_1$ and $n_t$ for a given $p$. The probability

of weak evidence for a probability $p$ at the end of the first stage is

$$\gamma_{1,p} = 1 - B(s_1, n_1^{**}, p) \tag{4.12}$$

and the probability of strong evidence for a given $p$ at the end of stage one is

$$\tau_{1,p} = B(s_1, n_1^{**}, p) \tag{4.13}$$

At the end of the second stage, the probability of weak evidence is

$$\gamma_{t,p} = \sum_{x=s_1+1}^{n_1^{**}} b(x, n_1^{**}, p) - B(s_1 - x, n_t^{**} - n_1^{**}, p) \tag{4.14}$$

and the probability of strong evidence is

$$\tau_{t,p} = \tau_{1,p} + \sum_{x=s_1+1}^{n_1^{*}} \left( b(x, n_1^{*}, p) \times B(s_1 - x, n_t^{**} - n_1^{*}, p) \right) \tag{4.15}$$

The probability of early termination under the null hypothesis is then

$$\text{PET}_0 = \tau_{1,p_0} \tag{4.16}$$

and the expected sample size under the null hypothesis is

$$\text{EN}_0 = n_1^{**} + \gamma_{1,p_0} \times (n_t^{**} - n_1^{**}) \tag{4.17}$$

Ayers and Blume [16] show that the Likelihood designs often preserve type I error rates, although, some cases exceed nominal rates with an increase in power. Under the Likelihood design, error rates tend to be less of an issue because the average of the error rates, $\frac{\alpha+\beta}{2}$, is minimized with the likelihood approach [16]. For the purpose of comparing methods, we do not consider the cases in which cohorts can be added after the second stage. We

also only consider Likelihood redesign methods to emulate traditional frequentist designs – to calculate new critical values – and do not consider pure likelihood method two-stage design as formerly introduced.

We consider new approaches to unplanned sample sizes in the first stage in both the traditional frequentist and likelihood settings. In the interest of prespecifying designs, we focus on deviation from the planned sample size only in the first stage because it is impractical to prespecify limitless combinations of unplanned sample sizes in both the first and second. Because it is desired to stay as closely to the original design as possible for financial and resource planning reasons, we investigate this method while maintaining the original second stage sample size ($n_2$) or original total sample size ($n_t$). Then, the two situations we consider are 1. $n_2^{**} = n_t + n_1^{**}$ and 2. $n_t^{**} = n_1^{**} + n_2$. Lastly, we refer to Chang *et al.*, OK, and Likelihood designs as "attained designs".

# Chapter 5

# Example

In order to compare these new methods for deviation of sample size in the first stage, we first introduce an example. An actual phase II cancer clinical trial was designed using a Likelihood two-stage design. In order to follow the convention, the trial would only stop early for futility. The planned design parameters are $n_1 = 17$, $n_t = 41$, $r_1 = 7$, $r_t = 21$, $p_0 \leq 0.4$, and $p_1 \geq 0.6$. This study design has an expected sample size of 25.6 and a probability of early termination of 64% under the null hypothesis. This is considered an Admissible design and meets the nominal type I error rate, $\alpha = 0.05$, and type II error rate, $\beta = 0.2$ where the actual type I error rate is 0.047. The authors provide alternative interim stopping rules for sample sizes that deviate from the planned design using the Likelihood approach as shown in Table 5.1. These new designs have a probability of early termination under the null that exceed 50% and preserve type I and type II error rates. Using the original likelihood design, but varying $n_1$, one can use Chang *et al.*'s method and Olson and Koyama's method to obtain similar results. The total sample size is equal to the planned total sample size, $n_t^{**} = n_t$ in this case. We compare attained methods' characteristics, in particular, type I error, power, probability of early termination under the null hypothesis, and expected sample size under the null hypothesis.

This example illustrates the comparability of the three attained design methods. Generally, the stopping rules between the Chang *et al.* design, OK design, and the Likelihood design are the same when the probability of early termination under the null exceeds 50%. When $n_1 = 16$, the Olson and Koyama's design gives a more conservative critical value; this is expected by design and the discreteness of the binomial distribution.

Table 5.1: Stopping rules for deviations from first stage planned sample size concrete example

| Design | $r_1$ | $n_1$ | $PET_0$ | $EN_0$ | Likelihood ratio favoring $H_0$ that corresponds to Simon's futility stopping rule |
|---|---|---|---|---|---|
| Likelihood | 6 | 16 | 53% | 27.8 | 1/5.062 |
| Chang *et al.* | 6 | 16 | 53% | 27.8 | |
| Olson and Koyama | 7 | 16 | 72% | 23.1 | |
| Likelihood | 7 | 17 | 64% | 25.6 | 1/3.375 |
| Chang *et al.* | 7 | 17 | 64% | 25.6 | |
| Olson and Koyama | 7 | 17 | 64% | 25.6 | |
| Likelihood | 7 | 18 | 56% | 28 | 1/5.062 |
| Chang *et al.* | 7 | 18 | 56% | 28 | |
| Olson and Koyama | 7 | 18 | 56% | 28 | |
| Likelihood | 8 | 19 | 67% | 26.3 | 1/3.375 |
| Chang *et al.* | 8 | 19 | 67% | 26.3 | |
| Olson and Koyama | 8 | 19 | 67% | 26.3 | |
| Likelihood | 8 | 20 | 60% | 28.5 | 1/5.062 |
| Chang *et al.* | 8 | 20 | 60% | 28.5 | |
| Olson and Koyama | 8 | 20 | 60% | 28.5 | |
| Likelihood | 9 | 21 | 69% | 27.2 | 1/3.375 |
| Chang *et al.* | 9 | 21 | 69% | 27.2 | |
| Olson and Koyama | 9 | 21 | 69% | 27.2 | |
| Likelihood | 10 | 23 | 71% | 28.2 | 1/3.375 |
| Chang *et al.* | 10 | 23 | 71% | 28.2 | |
| Olson and Koyama | 10 | 23 | 71% | 28.2 | |

# Chapter 6

# Results

We compare the methods of Chang *et al.*, Olson and Koyama, and the Likelihood by first selecting either an Admissible, Minimax, or Optimal two-stage design. Each method is applied to deviations in first stage sample size of up to $\pm10$. We suggest keeping the original total planned sample size or the original planned second stage sample size the same when utilizing Chang *et al.*'s and Olson and Koyama's methods to stay close to the original design. We choose to keep total sample size the same in our investigation because this approach is more realistic when resources and budget are fixed for the planned total sample size and this inhibits the ability to stray radically from the planned design.

We present results that are limited to our primary problem of interest in Tables 6.1 through 6.6. In each design table, the planned design is specified, and the first stage sample size varies from planned, while maintaining the original total sample size. We compare attained methods' characteristics, in particular, type I error, power, probability of early termination under the null hypothesis, and expected sample size under the null hypothesis.

Table 6.1 displays a planned Admissible design with varying first stage sample size $\pm$ 10. We notice that under low $p_0$ and $p_1$ (0.1 and 0.25, respectively), $s_1$ will vary between each method. Though, power and type I error are likely to be similar between and within each attained method, and expected sample size is also consistent. The Likelihood and Chang *et al.* method are at risk of low probability of early termination, especially when the sample size is lower than planned. Table 6.2 shows an Optimal design when $p_0$ is 0.5. Between attained designs, particularly when there is over-accrual, $s_1$ is inconsistent. We particularly see a large difference when $n_1^{**} = n_1 + 10$ or more between the Chang *et al.* design and the other two. The Likelihood design is anticonservative in type I error and con-

servative in type II error here, displaying a $\alpha^{**} > \alpha$ and $1 - \beta^{**} > 1 - \beta$. The Chang *et al.* and OK designs both have a conservative type I error for all sample size deviations, though Chang *et al.* and Likelihood designs maintain higher power than Olson and Koyama's. Probability of early termination is closest to the original under Olson and Koyama's and thus have a much lower expected sample size under the null hypothesis, with as much as a difference of approximately 23.

Table 6.3 displays results from a planned Minimax design when $p_0$ is larger than 0.5. Here, the Likelihood design has desirable properties where type I error and power are consistently close to the planned design for all deviations in sample size. Though, when the sample size is severely under-accrued, the probability of early termination nearly halves. The expected sample size is consistent between designs. The Chang *et al.* and OK designs stray from the planned nominal type I and type II error rates when there is over-accrual. The PET$_0$ for the Chang *et al.* design varies significantly between deviations.

Table 6.4 through 6.6 display results for planned designs when $\alpha = \beta = 0.1$. Table 6.4 displays attained design characteristics for deviations in sample size when the planned first stage sample size is low. In all three attained designs, we see that when the attained sample size is lower than planned, there is a significant drop in power and a moderate to severe drop in type I error. The probability of early termination almost occurs with probability 1 when the attained sample size is $n_1^{**} = 1$. In practice, though, accrual lower than planned here is not practical. When there is over-accrual, attained design characteristics are not concerning.

Table 6.5 illustrates the similarity between attained designs when $p_0 = 0.3$. All designs and their deviations are relatively consistent in type I error, power, and PET$_0$. Though, the Olson and Koyama's design is most consistent in the probability of early termination with the planned design, but we see a conservative deviation in type I error for large over-accrual. Table 6.6 displays similar results as Table 6.3.

Table 6.1: Attained design characteristics from deviation of Admissible II stage design ($p_0 = 0.1$, $p_1 = 0.25$, $\alpha = 0.05$, $\beta = 0.20$)

| Planned Design | | | | | | | | Attained Sample Size / Chang Design | | | | | | | Redesign / Olson and Koyama Design | | | | | | Likelihood Design | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $p_0$ | $p_1$ | $n_1$ | $n$ | $r_1$ | $r_t$ | $PET_0$ | $EN_0$ | $n_1^{**}$ | $s_1$ | $s_t$ | $\alpha^{**}$ | $1-\beta^{**}$ | $PET_0^{**}$ | $EN_0^{**}$ | $s_1$ | $s_t$ | $\alpha^{**}$ | $1-\beta^{**}$ | $PET_0^{**}$ | $EN_0^{**}$ | $s_1$ | $s_t$ | $\alpha^{**}$ | $1-\beta^{**}$ | $PET_0^{**}$ | $EN_0^{**}$ |
| 0.1 | 0.25 | 15 | 41 | 1 | 7 | 0.549 | 26.7 | 5 | 0 | 7 | 0.034 | 0.671 | 0.590 | 19.7 | 0 | 7 | 0.034 | 0.671 | 0.590 | 19.7 | 0 | 7 | 0.034 | 0.671 | 0.590 | 19.7 |
| 0.1 | 0.25 | 15 | 41 | 1 | 7 | 0.549 | 26.7 | 7 | 0 | 7 | 0.040 | 0.754 | 0.478 | 24.7 | 0 | 7 | 0.040 | 0.754 | 0.478 | 24.7 | 0 | 7 | 0.040 | 0.754 | 0.478 | 24.7 |
| 0.1 | 0.25 | 15 | 41 | 1 | 7 | 0.549 | 26.7 | 9 | 0 | 7 | 0.043 | 0.797 | 0.387 | 28.6 | 0 | 7 | 0.043 | 0.797 | 0.387 | 28.6 | 0 | 7 | 0.043 | 0.797 | 0.387 | 28.6 |
| 0.1 | 0.25 | 15 | 41 | 1 | 7 | 0.549 | 26.7 | 11 | 0 | 7 | 0.045 | 0.819 | 0.314 | 31.6 | 1 | 7 | 0.035 | 0.718 | 0.697 | 20.1 | 0 | 7 | 0.045 | 0.819 | 0.314 | 31.6 |
| 0.1 | 0.25 | 15 | 41 | 1 | 7 | 0.549 | 26.7 | 13 | 0 | 7 | 0.046 | 0.830 | 0.254 | 33.9 | 1 | 7 | 0.040 | 0.771 | 0.621 | 23.6 | 1 | 7 | 0.046 | 0.830 | 0.254 | 33.9 |
| 0.1 | 0.25 | 15 | 41 | 1 | 7 | 0.549 | 26.7 | 15 | 1 | 7 | 0.043 | 0.803 | 0.549 | 26.7 | 1 | 7 | 0.043 | 0.803 | 0.549 | 26.7 | 1 | 7 | 0.043 | 0.803 | 0.549 | 26.7 |
| 0.1 | 0.25 | 15 | 41 | 1 | 7 | 0.549 | 26.7 | 17 | 1 | 7 | 0.045 | 0.821 | 0.482 | 29.4 | 1 | 7 | 0.045 | 0.821 | 0.482 | 29.4 | 1 | 7 | 0.045 | 0.821 | 0.482 | 29.4 |
| 0.1 | 0.25 | 15 | 41 | 1 | 7 | 0.549 | 26.7 | 19 | 2 | 7 | 0.041 | 0.792 | 0.705 | 25.5 | 1 | 7 | 0.046 | 0.831 | 0.420 | 31.8 | 1 | 7 | 0.046 | 0.831 | 0.420 | 31.8 |
| 0.1 | 0.25 | 15 | 41 | 1 | 7 | 0.549 | 26.7 | 21 | 2 | 7 | 0.044 | 0.814 | 0.648 | 28.0 | 2 | 7 | 0.044 | 0.814 | 0.648 | 28.0 | 1 | 7 | 0.047 | 0.836 | 0.365 | 33.7 |
| 0.1 | 0.25 | 15 | 41 | 1 | 7 | 0.549 | 26.7 | 23 | 3 | 7 | 0.040 | 0.785 | 0.807 | 26.5 | 2 | 7 | 0.046 | 0.827 | 0.592 | 30.3 | 2 | 7 | 0.046 | 0.827 | 0.592 | 30.3 |
| 0.1 | 0.25 | 15 | 41 | 1 | 7 | 0.549 | 26.7 | 25 | 3 | 7 | 0.043 | 0.810 | 0.764 | 28.8 | 2 | 7 | 0.047 | 0.834 | 0.537 | 32.4 | 2 | 7 | 0.047 | 0.834 | 0.537 | 32.4 |

Table 6.2: Attained design characteristics from deviation of Simon's Optimal II stage design ($p_0 = 0.5$, $p_1 = 0.65$, $\alpha = 0.05$, $\beta = 0.2$)

| Planned Design | | | | | | | | Attained Sample Size / Chang Design | | | | | | | Redesign / Olson and Koyama Design | | | | | | Likelihood Design | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $p_0$ | $p_1$ | $n_1$ | $n$ | $r_1$ | $r_t$ | $PET_0$ | $EN_0$ | $n_1^{**}$ | $s_1$ | $s_t$ | $\alpha^{**}$ | $1-\beta^{**}$ | $PET_0^{**}$ | $EN_0^{**}$ | $s_1$ | $s_t$ | $\alpha^{**}$ | $1-\beta^{**}$ | $PET_0^{**}$ | $EN_0^{**}$ | $s_1$ | $s_t$ | $\alpha^{**}$ | $1-\beta^{**}$ | $PET_0^{**}$ | $EN_0^{**}$ |
| 0.5 | 0.65 | 28 | 83 | 15 | 48 | 0.714 | 43.7 | 18 | 8 | 49 | 0.036 | 0.815 | 0.407 | 56.5 | 10 | 48 | 0.037 | 0.685 | 0.760 | 33.6 | 9 | 48 | 0.048 | 0.796 | 0.593 | 44.5 |
| 0.5 | 0.65 | 28 | 83 | 15 | 48 | 0.714 | 43.7 | 20 | 10 | 48 | 0.050 | 0.811 | 0.588 | 46.0 | 11 | 48 | 0.039 | 0.716 | 0.748 | 35.8 | 10 | 48 | 0.050 | 0.811 | 0.588 | 46.0 |
| 0.5 | 0.65 | 28 | 83 | 15 | 48 | 0.714 | 43.7 | 22 | 11 | 49 | 0.034 | 0.788 | 0.584 | 47.4 | 12 | 48 | 0.042 | 0.743 | 0.738 | 38.0 | 11 | 48 | 0.051 | 0.824 | 0.584 | 47.4 |
| 0.5 | 0.65 | 28 | 83 | 15 | 48 | 0.714 | 43.7 | 24 | 12 | 49 | 0.034 | 0.798 | 0.581 | 48.7 | 13 | 48 | 0.044 | 0.765 | 0.729 | 40.0 | 12 | 48 | 0.052 | 0.835 | 0.581 | 48.7 |
| 0.5 | 0.65 | 28 | 83 | 15 | 48 | 0.714 | 43.7 | 26 | 14 | 48 | 0.045 | 0.785 | 0.721 | 41.9 | 14 | 48 | 0.045 | 0.785 | 0.721 | 41.9 | 13 | 48 | 0.053 | 0.845 | 0.577 | 50.1 |
| 0.5 | 0.65 | 28 | 83 | 15 | 48 | 0.714 | 43.7 | 28 | 15 | 48 | 0.047 | 0.802 | 0.714 | 43.7 | 15 | 48 | 0.047 | 0.802 | 0.714 | 43.7 | 15 | 48 | 0.047 | 0.802 | 0.714 | 43.7 |
| 0.5 | 0.65 | 28 | 83 | 15 | 48 | 0.714 | 43.7 | 30 | 16 | 48 | 0.049 | 0.816 | 0.708 | 45.5 | 16 | 48 | 0.049 | 0.816 | 0.708 | 45.5 | 16 | 48 | 0.049 | 0.816 | 0.708 | 45.5 |
| 0.5 | 0.65 | 28 | 83 | 15 | 48 | 0.714 | 43.7 | 32 | 17 | 49 | 0.033 | 0.793 | 0.702 | 47.2 | 17 | 49 | 0.033 | 0.793 | 0.702 | 47.2 | 17 | 48 | 0.050 | 0.828 | 0.702 | 47.2 |
| 0.5 | 0.65 | 28 | 83 | 15 | 48 | 0.714 | 43.7 | 34 | 19 | 48 | 0.043 | 0.782 | 0.804 | 43.6 | 18 | 49 | 0.034 | 0.803 | 0.696 | 48.9 | 18 | 48 | 0.051 | 0.839 | 0.696 | 48.9 |
| 0.5 | 0.65 | 28 | 83 | 15 | 48 | 0.714 | 43.7 | 36 | 20 | 48 | 0.045 | 0.798 | 0.797 | 45.5 | 19 | 49 | 0.035 | 0.811 | 0.691 | 50.5 | 19 | 48 | 0.053 | 0.848 | 0.691 | 50.5 |
| 0.5 | 0.65 | 28 | 83 | 15 | 48 | 0.714 | 43.7 | 38 | 21 | 48 | 0.047 | 0.813 | 0.791 | 47.4 | 20 | 49 | 0.035 | 0.818 | 0.686 | 52.1 | 20 | 48 | 0.054 | 0.856 | 0.686 | 52.1 |

Table 6.3: Attained design characteristics from deviation of Simon's Minimax II stage design ($p_0 = 0.75$, $p_1 = 0.9$, $\alpha = 0.05$, $\beta = 0.2$)

| | Planned Design | | | | | | | | Attained Sample Size | Chang Design | | | | | | Redesign Olson and Koyama Design | | | | | | Likelihood Design | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $p_0$ | $p_1$ | $n_1$ | $n$ | $r_1$ | $r_t$ | PET$_0$ | EN$_0$ | $n_1^{**}$ | $s_1$ | $s_t$ | $\alpha^{**}$ | $1-\beta^{**}$ | PET$_0^{**}$ | EN$_0^{**}$ | $s_1$ | $s_t$ | $\alpha^{**}$ | $1-\beta^{**}$ | PET$_0^{**}$ | EN$_0^{**}$ | $s_1$ | $s_t$ | $\alpha^{**}$ | $1-\beta^{**}$ | PET$_0^{**}$ | EN$_0^{**}$ |
| 0.75 | 0.9 | 22 | 39 | 17 | 33 | 0.677 | 27.5 | 12 | 8 | 34 | 0.019 | 0.648 | 0.351 | 29.5 | 9 | 33 | 0.045 | 0.763 | 0.609 | 22.5 | 8 | 33 | 0.050 | 0.805 | 0.351 | 29.5 |
| 0.75 | 0.9 | 22 | 39 | 17 | 33 | 0.677 | 27.5 | 14 | 10 | 33 | 0.050 | 0.800 | 0.479 | 27.0 | 11 | 33 | 0.042 | 0.738 | 0.719 | 21.0 | 10 | 33 | 0.050 | 0.800 | 0.479 | 27.0 |
| 0.75 | 0.9 | 22 | 39 | 17 | 33 | 0.677 | 27.5 | 16 | 12 | 33 | 0.048 | 0.792 | 0.595 | 25.3 | 12 | 33 | 0.048 | 0.792 | 0.595 | 25.3 | 11 | 33 | 0.051 | 0.809 | 0.370 | 30.5 |
| 0.75 | 0.9 | 22 | 39 | 17 | 33 | 0.677 | 27.5 | 18 | 13 | 34 | 0.019 | 0.650 | 0.481 | 28.9 | 14 | 33 | 0.047 | 0.782 | 0.694 | 24.4 | 13 | 33 | 0.051 | 0.807 | 0.481 | 29.0 |
| 0.75 | 0.9 | 22 | 39 | 17 | 33 | 0.677 | 27.5 | 20 | 15 | 34 | 0.019 | 0.650 | 0.585 | 27.9 | 15 | 34 | 0.019 | 0.650 | 0.585 | 27.9 | 15 | 33 | 0.050 | 0.805 | 0.585 | 27.9 |
| 0.75 | 0.9 | 22 | 39 | 17 | 33 | 0.677 | 27.5 | 22 | 17 | 33 | 0.050 | 0.802 | 0.677 | 27.5 | 17 | 33 | 0.050 | 0.802 | 0.677 | 27.5 | 17 | 33 | 0.050 | 0.802 | 0.677 | 27.5 |
| 0.75 | 0.9 | 22 | 39 | 17 | 33 | 0.677 | 27.5 | 24 | 19 | 33 | 0.049 | 0.798 | 0.753 | 27.7 | 19 | 33 | 0.049 | 0.798 | 0.753 | 27.7 | 18 | 33 | 0.051 | 0.810 | 0.578 | 30.3 |
| 0.75 | 0.9 | 22 | 39 | 17 | 33 | 0.677 | 27.5 | 26 | 21 | 33 | 0.048 | 0.791 | 0.816 | 28.4 | 20 | 34 | 0.019 | 0.650 | 0.663 | 30.4 | 20 | 33 | 0.051 | 0.810 | 0.663 | 30.4 |
| 0.75 | 0.9 | 22 | 39 | 17 | 33 | 0.677 | 27.5 | 28 | 23 | 33 | 0.046 | 0.782 | 0.865 | 29.5 | 22 | 34 | 0.019 | 0.650 | 0.736 | 30.9 | 22 | 33 | 0.051 | 0.810 | 0.736 | 30.9 |
| 0.75 | 0.9 | 22 | 39 | 17 | 33 | 0.677 | 27.5 | 30 | 25 | 33 | 0.043 | 0.770 | 0.902 | 30.9 | 23 | 34 | 0.019 | 0.650 | 0.652 | 33.1 | 23 | 33 | 0.051 | 0.810 | 0.652 | 33.1 |
| 0.75 | 0.9 | 22 | 39 | 17 | 33 | 0.677 | 27.5 | 32 | 26 | 34 | 0.019 | 0.650 | 0.847 | 33.0 | 25 | 34 | 0.019 | 0.650 | 0.722 | 33.9 | 25 | 33 | 0.051 | 0.810 | 0.722 | 33.9 |

Table 6.4: Attained design characteristics from deviation of Simon's Optimal II stage design ($p_0 = 0.05$, $p_1 = 0.25$, $\alpha = 0.1$, $\beta = 0.1$)

| | Planned Design | | | | | | | | Attained Sample Size | Chang Design | | | | | | Redesign Olson and Koyama Design | | | | | | Likelihood Design | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $p_0$ | $p_1$ | $n_1$ | $n$ | $r_1$ | $r_t$ | PET$_0$ | EN$_0$ | $n_1^{**}$ | $s_1$ | $s_t$ | $\alpha^{**}$ | $1-\beta^{**}$ | PET$_0^{**}$ | EN$_0^{**}$ | $s_1$ | $s_t$ | $\alpha^{**}$ | $1-\beta^{**}$ | PET$_0^{**}$ | EN$_0^{**}$ | $s_1$ | $s_t$ | $\alpha^{**}$ | $1-\beta^{**}$ | PET$_0^{**}$ | EN$_0^{**}$ |
| 0.05 | 0.2 | 9 | 24 | 0 | 2 | 0.630 | 14.5 | 1 | 0 | 0 | 0.050 | 0.200 | 0.950 | 2.2 | 0 | 0 | 0.050 | 0.200 | 0.950 | 2.2 | 0 | 2 | 0.016 | 0.192 | 0.950 | 2.2 |
| 0.05 | 0.2 | 9 | 24 | 0 | 2 | 0.630 | 14.5 | 3 | 0 | 1 | 0.097 | 0.484 | 0.857 | 6.0 | 0 | 1 | 0.097 | 0.484 | 0.857 | 6.0 | 0 | 2 | 0.043 | 0.465 | 0.857 | 6.0 |
| 0.05 | 0.2 | 9 | 24 | 0 | 2 | 0.630 | 14.5 | 5 | 0 | 2 | 0.064 | 0.635 | 0.774 | 9.3 | 0 | 2 | 0.064 | 0.635 | 0.774 | 9.3 | 0 | 2 | 0.064 | 0.635 | 0.774 | 9.3 |
| 0.05 | 0.2 | 9 | 24 | 0 | 2 | 0.630 | 14.5 | 7 | 0 | 2 | 0.081 | 0.741 | 0.698 | 12.1 | 0 | 2 | 0.081 | 0.741 | 0.698 | 12.1 | 0 | 2 | 0.081 | 0.741 | 0.698 | 12.1 |
| 0.05 | 0.2 | 9 | 24 | 0 | 2 | 0.630 | 14.5 | 9 | 0 | 2 | 0.093 | 0.805 | 0.630 | 14.5 | 0 | 2 | 0.093 | 0.805 | 0.630 | 14.5 | 0 | 2 | 0.093 | 0.805 | 0.630 | 14.5 |
| 0.05 | 0.2 | 9 | 24 | 0 | 2 | 0.630 | 14.5 | 11 | 0 | 3 | 0.028 | 0.714 | 0.569 | 16.6 | 0 | 3 | 0.028 | 0.714 | 0.569 | 16.6 | 0 | 2 | 0.102 | 0.843 | 0.569 | 16.6 |
| 0.05 | 0.2 | 9 | 24 | 0 | 2 | 0.630 | 14.5 | 13 | 0 | 3 | 0.029 | 0.727 | 0.513 | 18.4 | 0 | 3 | 0.029 | 0.727 | 0.513 | 18.4 | 0 | 2 | 0.108 | 0.864 | 0.513 | 18.4 |
| 0.05 | 0.2 | 9 | 24 | 0 | 2 | 0.630 | 14.5 | 15 | 1 | 2 | 0.086 | 0.802 | 0.829 | 16.5 | 0 | 3 | 0.029 | 0.733 | 0.463 | 19.8 | 0 | 2 | 0.112 | 0.876 | 0.463 | 19.8 |
| 0.05 | 0.2 | 9 | 24 | 0 | 2 | 0.630 | 14.5 | 17 | 1 | 2 | 0.098 | 0.842 | 0.792 | 18.5 | 1 | 2 | 0.098 | 0.842 | 0.792 | 18.5 | 0 | 2 | 0.114 | 0.882 | 0.418 | 21.1 |

Table 6.5: Attained design characteristics from deviation of Admissible II stage design ($p_0 = 0.3$, $p_1 = 0.45$, $\alpha = 0.1$, $\beta = 0.1$)

| Planned Design | | | | | | | | Attained Sample Size Chang Design | | | | | | | Redesign Olson and Koyama Design | | | | | | Likelihood Design | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $p_0$ | $p_1$ | $n_1$ | $n$ | $r_1$ | $r_t$ | $PET_0$ | $EN_0$ | $n_1^{**}$ | $s_1$ | $s_r$ | $\alpha^{**}$ | $1-\beta^{**}$ | $PET_0^{**}$ | $EN_0^{**}$ | $s_1$ | $s_r$ | $\alpha^{**}$ | $1-\beta^{**}$ | $PET_0^{**}$ | $EN_0^{**}$ | $s_1$ | $s_r$ | $\alpha^{**}$ | $1-\beta^{**}$ | $PET_0^{**}$ | $EN_0^{**}$ |
| 0.3 | 0.45 | 37 | 72 | 11 | 26 | 0.566 | 52.2 | 27 | 7 | 26 | 0.099 | 0.902 | 0.411 | 53.5 | 8 | 26 | 0.090 | 0.871 | 0.577 | 46.0 | 7 | 26 | 0.099 | 0.902 | 0.411 | 53.5 |
| 0.3 | 0.45 | 37 | 72 | 11 | 26 | 0.566 | 52.2 | 29 | 8 | 26 | 0.097 | 0.897 | 0.479 | 51.4 | 9 | 26 | 0.087 | 0.862 | 0.636 | 44.7 | 8 | 26 | 0.097 | 0.897 | 0.479 | 51.4 |
| 0.3 | 0.45 | 37 | 72 | 11 | 26 | 0.566 | 52.2 | 31 | 8 | 27 | 0.065 | 0.870 | 0.386 | 56.2 | 9 | 26 | 0.095 | 0.892 | 0.542 | 49.8 | 8 | 26 | 0.101 | 0.910 | 0.386 | 56.2 |
| 0.3 | 0.45 | 37 | 72 | 11 | 26 | 0.566 | 52.2 | 33 | 9 | 26 | 0.100 | 0.907 | 0.450 | 54.5 | 10 | 26 | 0.093 | 0.886 | 0.599 | 48.6 | 9 | 26 | 0.100 | 0.907 | 0.450 | 54.5 |
| 0.3 | 0.45 | 37 | 72 | 11 | 26 | 0.566 | 52.2 | 35 | 10 | 26 | 0.099 | 0.904 | 0.510 | 53.1 | 10 | 26 | 0.099 | 0.904 | 0.510 | 53.1 | 10 | 26 | 0.099 | 0.904 | 0.510 | 53.1 |
| 0.3 | 0.45 | 37 | 72 | 11 | 26 | 0.566 | 52.2 | 37 | 11 | 26 | 0.097 | 0.901 | 0.566 | 52.2 | 11 | 26 | 0.097 | 0.901 | 0.566 | 52.2 | 11 | 26 | 0.097 | 0.901 | 0.566 | 52.2 |
| 0.3 | 0.45 | 37 | 72 | 11 | 26 | 0.566 | 52.2 | 39 | 12 | 26 | 0.095 | 0.897 | 0.618 | 51.6 | 12 | 26 | 0.095 | 0.897 | 0.618 | 51.6 | 11 | 26 | 0.101 | 0.912 | 0.482 | 56.1 |
| 0.3 | 0.45 | 37 | 72 | 11 | 26 | 0.566 | 52.2 | 41 | 13 | 26 | 0.094 | 0.893 | 0.665 | 51.4 | 12 | 27 | 0.065 | 0.871 | 0.536 | 55.4 | 12 | 26 | 0.100 | 0.910 | 0.536 | 55.4 |
| 0.3 | 0.45 | 37 | 72 | 11 | 26 | 0.566 | 52.2 | 43 | 14 | 26 | 0.091 | 0.889 | 0.708 | 51.5 | 13 | 26 | 0.099 | 0.908 | 0.587 | 55.0 | 13 | 26 | 0.099 | 0.908 | 0.587 | 55.0 |
| 0.3 | 0.45 | 37 | 72 | 11 | 26 | 0.566 | 52.2 | 45 | 14 | 26 | 0.098 | 0.906 | 0.635 | 54.9 | 13 | 27 | 0.066 | 0.875 | 0.509 | 58.2 | 13 | 26 | 0.103 | 0.915 | 0.509 | 58.2 |
| 0.3 | 0.45 | 37 | 72 | 11 | 26 | 0.566 | 52.2 | 47 | 15 | 26 | 0.097 | 0.904 | 0.678 | 55.1 | 14 | 27 | 0.066 | 0.874 | 0.559 | 58.0 | 14 | 26 | 0.102 | 0.914 | 0.559 | 58.1 |

Table 6.6: Attained design characteristics from deviation of Simon's Minimax II stage design ($p_0 = 0.75$, $p_1 = 0.9$, $\alpha = 0.1$, $\beta = 0.1$)

| Planned Design | | | | | | | | Attained Sample Size Chang Design | | | | | | | Redesign Olson and Koyama Design | | | | | | Likelihood Design | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $p_0$ | $p_1$ | $n_1$ | $n$ | $r_1$ | $r_t$ | $PET_0$ | $EN_0$ | $n_1^{**}$ | $s_1$ | $s_r$ | $\alpha^{**}$ | $1-\beta^{**}$ | $PET_0^{**}$ | $EN_0^{**}$ | $s_1$ | $s_r$ | $\alpha^{**}$ | $1-\beta^{**}$ | $PET_0^{**}$ | $EN_0^{**}$ | $s_1$ | $s_r$ | $\alpha^{**}$ | $1-\beta^{**}$ | $PET_0^{**}$ | $EN_0^{**}$ |
| 0.75 | 0.9 | 27 | 40 | 20 | 33 | 0.529 | 33.1 | 17 | 11 | 33 | 0.096 | 0.900 | 0.235 | 34.6 | 12 | 33 | 0.094 | 0.895 | 0.426 | 30.2 | 11 | 33 | 0.096 | 0.900 | 0.235 | 34.6 |
| 0.75 | 0.9 | 27 | 40 | 20 | 33 | 0.529 | 33.1 | 19 | 13 | 33 | 0.096 | 0.900 | 0.332 | 33.0 | 14 | 33 | 0.092 | 0.890 | 0.535 | 28.8 | 13 | 33 | 0.096 | 0.900 | 0.332 | 33.0 |
| 0.75 | 0.9 | 27 | 40 | 20 | 33 | 0.529 | 33.1 | 21 | 15 | 33 | 0.095 | 0.899 | 0.433 | 31.8 | 15 | 33 | 0.095 | 0.899 | 0.433 | 31.8 | 14 | 33 | 0.096 | 0.900 | 0.256 | 35.1 |
| 0.75 | 0.9 | 27 | 40 | 20 | 33 | 0.529 | 33.1 | 23 | 16 | 33 | 0.096 | 0.900 | 0.346 | 34.1 | 17 | 33 | 0.095 | 0.898 | 0.532 | 31.0 | 16 | 33 | 0.096 | 0.900 | 0.346 | 34.1 |
| 0.75 | 0.9 | 27 | 40 | 20 | 33 | 0.529 | 33.1 | 25 | 18 | 33 | 0.096 | 0.900 | 0.439 | 33.4 | 18 | 33 | 0.096 | 0.900 | 0.439 | 33.4 | 18 | 33 | 0.096 | 0.900 | 0.439 | 33.4 |
| 0.75 | 0.9 | 27 | 40 | 20 | 33 | 0.529 | 33.1 | 27 | 20 | 33 | 0.096 | 0.900 | 0.529 | 33.1 | 20 | 33 | 0.096 | 0.900 | 0.529 | 33.1 | 20 | 33 | 0.096 | 0.900 | 0.529 | 33.1 |
| 0.75 | 0.9 | 27 | 40 | 20 | 33 | 0.529 | 33.1 | 29 | 22 | 33 | 0.096 | 0.900 | 0.613 | 33.3 | 22 | 33 | 0.096 | 0.900 | 0.613 | 33.3 | 21 | 33 | 0.096 | 0.900 | 0.443 | 35.1 |
| 0.75 | 0.9 | 27 | 40 | 20 | 33 | 0.529 | 33.1 | 31 | 24 | 33 | 0.096 | 0.900 | 0.688 | 33.8 | 23 | 33 | 0.096 | 0.900 | 0.527 | 35.3 | 23 | 33 | 0.096 | 0.900 | 0.527 | 35.3 |
| 0.75 | 0.9 | 27 | 40 | 20 | 33 | 0.529 | 33.1 | 33 | 26 | 33 | 0.096 | 0.900 | 0.753 | 34.7 | 25 | 33 | 0.096 | 0.900 | 0.606 | 35.8 | 25 | 33 | 0.096 | 0.900 | 0.606 | 35.8 |
| 0.75 | 0.9 | 27 | 40 | 20 | 33 | 0.529 | 33.1 | 35 | 28 | 33 | 0.096 | 0.900 | 0.808 | 36.0 | 26 | 33 | 0.096 | 0.900 | 0.526 | 37.4 | 26 | 33 | 0.096 | 0.900 | 0.526 | 37.4 |
| 0.75 | 0.9 | 27 | 40 | 20 | 33 | 0.529 | 33.1 | 37 | 30 | 33 | 0.096 | 0.900 | 0.853 | 37.4 | 28 | 33 | 0.096 | 0.900 | 0.600 | 38.2 | 28 | 33 | 0.096 | 0.900 | 0.600 | 38.2 |

Figures 6.1 and 6.2 display Monte Carlo simulation results for an average of 20 two-stage designs that are Minimax, Optimal, or Admissible with $\alpha = 0.05, \beta = 0.20$ and $n_t^{**} = n_1^{**} + n_2$ and $n_t^{**} = n_t$, respectively. Therefore, each point will be an average of attained type I error or power for 20 two-stage designs under different sample size deviations. These simulations further support our choice to keep the planned total sample size as planned because the error rates are closer to the original design more often for sample size deviations $\pm 10$ when maintaining the original total sample size rather than original second stage sample size (Figures 6.1 and 6.2). In Figure 6.1 we see that the average type I error rate for the Likelihood design is often above 0.05; we see a parabolic decrease in power in Chang *et al.*'s method and a sharp linear increase in power in the OK method, which is not a fruitful result.

When keeping the original total sample size as planned, Figure 6.2 shows that, on average, the Likelihood two-stage design has power above the nominal power and type I error rate below the nominal $\alpha$-level for all sample size deviations $\pm 10$. Both Chang *et al.*'s and Olson and Koyama's methods are conservative in type I error for all sample size deviations and thus are more likely to suffer in power. Figure 6.3 shows a similar Monte Carlo simulation, though it displays the average of the average type I error rate and the average type II error rate. The simulation confirms that the Likelihood design minimizes the average of the error rates, while Chang *et al.*'s method performs better in this sense when there is under-enrollment, while Olson and Koyama performs better when there is over-enrollment.
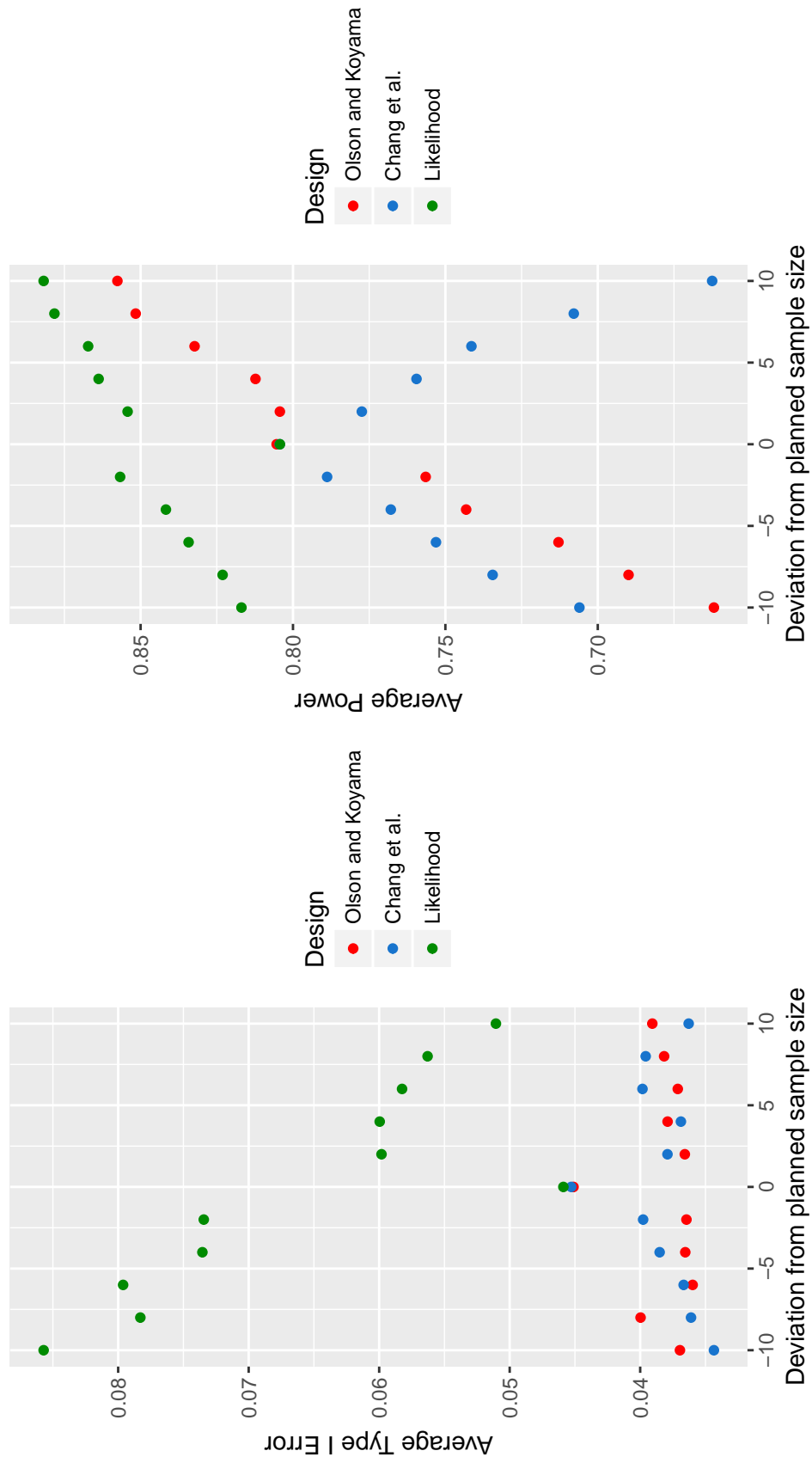
Figure 6.1: Monte Carlo Simulation of Average Type I Error Rates and Power of 20 Simon-like Designs when Stage I Sample Size Deviates from Planned for Attained Designs ($n_t^{**} = n_1^{**} + n_2$) Number of Simulations = 10000.
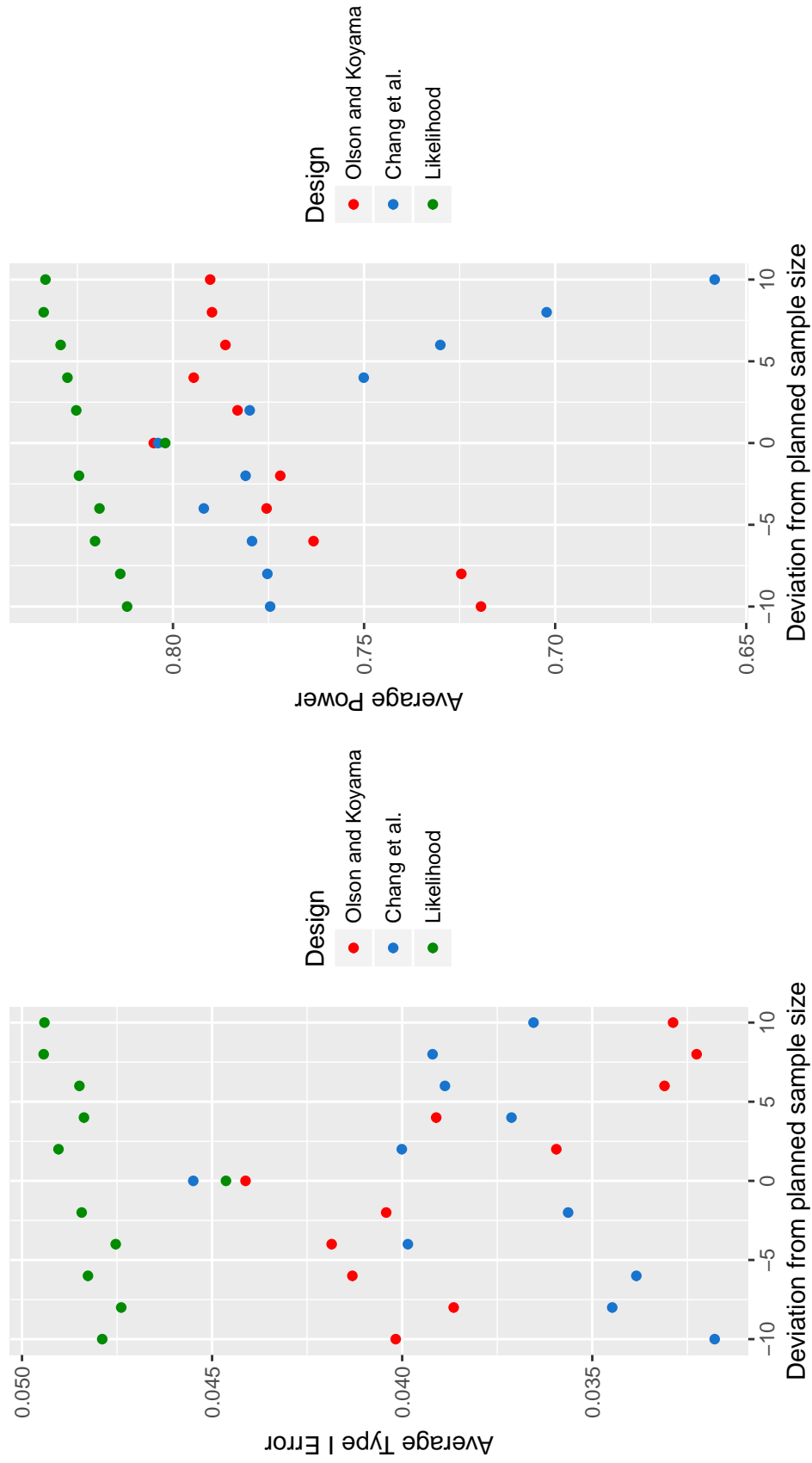
Figure 6.2: Monte Carlo Simulation of Average Type I Error Rates and Power of 20 Simon-like Designs when Stage I Sample Size Deviates from Planned for Attained Designs ($n_t^{**} = n_t$) Number of Simulations = 10000.

Figure 6.3: Monte Carlo Simulation of the Average of the Average Type I and Type II Error Rates of 20 Simon-like Designs when Stage I Sample Size Deviates from Planned for Attained Designs ($n_t^{**} = n_t$) Number of Simulations = 10000.
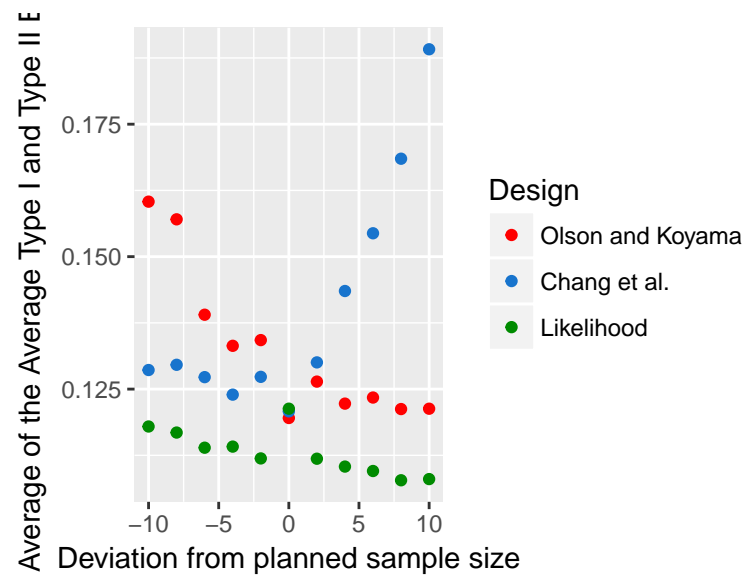
# Chapter 7

# Discussion and Conclusion

Deviations from the planned second stage sample size has been better studied than deviations from the planned first stage sample size. Some methods have been proposed on decision rules, and Koyama and Chen had introduced a redesign when the first stage sample size is as planned. Because the calculation of a proper p-value in this case is more straightforward than when stage I differs, there may be less literature proposing these redesigns. Though, one could calculate a p-value ignoring the sample paths for deviations in either the first or second stage as if the two-stage design were a single stage with attained values. If this were done, the decision could be different than that of the proper p-value which considers sample paths, and how to decide how small the p-value is needed in order to determine statistical significance is unclear.

We focused our investigation and results on deviations from the planned first stage sample size. One argument against redesigning these trials in the first place could be that researchers always have the option to simply wait until stage I sample size is met. In practice, though, some ethical matters may arise that would give the researcher incentive to evaluate the first stage prematurely. For instance, if a new regimen appears to be more beneficial than historical treatments, but statistical requirements prevent new subjects from being enrolled until all currently enrolled subjects record responses, a researcher may consider this unethical. In this case, $n_1^{**} < n_1$ where $n_1^{**}$ would be subjects who have recorded responses. Having a decision rule for a case such as this would alleviate some discomfort from both the researcher and statistician, though abuse of new decision rules would be discouraged.

We suggest a redesign of a trial for attained sample sizes using the planned total sample size for practical reasons such as resources and inability to stray far from the planned

design. A numerical study also suggested that it may be desirable to redesign trials using the planned total sample size because it better controls type I error for all attained methods and power is closer to the nominal power more often. Furthermore, if one employs Chang *et al.*'s method, the sample size is radical in the first stage, and the resulting probability of continuing is less than 0.05, it will be impossible for a type I error to occur. This will then reduce the two-stage design to a one-stage design which is considerably different than the planned design, so this is further support for using the planned total sample size. Assuming that redesigns use the planned total sample size, results from different two-stage designs were presented. Chang *et al.* and the Olson and Koyama methods primarily differ when there are extreme sample size shifts. This is most likely due to the nature of their methods and their primary goals of maintaining type II error spending versus probability of early termination.

Recommending the use of these designs in practice will depend on the desire of statistical approach of the researcher. Blume states, "Hypothesis testing procedures do not place any interpretation on the numerical value of the likelihood ratio. The extremeness of an observation is measured, not by the magnitude of the likelihood ratio, but by the probability of observing a likelihood ratio that large or larger. It is the tail area, not the likelihood ratio, that is the meaningful quantity in hypothesis testing" [17]. Thus, the primary argument for using Chang *et al.*'s or the OK method is in the case that the researcher does not want to abandon hypothesis testing. If the researcher prefers to use a traditional frequentist approach with hypothesis testing, it may be recommended that the Olson and Koyama's approach is used because it results in higher average power across deviations. Because it may be of concern that researchers take advantage of the ability to deviate from the planned design, Olson and Koyama's method also penalizes deviation by resulting in a higher probability of early termination when there is under-accrual than Chang *et al*'s method.

Overall, there are many advantages to using a Likelihood two-stage design. One advantage to the Likelihood design is that it is able to add cohorts of subjects at the end of the

second stage if weak evidence is obtained without threatening frequentist properties such as type I error. Another advantage to the Likelihood approach is that inference is more straightforward because one is not concerned with error rates or p-values. The likelihood is unaffected by the number of looks at the data and the evidence is independent of the probability of observing misleading evidence [17]. Though we don't consider calculating p-values when stage I differs from planned, it would be complicated if one wished to do so, whereas Likelihood methods would not require this because evidence is represented by a simple calculation of the likelihood ratio. Likelihood designs are also more generalizable. The Likelihood two-stage approach could be generalized easily to three stages, whereas the Chang *et al.* and OK designs would not be able to generalize in such a way. In this paper, though, we are very much restricting the Likelihood design and not taking full advantage of its natural characteristics. One could simply use a pure Likelihood design and avoid traditional frequentist issues altogether.

In general, though, if a researcher wanted to design a study using the restricted designs as described in this paper, the OK design and the Likelihood design are highly competitive and achieve similar results when the probability of early termination is reasonable (i.e. above 50%). One design may be more favorable than the other depending on the hypotheses being tested and the amount of over- or under-enrollment.

We do not consider redesigns when both the first and second stage accrual are not as planned because if one is interested in prespecifying stopping criteria for sample size deviations, the number of combinations needed to be specified in order to prespecify the exact combination that will occur is unreasonable. Though, these attained designs are able to accommodate if this is desired. Lastly, a primary concern that we have with redesigning trials for unplanned sample sizes is that researchers could take advantage of the these new stopping criteria and stray from the planned design too often. It is for this reason that one may consider adapting Chang *et al.*'s design using a very conservative rule in the first stage and have the probability of early termination under the null always be higher than planned.

When deviations are extreme, especially where there is underacrrual, evaluating the trial early would be highly penalized by potentially having a very high probability of early termination. Overall, intentional early or late evaluation of the first stage without sound reason is highly discouraged and will not result in optimal statistical properties.

# REFERENCES

[1] R. Porcher and K. Desseaux, "What inference for two stage phase II trials?," *BMC Medical Research Methodology*, vol. 12, p. 117, 2012.

[2] R. Simon, "Optimal two-stage designs for phase II clinical trials," *Controlled Clinical Trials*, vol. 10, pp. 1–10, 1989.

[3] T. Koyama and H. Chen, "Proper inference from simon's two-stage designs," *Statistics In Medicine*, vol. 27, pp. 3145–3154, 2008.

[4] S.-H. Jung, Y. Lee, K. Kim, and S. L. George, "Admissible two-stage designs for phase II cancer clinical trials," *Statistics In Medicine*, vol. 23, pp. 561–569, 2004.

[5] J. Whitehead, "On the bias of maximum likelihood estimation following a sequential test," *Biometrika*, vol. 73, pp. 573–581, 1986.

[6] M. Chang, H. Wieand, and V. Chang, "The bias of the sample proportion following a group sequential phase II trial," *Statistics in Medicine*, vol. 24, pp. 1009–1027, 1989.

[7] H. Guo and A. Liu, "a simple and efficient bias-reduced estimator of response probability following a group sequential phase II trial," *J Biopharm Stat*, vol. 15, pp. 773–781, 2005.

[8] S. Jung and K. Kim, "On the estimation of the binomial probability in multistage clinical trials," *Statistics in Medicine*, vol. 23, pp. 881–896, 2004.

[9] W. Tsai, Y. Chi, and C. Chen, "Interval estimation of binomial proportion in clinical trials with a two-stage design," *Statistics in Medicine*, vol. 23, pp. 15–35, 2008.

[10] S. Jung, K. Owzar, S. George, and T. Lee, "P-value calculation for multistage phase II cancer clinical trials," *J Biopharm Stat*, vol. 16, pp. 765–775, 2006.

[11] D. Zeng, F. Gao, K. Hu, C. Jia, and J. Ibrahim, "Hypothesis testing for two-stage designs with over or under enrollment," *Statistics in Medicine*, vol. 34, pp. 2417–2426, 2015.

[12] S. Green and S. Dahlberg, "Planned versus attained design in phase II clinical trials," *Statistics in Medicine*, vol. 11, pp. 853–862, 1992.

[13] T. Chen and T. Ng, "Optimal flexible designs in phase II clinical trials," *Statistics in Medicine*, vol. 17, pp. 2301–2312, 1998.

[14] Y. Li, R. Mick, and D. Heitjan, "A bayesian approach for unplanned sample sizes in phase II clinical trials," *Clinical Trials*, vol. 9, pp. 293–302, 2012.

[15] M. Chang, Y. Li, and Q. An, "Alternative designs for phase II clinical trials when attained sample sizes are different from planned sample sizes," *Biometrics and Biostatistics*, vol. 6, p. 229, 2015.

[16] D. Ayers and J. Blume, "Likelihood designs for single-arm phase II clinical trials with a binary endpoint," Manuscript in preparation.

[17] J. Blume, "Tutorial in biostatistics, likelihood methods for measuring statistical evidence," *Statistics in Medicine*, vol. 21, pp. 2563–2599, 2002.

[18] J. Blume *Personal Communication*, Spring 2016.

[19] J. Blume, "How often likelihood ratios are misleading in sequential trials," *Communications in Statistics - Theory and Methods*, vol. 37, no. 8, pp. 1193–1206, 2008.